Advancing Semantic Communication Systems through Knowledge Graphs, Generative AI, and Safeguarded AI

Chengsi Liang ( )

Submitted in the fulfilment of the requirements for the
Degree of Doctor of Philosophy

James Watt School of Engineering
College of Science and Engineering
University of Glasgow

October 2025

University of Glasgow

College of Science & Engineering

Statement of Originality

Name: Chengsi Liang

Registration Number:

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature:

Date: 29th October 2025

Abstract

Semantic communication (SemCom) represents a paradigm shift that prioritizes conveying information meaning over traditional bit-level transmission. However, SemCom faces fundamental challenges in explicitly and logically characterizing semantics within coding models. Knowledge graphs (KGs) emerge as a promising solution by encapsulating entity attributes and relational logic through structured triples of entities and relationships. The multi-modal nature of KGs, encompassing text, images, and audio data, enables comprehensive semantic representation across diverse communication scenarios.

Despite their potential, integrating KGs into SemCom systems presents three critical challenges. First, developing effective methods to align and integrate source data with KG information for coherent semantic representations remains complex. Second, reconstructing original data from KGs proves particularly difficult under adverse communication conditions. Third, the KG integration inevitably introduces additional transmission overhead that must be carefully managed.

This thesis addresses these challenges by designing and optimizing KG-based SemCom frameworks across multiple data formats. The research comprises five interconnected contributions. The first work establishes a KG-based SemCom framework for video delivery that provides foundational principles for subsequent VR applications. The second work investigates generative AI-driven SemCom networks incorporating KG utilization. The third work develops a KG-based SemCom framework for audio delivery in Internet of Sounds environments. The fourth work presents a comprehensive KG-enabled SemCom framework with detailed KG fusion methodology. The final work addresses AI safety concerns by proposing a safeguarded AI SemCom framework for secure SemCom systems.

In conclusion, the work presented in this thesis provides insight into the design of KG-empowered SemCom systems for multi-modal data transmission, which can be viewed as a foundational step towards achieving efficient semantic representation and reconstruction.

Contents

Statement of Originality i

Abstract ii

List of Publications vi

List of Tables viii

List of Tables viii

List of Figures ix

List of Figures ix

List of Acronyms xi

Acknowledgements xiv

1 Introduction 1

1.1 Background 1

1.1.1 Knowledge Graphs based Semantic Communication Systems 3

1.1.2 Text Delivery 4

1.1.3 Video Delivery 5

1.2 SemCom Systems via Generative AI 6

1.3 AI Safety in SemCom Systems 7

1.4 Motivations 9

1.5 Objectives 13

1.6 Research Contributions 15

1.7 Thesis Outline 17

2 Overview and Literature Review **19**
2.1 Semantic Representation in Natural Language Processing 19
2.2 Knowledge Graph in Deep Learning Models 20
2.3 GAI Models in SemCom 20
2.3.1 GAI Models 21
2.3.2 Multimodal Models 25
2.3.3 GAI, AIGC in SemCom systems 27
2.4 Information Theory for SemCom 28
2.5 Transceiver Design in SemCom 29
2.5.1 AI Safety in SemCom 32

3 Knowledge Graph-based SemCom frameworks **33**
3.1 KG-SemCom Framework 33
3.1.1 Knowledge Graph Preprocessing 33
3.1.2 KG Fusion based Semantic Communication Model 35
3.1.3 Semantic Channel Capacity Modeling 37
3.2 Transceiver Design in KG-SemCom 39
3.2.1 Semantic Representation 40
3.2.2 Conventional Wireless Transmission Module 43
3.2.3 Data Reconstruction 44
3.3 Simulation results and Discussion 46
3.3.1 Dataset and Simulation Settings 46
3.3.2 Performance Metrics 47
3.3.3 Simulation Results 48

4 KG-based SemCom framework for Video Delivery **54**
4.1 Video Transmission Framework in VISTA 54
4.2 SLG-based Transceiver Design in VISTA 56
4.2.1 Semantic Segmentation Module 56
4.2.2 JSCC Module 58
4.2.3 Frame Interpolation Module 59
4.3 Simulation results and discussions 61
4.3.1 Simulation Setting 61
4.3.2 VISTA Framework Performance Evaluation 61

5 GAI-driven SemCom Networks **67**
5.1 GAI-driven SemCom Network Architecture 68

5.1.1 Novel Layers in GAI-driven SemCom Architecture 70
5.1.2 Knowledge Management in GAI-driven SemCom Networks 72
5.2 Problem Formulations 73
5.2.1 Semantic Encoder 73
5.2.2 Semantic Decoder 75
5.3 Experiments and Results 75
5.4 Use Cases 78
5.4.1 Autonomous Driving 78
5.4.2 Smart City 79
5.4.3 Metaverse 79

6 Safeguarded AI-driven Semantic Communication 81
6.1 Safeguarded AI Design Principles 81
6.2 Safeguarded-AI in Semantic Communication Networks 83
6.2.1 World Models 83
6.2.2 Safety Specifications 86
6.2.3 Gatekeeper Design 87
6.3 Safeguarded AI-driven Semantic Communication Framework 88
6.4 Case Study 90

7 Conclusion and Future Work 93
7.1 Conclusion 93
7.2 Future Work 94
7.2.1 KG-empowered SemCom Systems 94
7.2.2 GAI-assisted SemCom Systems 95
7.2.3 Safeguarded AI-assisted SemCom Systems 96

Bibliography 97

List of Publications

Journal

- Le Xia, Yao Sun, **Chengsi Liang**, Daquan Feng, Runze Cheng, Yang Yang, Muhammad Ali Imran, "WiserVR: Semantic Communication Enabled Wireless Virtual Reality Delivery", IEEE Wireless Communications 30.2 (2023): 32-39.
- **Chengsi Liang**, Hongyang Du, Yao Sun, Dusit Niyato, Jiawen Kang, Dezong Zhao, Muhammad Ali Imran, 2024, "Generative AI-driven Semantic Communication Networks: Architecture, Technologies and Applications", IEEE Transactions on Cognitive Communications and Networking (2024).
- Le Xia, Yao Sun, **Chengsi Liang**, Lei Zhang, Muhammad Ali Imran, Dusit Niyato, "Generative AI for Semantic Communication: Architecture, Challenges, and Outlook", IEEE Wireless Communications 32.1 (2025): 132-140.
- **Chengsi Liang**, Xiangyi Deng, Yao Sun, Runze Cheng, Le Xia, Dusit Niyato, Muhammad Ali Imran, "Semantic Communication for the Internet of Sounds: Architecture, Design Principles, and Challenges", IEEE Wireless Communications (2025).
- **Chengsi Liang**, Yao Sun, Dusit Niyato, Muhammad Ali Imran, "Knowledge Graph Fusion Based Semantic Communication Framework", IEEE Transactions on Mobile Computing (2025).
- **Chengsi Liang**, Yao Sun, Dongzhu Liu, Dahui Yu, Muhammad Ali Imran, "Safeguarded AI-driven Semantic Communication: Design Principles, Architecture, and Challenges", IEEE Communications Standards Magazine.

Conference

- **Chengsi Liang**, Xiangyi Deng, Yao Sun, Runze Cheng, Le Xia, Dusit Niyato, Muhammad Ali Imran, "VISTA: Video Transmission over A Semantic Communication Ap-

proach", 2023 IEEE International Conference on Communications Workshops (ICC Workshops).

- Zhixiang Qiao, Yao Sun, Kairong Ma, Runze Cheng, Yixuan Fan, **Chengsi Liang**, Muhammad Ali Imran, "Channel Assignment for Image Transmission in Polar Code Based Semantic Communication", 2025 IEEE Global Communications Conference (GLOBECOM) (accepted for publication).

List of Tables

2.1 Overview of state-of-the-art AIGC applications and models. 23

3.1 Main notations with descriptions. 34
3.2 The comparison of the reconstructed sentences from KG-SemCom, DeepSC and Huffman + LDPC schemes (SNR = −3 dB). 47
3.3 The training details of KG-SemCom, DeepSC and Huffman+LDPC. 52

List of Figures

1.1 Three levels of communication systems. 1
1.2 The structure of common SemCom systems. 2
1.3 Examples of triplets in a KG. 4
1.4 Three problems of existing AI models in SemCom networks. The upper layer demonstrates end-to-end semantic transmission, where problems of “best effort" and overconfidence arise. The bottom layer represents a real-world dynamic SemCom network. The combination of network dynamics and the inaccurate output generated at the upper layer manifests as the “lack of generalization" problem. 8

3.1 The framework of KG-SemCom, including KG preprocessing, semantic representation, data transmission, and context- and knowledge-based reasoning. 34
3.2 The transceiver design of KG-SemCom. 40
3.3 The neural network structure of the semantic representation network in KG-SemCom. 41
3.4 (a) Named token similarity and (b) BERT-based sentence similarity of predictions utilizing Wikipedia (top) and TACRED (bottom) over AWGN channels and Rayleigh fading channels versus varying SNRs. 49
3.5 Named token similarity (a) and sentence similarity (b) of recovered text with 0%, 50% and 100% KGs versus varying SNRs. 50
3.6 The bit counts of transmitted messages generated by three schemes versus original messages. 51
3.7 The comparison of runtimes between KG-SemCom, DeepSC, and Huffman+LDPC coding schemes. 53

4.1 The diagram of transceiver in VISTA. 56
4.2 The examples of SLGs in original video. 61
4.3 The frame samples recovered by VISTA under varying SNRs from -9 to 6 dB. 62

4.4 Visual comparison on frame samples for original video, recovered video by LDPC codes, VISTA and JSCC-VFI with 0%, 50%, and 75% interpolation at a SNR of 0 dB. 63
4.5 PSNR performance of recovered video frames versus varying SNRs from −3 to 18 dB. 64
4.6 Total processing time (a) and transmission bits (b) for 20 consecutive video frames under different interpolation proportions. 65

5.1 The architecture of the GAI-driven SemCom networks involving three perspectives: data plane, physical infrastructure and network control plane. The data plane layer includes information creation, AIGC-based SemCom transmission and information effectiveness. The network control plane layer includes GAI-driven SemCom network management. 67
5.2 Two types of GAI models for information creation: unimodal and multimodal. Unimodal GAI models specialize in processing a single type of data, while multimodal GAI models integrate and interpret multiple data types. . . 70
5.3 GAI-driven SemCom networks with physical, semantic, and generation levels. 71
5.4 Performance comparison of GAI-SemCom and SemCom for image transmission over AWGN channel at different SNR levels ($0-20$ dB) 76
5.5 SSIM performance of images reconstructed by a GAI-driven SemCom system, a classical SemCom system and a traditional wireless communication system versus varying SNRs. 77
5.6 Processing time and total transmission bits for 10 images of a GAI-driven SemCom system, a classical SemCom system and a traditional wireless communication system. 77

6.1 The architecture of safeguarded AI framework. 83
6.2 The framework of the safeguarded-AI driven SemCom framework. 88
6.3 BLEU (1-gram) scores for the proposed framework with varying thresholds ($\tau = 0.5, 0.6, 0.7$) vs. DeepSC. 91
6.4 Failure rates of the proposed framework with varying thresholds ($\tau = 0.5, 0.6, 0.7$). 91

List of Acronyms

AoI	Age of information
AoII	Age of Incorrect Information
AIGC	Artificial intelligence-generated content
AP	Access point
API	Application programming interface
AR	Augmented reality
AV	Autonomous vehicle
AWGN	Additive white Gaussian noise
BEP	Bit error probability
BER	Bit-error rate
BLEU	Bilingual evaluation understudy
BOW	Bag of words
BPSK	Binary phase-shift keying
BS	Base station
CAV	Connected and autonomous vehicle
CBOW	Continuous bag of words
CL	Causal learning
CLIP	Contrastive language-image pre-training
CNN	Convolutional neural network
CSWA	Cross-scale window-based attention
dEA	Denoising entity auto-encoder
DDIM	Denoising diffusion implicit models
DDPM	Denoising diffusion probabilistic models
DEP	Detection error probability
DL	Deep learning
DRL	Deep reinforcement learning
FL	Federated learning
GAI	Generative artificial intelligence

GAN	Generative adversarial network
GDM	Generative diffusion model
GELU	Gaussian error linear unit
GRU	Gated recurrent unit
IC	Information content
IoS	Internet of Sounds
IoT	Internet of Things
JSCC	Joint source-channel coding
KB	Knowledge base
KF	Kalman filter
KG	Knowledge graph
KGE	Knowledge graph embedding
KL	Kullback–Leibler
KRL	Knowledge representation learning
LCS	Least common subsumer
LDPC	Low-density parity-check code
LSTM	Long short-term memory
MEC	Mobile edge computing
MED	Mixture of encoder-decoder
ML	Machine learning
MLM	Masked language model
MSE	Mean squared error
NER	Name entity recognition
NLG	Natural language generation
NLP	Natural language processing
NLU	Natural language understanding
NSP	Next sentence prediction
PFM	Pre-trained foundation model
PSNR	Peak signal-to-noise ratio
QA	Question-answering
QoS	Quality of service
RDF	Resource description framework
RL	Reinforcement learning
RNN	Recurrent neural networks
SCM	Structural causal mode
SemCom	Semantic communication

Seq2Seq	Sequence-to-sequence
SER	Symbol-error rate
SIT	Semantic information theory
SLG	Semantic location graph
SNR	Signal-to-noise ratio
S-QoS	Semantic quality of service
SSIM	Structural similarity index measure
TD	Terminal device
TFB	Transformer blocks
TFL	Transformer layers
UAV	Unmanned aerial vehicle
URLLC	Ultra-reliable and low-latency communication
VAE	Variational autoencoder
ViT	Vision Transformer
VoI	Value of information
VR	Virtual reality
XR	Extended reality

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Yao Sun and Prof. Muhammad Imran, for their valuable guidance, insightful feedback, and constant support throughout the course of my research. Their expertise and knowledge have been instrumental in shaping my works.

I would also like to thank University of Glasgow for providing the necessary resources and facilities to conduct my research. Special thanks go to my colleagues and peers for their constructive comments and helpful suggestions that contributed to the improvement of this manuscript.

Finally, I extend my heartfelt appreciation to my family, group members and friends for their encouragement, understanding, and patience during the completion of this project.

Chapter 1

Introduction

1.1 Background

According to Shannon and Weaver's framework [1], communication systems can be classified into three distinct hierarchical levels as shown in Fig. 1.1: i) technical level; ii) semantic level; iii) effectiveness level. The foundational level primarily addresses the reliable transfer of symbols between transmitter and receiver, with performance evaluated through bit or symbol-level transmission accuracy. The intermediate level encompasses the conveyance and interpretation of semantic content, constituting what is termed semantic communication (SemCom). The highest level examines how communication effectiveness enables receivers

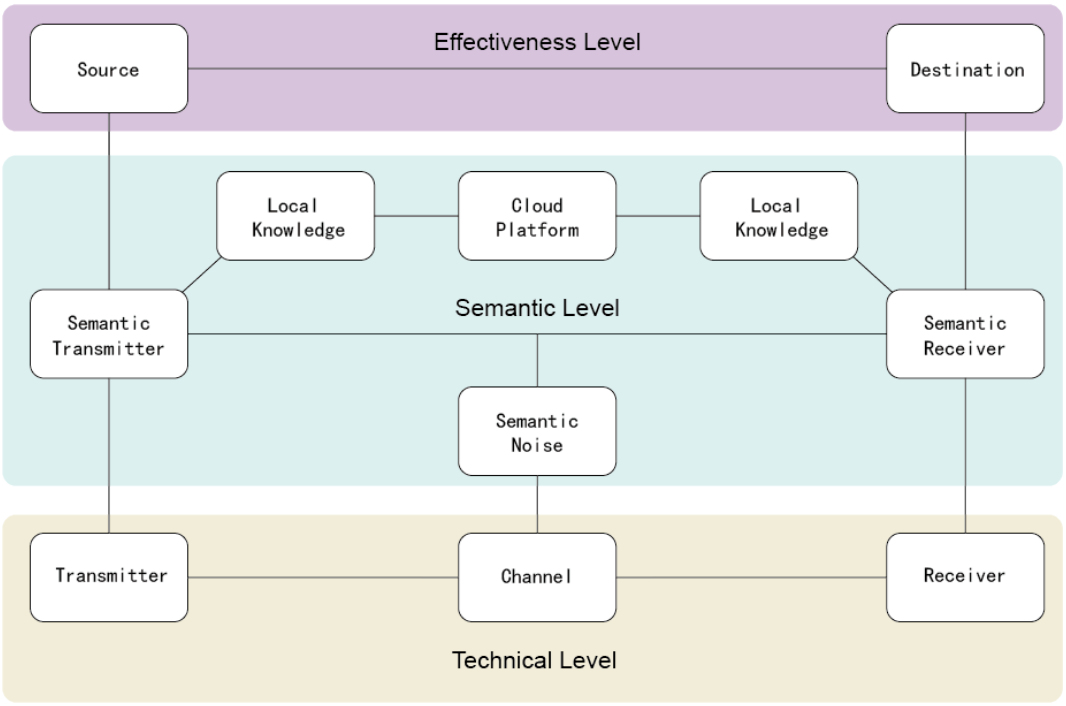


Figure 1.1: Three levels of communication systems.

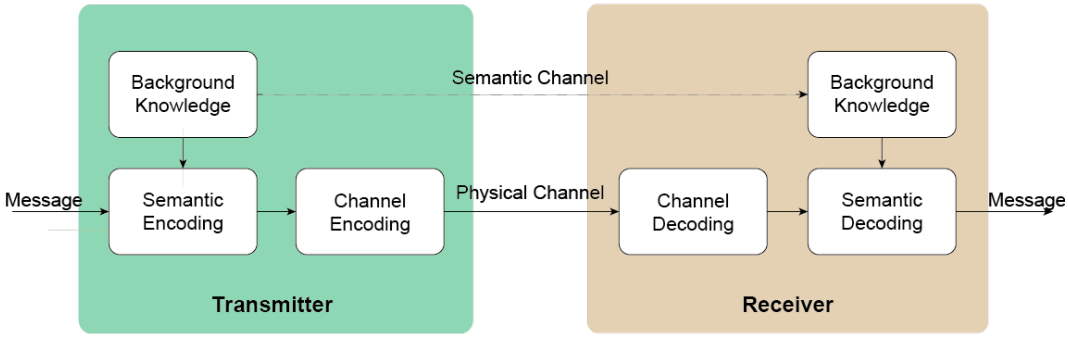


Figure 1.2: The structure of common SemCom systems.

to execute intended tasks as desired by transmitters.

Conventional communication research are predominantly at the first level, concentrating on the achieving accurate and efficient symbol transmission, typically quantified in bits, from source to destination. Such frameworks commonly employ bit-error rate (BER) or symbol-error rate (SER) as primary performance indicators. The evolutionary progression from first-generation through fifth-generation systems has yielded transmission rate improvements spanning several orders of magnitude, with system capacities increasingly approaching theoretical Shannon bounds [2]. However, emerging applications envisioned for future 6G communication systems, including augmented/virtual reality (AR/VR), autonomous driving, and metaverse environments, present unprecedented challenges that transcend traditional capacity limitations. These interconnected systems not only generate massive data volumes but also demand extensive connectivity across constrained spectral resources while maintaining stringent latency and reliability requirements, thereby creating a fundamental mismatch between application needs and conventional communication paradigms.

SemCom, deemed as a groundbreaking paradigm shift, addresses these limitations by operating within the semantic domain, extracting meaningful information while eliminating redundant, irrelevant, and non-essential data, thus achieving compression while maintaining semantic integrity. The structure of common SemCom systems is illustrated in Fig. 1.2. Typically, a transmitter in SemCom begins by extracting the hidden semantics from source data and then adapting the encoding bits to the wireless channel conditions [3, 4]. The information is then transmitted through a wireless channel, with the receiver working to recover the source's meaning aiming to minimize semantic ambiguity. Furthermore, SemCom demonstrates enhanced resilience in challenging channel conditions, particularly low signal-to-noise ratio (SNR) environments, making it well-suited for reliability-critical applications. These

considerations drive the development of intelligent communication frameworks that leverage semantic understanding to improve both accuracy and efficiency. SemCom also enables substantial data traffic reduction and proves particularly valuable in bandwidth-constrained environments, low SNR conditions, or scenarios with elevated BER/SER in conventional communication systems.

1.1.1 Knowledge Graphs based Semantic Communication Systems

In particular, background knowledge represents a critical element in SemCom systems, typically stored in the knowledge bases (KBs) of transceivers, fundamentally determining the accuracy of semantic encoding and decoding processes. Various forms of background knowledge exist, including raw corpora, ontologies, and knowledge graphs (KGs). However, the majority of existing work treats knowledge background merely as large-scale raw corpora for pretraining (e.g., sentences, audio clips, and videos), which exhibits weak generalizability and limited adaptability. Consequently, these corpus-based approaches suffer from poor compatibility with heterogeneous data sources and require substantial retraining time when incorporating new datasets, significantly hindering their practical deployment.

To address these limitations and enhance semantic translation accuracy, we propose replacing traditional KBs with KGs that explicitly emphasize structured relationships between informative knowledge facts. Specifically, we leverage the resource description framework (RDF) [5], which represents knowledge as factual triples in the form <head, relation, tail>, capturing explicit relationships between entities. KGs offer several substantial benefits for SemCom systems. First, KGs provide explicit semantic structures that enable logical reasoning and relationship inference, significantly improving semantic disambiguation capabilities. Second, the graph-based representation facilitates efficient knowledge sharing and fusion across different data modalities, enabling unified semantic encoding for heterogeneous sources. Third, KGs exhibit superior generalizability through their structured entity-relationship representations, allowing seamless integration with new domains without extensive retraining. Fourth, the relational logic embedded in KG triples enhances robustness against channel noise by providing contextual constraints for semantic reconstruction. Finally, KGs enable interpretable semantic processing through traceable entity relationships, facilitating debugging and optimization of SemCom systems.

Motivated by these substantial benefits that KGs offer for structured semantic representation, contextual understanding, and robust multimodal communication, the integration of KGs with SemCom systems has emerged as a critical research direction. Specifically, this research focuses on KG-empowered SemCom frameworks across three primary data modal-

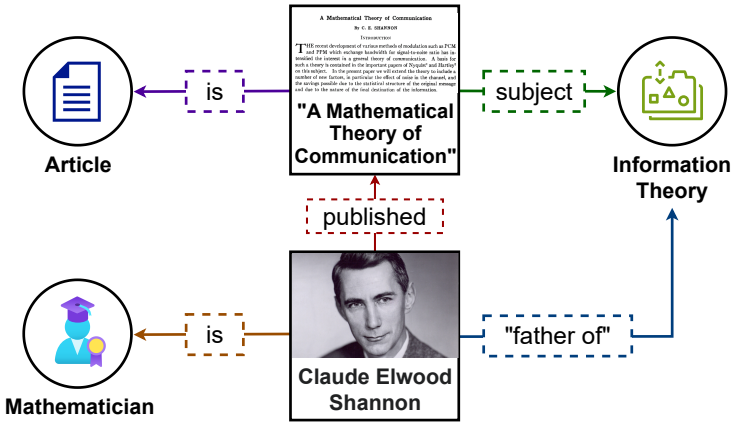


Figure 1.3: Examples of triplets in a KG.

ities: text, image and video delivery, which are presented in the following subsections.

1.1.2 Text Delivery

Recently, lots of efforts have been devoted to SemCom transceiver design for text delivery by leveraging advanced deep learning (DL) techniques, such as Transformers [6] and BERT [7]. These approaches utilize implicit reasoning methods through self-attention mechanisms to capture contextual relationships between tokens. While demonstrating promising results, these methods face several fundamental limitations. First, they represent semantics primarily through attention weights that dynamically shift based on relationships with other tokens in the sequence, resulting in semantic representations that can be both unstable and difficult to interpret. Consequently, when these systems encounter syntactic errors introduced by channel noise during transmission, they may fail to identify and correct these errors. Second, their generalization capabilities are inherently limited, as the semantic representations they develop are constrained by the distribution and characteristics of their training data. Third, these models require extensive training on large corpora to achieve accurate semantic encoding and update models, leading to considerable computational and energy resource consumption.

Consequently, we explore KGs in SemCom for text delivery, which provide an explicit semantic representation method. Unlike attention-based methods, KGs provide structured frameworks composed of numerous factual triplets that explicitly encode relationships between entities. These triplets are in the form of $\langle head, relation, tail \rangle$, where head and tail are pre-defined entities, and the relation describes the connection between them. For example, as shown in Fig. 1.3, the triplet $\langle Claude\ Elwood\ Shannon, published, "A\ Mathematical\ Theory\ of\ Communication" \rangle$ is a fact predefined in a public KG. In contrast to attention-

based approaches, KG-based semantic encoding methods facilitate more accurate, stable, and generalizable semantic representations by incorporating facts in KGs. The structured nature of KGs also enhances logical reasoning and inference capabilities in SemCom systems. For instance, even if "*Claude Elwood Shannon*" is syntactically ambiguous, the SemCom systems can correctly identify the intended entity by leveraging existing knowledge connections to "*A Mathematical Theory of Communication*" and the "*published*" relationship between them. KGs provide explicit entity-relationship structures that facilitate precise semantic disambiguation and enable logical reasoning for correcting transmission errors. The triple-based representation establishes direct mappings between textual mentions and structured knowledge entities, supporting efficient reconstruction of complex linguistic structures without exhaustive corpus coverage. The explicit semantic encoding approach through KGs enables SemCom systems to effectively resolve polysemy and ambiguity through structured knowledge representation.

1.1.3 Video Delivery

With the prosperity of multimedia services, it was witnessed that video streaming has occupied approximately 82 percent of all Internet traffic in 2022 [8] to cover a wide range of applications including live streaming, AR/VR, virtual meeting, etc. To further improve the quality of services (QoS) for users, ultra-reliable and low-latency communication (URLLC) is required, especially for real-time video applications. However, since the traditional wireless video transmission focuses on video compression and recovery via image pixels encoding and decoding, which consumes unprecedented amount of wireless spectrum and transmission time. Additionally, it may fail to achieve a satisfactory visual perception due to unstable wireless channel condition.

Motivated by the aforementioned benefits of KG-based SemCom systems, KG-based SemCom systems are expected to dramatically reduce transmitted data volume, thereby significantly conserving wireless resource consumption [9–11]. This efficiency gain becomes particularly valuable for bandwidth-intensive video delivery applications. Furthermore, recognizing that adjacent video frames exhibit strong semantic coupling at the content level, SemCom can achieve enhanced robustness especially under poor channel conditions by leveraging semantic decoders to correctly reconstruct degraded video pixels based on preserved semantic information from neighboring frames. This semantic-level temporal redundancy exploitation enables SemCom systems to maintain video quality even when individual frames experience transmission errors or distortions.

The KGs for videos exhibit greater complexity than textual due to the multidimensional

nature of visual content. Video KGs represent structured semantic knowledge that captures complex spatio-temporal relationships within video sequences, fundamentally differing from the static conceptual representations in textual KGs. Specifically, video KGs encode multiple interconnected layers of information simultaneously, including object entities with their spatial locations and visual attributes, temporal evolution and motion trajectories across frames, inter-object relationships such as interactions and occlusions, and hierarchical scene-level context that provides semantic grounding for the entire sequence.

1.2 SemCom Systems via Generative AI

The training of SemCom models and the construction of comprehensive KGs present significant technical challenges that demand substantial computational resources, large-scale annotated datasets, and domain expertise. Traditional SemCom model training requires extensive paired datasets of source content and corresponding semantic representations across different modalities, which are often scarce or expensive to obtain. Similarly, KG construction involves labor-intensive processes including entity extraction, relation identification, knowledge validation, and ontology alignment, particularly for specialized domains or emerging applications. These challenges are further compounded when dealing with multimodal data where cross-modal semantic alignment must be established and maintained.

Generative artificial intelligence (GAI) offers promising solutions to address these challenges through its remarkable capabilities in automated content generation, semantic understanding, and knowledge inference. GAI models, particularly large language models and multimodal foundation models, can significantly accelerate KG construction by automatically extracting entities and relationships from raw data, generating synthetic training samples to augment limited datasets, and performing zero-shot or few-shot knowledge inference to expand KG coverage. Furthermore, GAI techniques enable automated semantic representation learning without requiring exhaustive manual annotation, facilitate cross-modal knowledge alignment through learned embeddings, and support adaptive KG updates based on evolving communication contexts. By leveraging GAI's generative and reasoning capabilities, SemCom systems can overcome data scarcity issues, reduce manual construction efforts, and achieve more robust semantic encoding and decoding across diverse communication scenarios, thereby bridging the gap between theoretical frameworks and practical deployment.

Within the realm of GAI, AI-Generated Content (AIGC), i.e., digital content including text, image, audio and video is generated by machine learning (ML) algorithms automatically, stands as a notable application in information technology field. Recently, many AIGC products with high efficiency and knowledgeability meeting the huge demand from people

on data acquisition, have attracted much attention. One of the most significant reasons is that AIGC services are capable to deal with large-scale database in a short time due to the advanced computing ability. For instance, Claude, delivered by Anthropic's GAI, could process 100,000 tokens of text (equal to about 75,000 words in a minute) by May 2023 [12]. Especially, some advanced AIGC services are realized by sophisticated multimodal GAI models which can cope with more than one data formats. A renowned example, ChatGPT-4 [13], allows users to share images and engage in voice conversations. It significantly enriches the user experience, offering a more dynamic and interactive form of communication compared to ChatGPT-3.5's primarily text-based interactions.

Considering the superiorities of SemCom, it should be expected as a promising paradigm for AIGC transmission. It is observed that the structure and logic of the AIGC are inherently tied to GAI models, making it possible for meaning inference according to immediate context. This is highly compatible with the framework of SemCom. Meanwhile, SemCom systems enable to handle high-volume data and diverse content types with sustainable resource consumption, sufficing the needs of intricate AIGC services while alleviating internet strain. By leveraging the knowledge collected from user's history and sensing data, SemCom systems allow more intelligent personalized services.

1.3 AI Safety in SemCom Systems

Beyond the design and optimization of SemCom systems, AI safety emerges as a critical concern due to the widespread integration of AI models throughout SemCom architectures. Typically, SemCom systems are inherently safety-critical because communication infrastructure serves as the foundational backbone enabling numerous safety-critical applications and services. This criticality becomes particularly pronounced as SemCom systems increasingly support essential infrastructures including electrical power grids, intelligent transportation systems, air traffic control networks, autonomous vehicle (AV) coordination, and industrial automation platforms. These domains demand URLLC to ensure the timely and accurate exchange of information, where even slight communication errors can lead to severe consequences. SenCom, underpinned by AI models, has emerged as a key enabling technology to meet the stringent URLLC requirements. Unlike conventional communication systems that transmit syntactic symbols, SemCom systems transmit task-relevant semantic representations, which introduces new safety considerations. The semantic encoder and decoder must jointly ensure robustness and safety, as errors introduced during compression, transmission, or reconstruction can lead to semantic inconsistencies that traditional error control mechanisms cannot easily detect or correct. Even minor perturbations in semantic representations

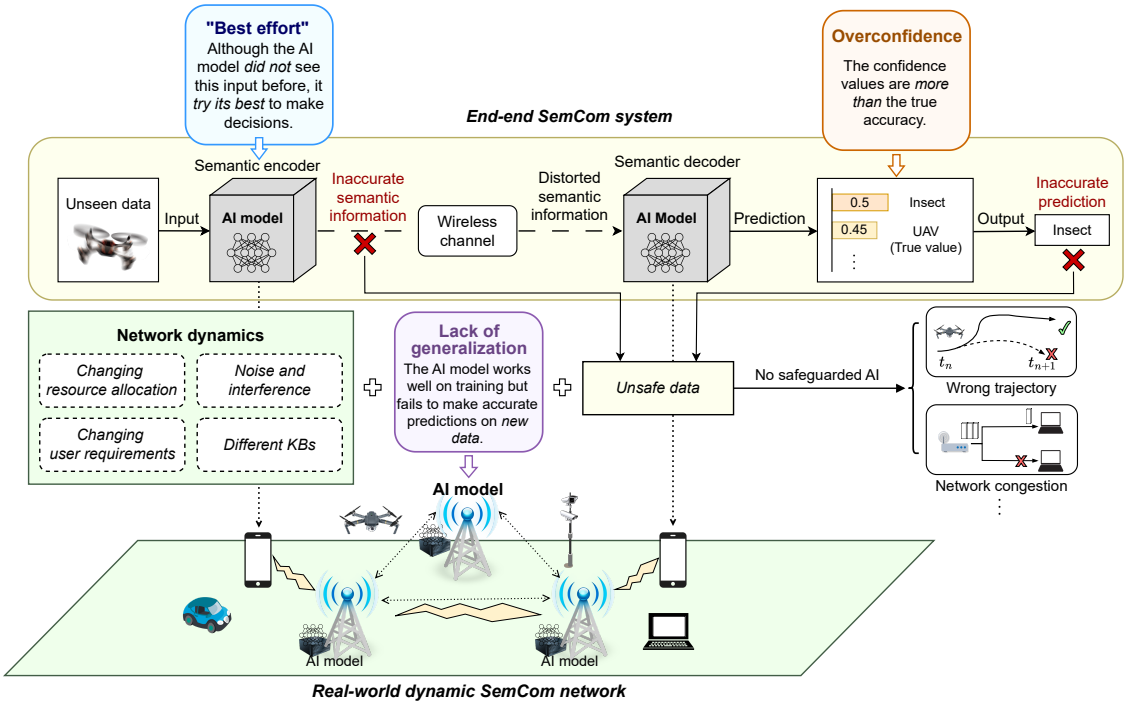


Figure 1.4: Three problems of existing AI models in SemCom networks. The upper layer demonstrates end-to-end semantic transmission, where problems of “best effort" and over-confidence arise. The bottom layer represents a real-world dynamic SemCom network. The combination of network dynamics and the inaccurate output generated at the upper layer manifests as the “lack of generalization" problem.

can result in significant misinterpretations, making end-to-end integrity critical in safety-sensitive applications.

Current AI models provide superior semantic processing tools for SemCom systems through perfect training in ideal and pre-defined communication conditions. However, in real-world deployment, AI models in SemCom networks face unforeseen network dynamics and unseen message input that expose critical safety vulnerabilities. Especially, as shown in Fig. 1.4, these models frequently lack rigorous safeguards to address safety concerns in three fundamental aspects:

- AI models operate on a "best effort" basis, such as maximum likelihood estimation, which means that models do their best to produce an output, regardless of whether it’s reliable. Even when confronted with previously unseen inputs, these models still generate outputs without guarantees. It is crucial when these models are adopted as semantic encoders, as it can result in inaccurate semantic information, further affecting the data reconstruction in semantic decoder.

- Sometimes AI models are overconfident during prediction, assigning high confidence scores to incorrect outputs. Since these models typically make decisions solely based

on the highest confidence scores, in the semantic decoder, this overconfidence manifest s as inaccurate message reconstruction.

- Network dynamics, encompassing changing channel conditions, resource allocation, and evolving user requirements, are frequent and inherent to SemCom networks. However, many AI models lack the generalization capabilities to adapt to these dynamic conditions. Under such network dynamics, the problems of "best effort" basis and overconfidence become particularly severe. Additionally, the limited generalization capabilities create specific challenges when dealing with diverse KBs in SemCom networks, which is critical to semantic encoding and decoding.

To address these challenges, the concept of "*safeguarded AI*" is introduced. Safeguarded AI is a framework aimed at ensuring that AI systems operate under strict safety protocols, especially in safety-critical applications. Building upon the broader goals of AI safety, which seeks to create systems that are human-centered, controllable, interpretable, robust, and aligned with human values, safeguarded AI distinguishes itself by employing mathematical rigor and formal verification methods over empirical safety testing to enforce these safety objectives. The framework's core innovation is an external *gatekeeper* module that provides mathematical guarantees for keeping AI outputs within proven safety boundaries [14]. Rather than relying on the model's internal confidence scores, which are often prone to overconfidence or "best-effort" decisions, the gatekeeper performs independent assessments using separate safety criteria. It acts as a safety layer that monitors real-world interactions and enforces predefined safety boundaries tailored to the system's intended use. To enable adaptive evaluation of AI behavior under dynamic conditions, the gatekeeper continuously collects network condition parameters and simulates prediction distributions. Furthermore, it can feed the collected real-time data back into the AI pipeline, enriching the training dataset and thereby enhancing the model's robustness and generalization performance over time.

1.4 Motivations

The integration of KGs, GAI, and safeguarded AI within SemCom systems demonstrates immense promise for addressing fundamental challenges in multimodal content delivery, semantic representation accuracy, and system reliability. However, translating this conceptual promise into practical implementations raises numerous critical design questions that remain unexplored. First, there is a fundamental disparity between multimodal source data and KGs in terms of representation formats. Video content consists of temporal sequences of visual frames with motion dynamics, image data comprises spatial pixel arrangements with visual

features, and audio signals contain spectral-temporal characteristics, while KGs are structured as semantic triplets of entities and relationships. These disparate data formats exist in distinct vector spaces with different dimensionalities and semantic structures, making their seamless integration and unified encoding in multimodal semantic encoders an extremely complex task. The challenge becomes even more pronounced when attempting to align visual objects in videos and images, audio events, and textual entities with corresponding KG nodes while preserving their inherent semantic relationships. Moreover, GAI techniques demonstrate remarkable capabilities in semantic understanding, content generation, and multimodal reasoning, which align well with KG-empowered SemCom requirements. Specifically, the detailed architectures and methodologies for GAI-driven SemCom frameworks, particularly for AIGC delivery, require systematic investigation. In the following, we identify and analyze the key challenges inherent in KG-empowered SemCom systems, which collectively provide the motivation and foundation for our research contributions.

- **KG-based SemCom for Text Delivery** How to leverage KG in SemCom systems is quite challenging. First, there is a fundamental disparity between large-scale textual corpora (consisting of sentences) and KGs (consisting of triplets). These two data formats exist in distinct vector spaces, making their seamless integration and encoding in the semantic encoder an extremely complex task. Second, while the fusion of semantics and public knowledge provides more comprehensive information, it also increases the volume of data to be transmitted, potentially degrading the superiority of SemCom. Moreover, both semantic and knowledge information are susceptible to distortion from noise and interference in the physical channel. This raises a critical question: how can we ensure that the potentially distorted knowledge effectively assists the decoder in correcting both transmission error and semantic ambiguity?

  Previous studies on KG-based SemCom systems [15–17] focus on converting source data into triplets based on KGs, transmitting these triplets and recovering original data from triplets. Although these approaches significantly reduce the volume of transmitted data and leverage well-trained large language models for data reconstruction, their accuracy can be compromised when relying exclusively on transmitted triplets. In such cases, when entities fall outside the KG's coverage, their decoder may fail to reconstruct messages accurately. Moreover, triplet representation is unsuitable for long and complex sentences. Long sentences with multiple clauses are difficult to capture accurately using these method [18]. As sentence complexity increases, the number of required triplets grows exponentially, potentially leading to data management and processing issues.

- **KG-based SemCom framework for Video Delivery**

  Despite many superiorities of SemCom-enabled video transmission, there are several inevitable challenges. It should be first noted that static and dynamic objects may coexist in multiple consecutive video frames, where the semantics implicit in each static object between different frames are normally identical and the change process of each dynamic object in consecutive frames should be regular. Hence, how to realize efficient semantic representation and reconstruction for consecutive frames is the first nontrivial challenge. Besides, signal attenuation and distortion in wireless channels may impose severe semantic ambiguity on transmitted videos and further greatly affect the final rendered video quality, thus the second challenge should be how to take into account different channel status in SemCom-enabled video transmission. A few pioneering works on DL-based video transmission in SemCom have been recently presented [6, 19, 20]. The authors in [6] design a deep joint source-channel coding (JSCC) framework aiming at transmitting semantics of the whole video over arbitrary wireless channels. [19] focuses on transmitting keypoints for semantic video conferencing and proposes an incremental redundancy hybrid automatic repeat-request framework to adapt varying channels. Furthermore, [20] discusses the prospect of URLLC in semantic VR delivery between the mobile edge computing server and VR users. However, these works do not leverage KGs in semantic encoding and decoding processes.

- **GAI-driven SemCom framework** A fundamental limitation of conventional SemCom systems is that, despite their ability to extract and transmit task-relevant semantic features, the receiver remains entirely dependent on what is explicitly sent over the channel. Every detail that must appear in the reconstructed output must, in some form, have been transmitted from the source. This constraint places a hard floor on the communication overhead that semantic compression alone cannot overcome, particularly for high-dimensional data modalities such as images, video, and audio, where perceptual quality demands rich and detailed reconstruction.

  Generative AI fundamentally breaks this constraint by enabling a paradigm we term 'Communicate less, synthesize the rest'. Rather than transmitting a compressed but complete representation of the source content, a GAI-empowered SemCom system transmits only the most compact semantic essence, such as structural layouts, object relationships, or high-level scene descriptions, and relies on a powerful generative model at the receiver to synthesize the remaining perceptual detail from learned priors. This shifts a significant portion of the communication burden from the channel to the receiver's generative capacity, achieving transmission efficiency that is fundamentally

beyond the reach of non-generative approaches.

To appreciate the practical significance of this distinction, consider a direct comparison between generative and non-generative SemCom under the same channel conditions and bandwidth constraints. Non-generative systems, which reconstruct outputs deterministically from received features, suffer progressive degradation in perceptual metrics such as PSNR and SSIM as bandwidth decreases, since fewer transmitted features means less information available for reconstruction. Generative systems, by contrast, can maintain substantially higher perceptual quality at the same or lower transmission overhead, because the generative model compensates for missing transmitted information through synthesis. Furthermore, by reducing the volume of data that must traverse the channel, generative SemCom can also achieve lower end-to-end latency, a critical advantage in delay-sensitive applications. These performance dimensions, namely reconstruction quality, bandwidth efficiency, and transmission latency, collectively define the space in which generative SemCom offers measurable and systematic gains over its non-generative counterparts, and motivate the framework developed in this thesis.

However, the synthesis of SemCom and GAI in intelligent wireless communication networks inevitably encounters many challenges, including:

Challenge 1: How to construct the SemCom systems fusing GAI to process any data format? The associated semantics are produced and interpreted by GAI through semantic encoder/decoder in SemCom. However, basic data processing falls short of meeting the demands posed by data-intensive AIGC services, necessitating the use of multimodal algorithms to handle diverse data types. Additionally, the computational time and power required for training must be factored in. During transmission, channel encoders and decoders should adaptively compress semantic information based on varying channel conditions.

Challenge 2: How to measure the effectiveness of information generated by GAI in SemCom-based networks? As SemCom emphasizes the conveyed message's meaning rather than transmitted bits, conventional performance indicators derived from Shannon's framework, are not suitable for evaluating SemCom networks [21]. To offer enhanced services, information effectiveness measurement is tied to the achievement of specific objectives and time. Additionally, interactions now incorporate both human-to-machine and machine-to-machine, moving beyond just human-to-human communication. Thus, determining appropriate metrics considering different goals and scenarios, poses another challenge.

Challenge 3: How to manage SemCom-based networks with GAI technologies? The rising ubiquity of ML tools across all network nodes necessitates coordinated

management of resources for computation, communication, and control. Significantly, SemCom heavily relies on background knowledge for semantic representation and interpretation. In this context, knowledge can be considered a resource that requires storage and bandwidth costs for construction and sharing. However, devices with limited storage capacity may need to reduce the size of their KBs. Furthermore, finding the right balance between knowledge freshness and update cost is crucial, as updating the KBs too frequently incurs high costs, while updating it too infrequently may result in outdated or stale knowledge. Consequently, the development and implementation of efficient knowledge management strategies are essential in SemCom-based networks, presenting the third challenge.

- **Safeguarded AI** As AI systems become increasingly autonomous and deeply embedded in critical infrastructure, ensuring their safe and reliable operation has emerged as a fundamental challenge. In semantic communication networks, this challenge is particularly acute: the AI-driven components responsible for semantic encoding, decoding, and inference are not only complex and opaque, but their failures can propagate directly into communication errors, misinterpretations, or unintended actions at the application layer. Unlike traditional communication systems where failures are largely bounded and predictable, an unsafe or unverified AI component in a SemCom network can compromise the integrity of the entire transmission pipeline. Moreover, the absence of formal safety guarantees in AI-driven SemCom systems makes it difficult to certify their behavior under distribution shifts, adversarial conditions, or unforeseen operational contexts. These vulnerabilities highlight a critical gap: while the communication community has embraced AI as a powerful tool for semantic processing, the rigorous safety frameworks necessary to govern its deployment remain largely absent. Addressing this gap requires moving beyond performance-driven design and establishing principled mechanisms that can provide verifiable, bounded guarantees on AI behavior within SemCom networks — a challenge that this dissertation directly tackles.

1.5 Objectives

This thesis is driven by the need to overcome several key challenges that hinder the practical deployment of KG-empowered SemCom systems. The primary objective is to systematically explore and propose innovative solutions that enable the effective integration of KGs in enhancing SemCom performance across diverse data modalities.

First, the research objectives of the first work are to develop an efficient semantic video

transmission framework that addresses key challenges in wireless communication systems. The primary objective is to design a semantic segmentation methodology that can accurately detect and recognize dynamic objects and static backgrounds within video frames while constructing semantic location graphs (SLGs) to capture spatial relationships and extract meaningful semantics. A secondary objective is to optimize data transmission efficiency by intelligently separating video content into environment and behavior segments, thereby reducing transmission overhead by requiring only one environment frame and select key behavior segments. Additionally, the work aims to develop robust frame interpolation techniques at the receiver that can accurately reconstruct original video content using transmitted segments and semantic information. Finally, the research objectives include demonstrating improved performance metrics in terms of data volume reduction, processing efficiency, and video quality enhancement, particularly under challenging low SNR conditions compared to existing benchmark methods.

Second, the research objectives of the second work are to develop and validate the first comprehensive GAI-driven SemCom framework specifically designed for AIGC delivery systems. The primary objective is to establish a complete framework architecture that integrates SemCom systems with GAI algorithms, including the definition of essential components, key performance indicators, and network management methodologies. A critical objective involves conducting systematic taxonomic analysis of GAI models across unimodal and multimodal domains, categorizing them into distinct operational classes to enable targeted optimization strategies. The work aims to formulate mathematical foundations and design optimized transceivers that maximize SemCom's benefits for AIGC delivery applications. Additionally, the research objectives include developing comprehensive evaluation methodologies for AIGC information effectiveness through task-oriented systems, age of information (AoI), value of information (VoI), and causal control frameworks. Furthermore, the study seeks to create innovative architectures and algorithms for joint optimization of communication and computing resources while establishing robust knowledge management protocols encompassing construction, updating, and sharing mechanisms. Finally, the research aims to demonstrate practical viability through comprehensive use case analysis across diverse application domains including autonomous driving, smart cities, and Metaverse environments.

Third, we investigate how to exploit knowledge in SemCom to enhance transmission reliability through the integration of contextual information and KGs. In particular, first, our framework simultaneously embeds tokens of source data and the aligned KG entities. As such, it is able to capture contextual information from the source, which differs fundamentally from previous KG-based SemCom methods. Second, during semantic decoding, our

framework can reconstruct data by reasoning from the contextual and knowledge relationships. This reasoning approach allows for more accurate prediction of missing or distorted tokens learning from contextual and knowledge relationships. Third, our hybrid approach ensures that even when entities fall outside the KG, overall accuracy remains relatively stable, as the textual encoder provides an effective fallback mechanism for handling new entities. Finally, our hybrid transmission method efficiently processes long and complex sentences that traditional triplet-based approaches struggle to represent.

Eventually, the research objectives of the final work are to develop and implement a comprehensive safeguarded AI framework that ensures the safety and security of SemCom systems. The primary objective is to design a robust safeguarded AI architecture comprising three essential components: world model development for environmental understanding, safety specifications for operational constraints, and gatekeeper mechanisms for monitoring and control within SemCom systems. The work aims to establish clear design principles for each framework component specifically tailored to SemCom requirements and constraints. A critical objective involves proposing an integrated safeguarded AI-driven SemCom framework that combines safety mechanisms with SemCom capabilities, including detailed architectural design and practical implementation methodologies. Additionally, the research seeks to validate the proposed framework's effectiveness through comprehensive simulation studies that demonstrate measurable improvements in semantic accuracy compared to conventional SemCom approaches. Finally, the work aims to establish foundational principles for safe AI integration in SemCom systems, providing a benchmark for future developments in secure and reliable SemCom technologies.

1.6 Research Contributions

The aforementioned objectives indicate that this thesis aims to promote further theoretical research and industrial applications of KG-driven SemCom by systematically addressing the fundamental challenges that arise in its practical deployment across diverse data modalities. Rather than treating these challenges in isolation, the contributions of this thesis collectively build toward a comprehensive SemCom framework that progresses from efficient multimodal semantic transmission, through AI-driven content generation and delivery, to rigorous semantic representation and, ultimately, to guaranteed operational safety.

Toward this end, we first establish the foundation for efficient multimodal semantic transmission by proposing VISTA, a novel VIdeo transmission framework over Semantic communicaTion Approach. VISTA addresses the core challenge of transmitting semantically rich video content over bandwidth-constrained wireless channels by integrating three tightly

coupled modules: a semantic segmentation module that detects and recognizes dynamic objects and static backgrounds, a JSCC module that manages SNR-adaptive wireless transmission, and a frame interpolation module that accurately reconstructs video at the receiver using transmitted segments and semantic scene graphs. Critically, VISTA introduces an SLG-based representation that separates video frames into environment segments and behavior segments, enabling dramatic reductions in transmitted data volume without sacrificing perceptual quality. Performance evaluations on real video datasets confirm VISTA's superiority in data volume reduction, processing efficiency, and quality enhancement, particularly under low SNR conditions.

Building upon this efficient transmission foundation, we then confront the emerging challenge of integrating generative AI into SemCom systems. As GAI models become capable of synthesizing high-quality multimodal content, a new paradigm arises in which SemCom networks must not only transmit existing data but also manage the delivery and coordination of AI-generated content. To this end, we present a pioneering GAI-driven SemCom framework for AIGC delivery, representing the first comprehensive investigation into framework architecture, components, key performance indicators, and network management approaches for this integration. This contribution provides a systematic taxonomic review of GAI models from unimodal and multimodal perspectives, formulates the mathematical theory underpinning SemCom's benefits for AIGC delivery, and introduces novel architectures for optimizing communication and computing resource allocation alongside knowledge construction, updating, and sharing protocols. The framework's practical value is validated through diverse use cases including autonomous driving, smart cities, and Metaverse environments.

However, efficient transmission and GAI-enhanced delivery alone are insufficient if the underlying semantic representations are ambiguous or unreliable. This motivates our third contribution, which addresses the theoretical and architectural foundations of semantic fidelity through a rigorous KG-based SemCom design. We mathematically model KG-SemCom and formulate the semantic channel capacity problem, explicitly accounting for syntactic channel equivocation and semantic ambiguity in both encoding and decoding. To minimize semantic ambiguity, we design a comprehensive semantic representation network leveraging KGs through a dual-component architecture comprising textual and knowledge encoders at the transmitter, and symmetrically, textual and knowledge decoders at the receiver that enable syntactic error correction through contextual and knowledge-based reasoning. Comprehensive simulations on open datasets demonstrate significantly improved accuracy for both sentence tokens and named tokens, establishing a principled semantic representation layer upon which reliable SemCom systems can be constructed.

Yet, even a semantically accurate and efficiently designed SemCom system cannot be

trusted for deployment in safety-critical applications unless its AI-driven components operate within verifiable behavioral bounds. This recognition motivates our final and culminating contribution, which guarantees the safety of the end-to-end SemCom system through a safeguarded AI framework. We present an architecture comprising three major components, namely the world model, safety specifications, and gatekeepers, and explain their design principles in the SemCom context. A safeguarded AI-driven SemCom framework is then proposed, with simulation results from a case study demonstrating significant improvements in semantic accuracy compared to conventional methods. This contribution closes the loop of the thesis: having established efficient transmission, intelligent content delivery, and reliable semantic representation in the preceding contributions, we now ensure that the entire system operates safely and accountably, thereby completing the path from foundational semantic communication design to trustworthy real-world deployment.

1.7 Thesis Outline

The rest of this thesis is as follows. In Chapter 2, we explore relevant literature across five key areas: semantic representation techniques in natural language processing (NLP), KG integration within DL models, generative artificial intelligence architectures, information-theoretic foundations for SemCom, and transceiver design methodologies in SemCom systems.

Chapter 3 investigates how to leverage knowledge in SemCom systems to enhance transmission reliability through the integration of contextual information and KGs. A KG-SemCom framework for text delivery is then presented, followed by detailed explanations of the KG fusion-based transceiver design methodology. Finally, simulation results are provided to validate the proposed framework's effectiveness.

Chapter 4 presents the architecture of an SLG-based SemCom system and describes the corresponding challenges addressed. The implementation details are subsequently provided, including the design of key modules and their operational mechanisms. Evaluation results are then presented to demonstrate the performance and effectiveness of the proposed model.

Chapter 5 presents a GAI-driven SemCom framework specifically designed for AIGC delivery. The chapter conducts a thorough investigation of AIGC information effectiveness from three critical perspectives: task-oriented systems, AoI, VoI, and causal control mechanisms. Furthermore, a novel architecture and associated algorithms are introduced for optimizing communication and computing resource allocation alongside knowledge management strategies, encompassing knowledge construction, updating, and sharing protocols necessary for operating and maintaining GAI-driven SemCom networks.

Finally, Chapter 6 presents conclusions summarizing the key contributions and findings

of this thesis, followed by an outlook on promising future research directions for SemCom systems.

Chapter 2

Overview and Literature Review

2.1 Semantic Representation in Natural Language Processing

The techniques of semantic coding can be derived from some eminent works in NLP. The most fundamental step is word embedding which translate human language into a form that machine can understand. Primarily, [22] computed semantic similarity between two entities represented by bag of words (BOW). However, BOW is lack of complexity and it ignore the association with the context. Later, [23] proposed a new representation method named word2vec which combines skip-gram and continuous bag of words (CBOW) algorithms aiming at training huge corpus and improve the translation accuracy at much lower computational cost. Furthermore, it embeds words into high dimensional vector space and explore the subtle semantic relationship between words. But it also suffer from the syntactic and synonymous problems. As a great solution, Transformer [24] was delivered with attention mechanism which is capable to capture the underlying meaning of words. Attention is the weight between each words and it can recognize synonyms. Notably, The position embedding also contributes to researching on the relationship between adjacent sentences. Bidirectional encoder representations from Transformers and BERT are promoted from transformer with masked language model (MLM). It can implement various NLP tasks including name entity recognition (NER), question answering and language inference after pretraining and fine-tuning.

2.2 Knowledge Graph in Deep Learning Models

In SemCom, KBs play the important role for transmitter and receiver which support them to encode and decode the semantic meaning. KG is constructed before the communication and they are expected to upgrade from the Cloud/edge which gathers the global shared knowledge. Commonly, KB is the same as KG, like in our project. However, there are some kinds of KB unqualified as KGs. [25] defines KG as a structured representation of facts, consisting of entities, relations and semantic descriptions. Specifically, entities are the objects and abstract concepts in the real world. Relationships represent the relationships between entities, semantic descriptions of entities and their relationships contain types and properties. Generally, KG is a directed graph including the nodes representing the entities and the lines representing the relations which have properties and attributes as well. There are many famous real-world KGs product, i.e. WordNet [26], Freebase [27], DBpedia [28], Wikidata [29], Microsoft's Satori and Google's Knowledge Graph [30]. The latest research focus on knowledge representation learning (KRL) or knowledge graph embedding (KGE) whose goal is mapping entities and relations into low-dimensional vectors which helps us extract the semantic information more easily and efficiently [31, 32]. Therefore, knowledge in KG can be compressed as a triple in the form like (head, relation, tail) or (subject, predicatem object) under the resource description framework (RDF) [25].

2.3 GAI Models in SemCom

Non-GAI has been extensively explored as the algorithmic backbone of SemCom systems, with a range of discriminative and sequential modeling approaches proposed to extract, encode, and reconstruct semantic content. The most prominent line of work centers on autoencoder-based JSCC systems [33], in which a neural encoder at the transmitter compresses source data into compact semantic features and a paired decoder at the receiver deterministically reconstructs the original content from the received representation. The landmark DeepSC [2] system exemplifies this paradigm, employing Transformer encoders and decoders to maximize semantic fidelity by minimizing sentence-level errors rather than bit-level distortions, without any generative capacity at the receiver. For visual modalities, convolutional neural networks (CNNs) [34] have been widely adopted for semantic feature extraction, segmentation, and object detection at the transmitter, with symmetric CNN decoders reconstructing visual content from transmitted feature maps, where reconstruction quality is directly bounded by the richness of what was sent. For sequential modalities such as text and speech, recurrent neural networks (RNNs)[35] and long short-term memory (LSTM) [36] net-

works have been employed to encode temporal semantic dependencies into fixed-dimensional representations, with symmetric recurrent decoders recovering the original sequence at the receiver. Beyond transceiver design, graph neural networks (GNNs) [37], including graph attention networks and graph convolutional networks, have been applied in knowledge graph-based SemCom to encode relational semantic structures into embeddings that preserve entity and relationship information across the channel. At the network management level, reinforcement learning (RL) [38] has been leveraged for adaptive resource allocation, dynamic bandwidth management, and channel-aware transmission scheduling, optimizing semantic transmission policies without contributing to content reconstruction.

2.3.1 GAI Models

While non-generative SemCom systems have demonstrated promising gains in transmission efficiency by extracting and compressing task-relevant semantic features, they are fundamentally constrained by a hard dependency on received information: the receiver can only reconstruct what has been explicitly transmitted, meaning that perceptual quality degrades inevitably as channel bandwidth decreases or channel conditions deteriorate. This ceiling becomes especially pronounced for high-dimensional modalities such as images, video, and audio, where faithful reconstruction demands rich representational detail that compression alone cannot preserve at low transmission rates. GAI offers a principled escape from this constraint by equipping the receiver with the capacity to synthesize perceptual detail that was never transmitted at all, drawing instead on learned priors about the structure and statistics of natural data. Under this paradigm, the transmitter need only convey a compact semantic essence, such as structural layouts, object identities, or scene-level descriptions, while the generative model at the receiver reconstructs a perceptually faithful output by filling in the missing information through synthesis. Specifically, the semantic information can be created by two kinds of GAI models, i.e., unimodal and multimodal models. Unimodal generative models focus on learning the distribution within a single modality and generating new samples that resemble the training data. In contrast, multimodal generative models learn the joint distribution across multiple modalities simultaneously, capturing the relationships and dependencies between them to generate samples that exhibit coherence and alignment among the modalities.

As shown in Table 3.2, we classify unimodal models based on the type of input data they work with, including text, vision (image and video) and audio.

Text-to-Text

Text-to-text GAI models are particularly effective in tasks like text generation, machine translation, summarization and question-answering (QA). These models can be divided into four categories: autoregressive models, variational autoencoder (VAE)-based models, generative adversarial network (GAN)-based models and diffusion-based models.

Autoregressive models, such as sequence-to-sequence (Seq2Seq) models [39–42] and Transformer-based models [43–46], process text sequences and predict the next textual element based on the previously generated text. Particularly, Seq2Seq models are designed to handle variable-length input and output sequences with the architectures such as LSTM [36], RNN [35], and gated recurrent unit (GRU) [47]. They are straightforward to train but tend to generate generic text. Transformer-based models are stemmed from Transformers [48], have become the backbone of many state-of-the-art GAI models (e.g., GPT 2-3 [43, 44], bidirectional encoder representations from Transformers, BERT [45], and T5 [46]) with their self-attention mechanism, excelling at handling long-range dependencies in text. Although autoregressive models can generate high-quality and coherent outputs, their sequential generation process often results in slower inference times compared to other types of generative models.

VAE-based models [49–52] function by encoding an input sequence into a fixed-dimensional representation and subsequently decoding it. They provide a probabilistic framework for learning latent representations and can be used for controlled text generation by manipulating the latent space. However, they struggle to capture long-range dependencies in text sequences due to their use of fixed-size latent representations. GAN-based models [53–57] are featured by GAN which consists of two neural networks, a generator and a discriminator, that compete with each other utilize a game-theoretic approach. Most of these GAN-based models and VAE-based models mentioned before employ RNNs in their generator and discriminator/encoder and decoder to generate text. Though GAN-based models can generate more diverse and generative content, they may be more challenging to train.

Diffusion-based models, originally designed for image generation, have recently been adapted for text generation tasks [58, 59]. These models learn to generate text by iteratively denoising a sequence of randomly corrupted text samples. At each step, these models learn to recover the original data distribution by estimating the noise that was added and removing it from the corrupted input. However, these models are computationally expensive and have a slower sampling process compared to other generative models.

Table 2.1: Overview of state-of-the-art AIGC applications and models.

Model Types			AIGC Applications	GAI Models	Model Architectures
Uni-modal	Text-to-Text		ChatGPT-3.5 [60], Bing AI [61], Megatron-Turing NLG [62], Claude 3 [63]	[39–42], GPT-2,3 [43, 44], BERT [45], T5 [46], [49–52], SeqGAN [53], [54, 55], VGAN [56], TranGAN [57], Diffusion-LM [58], DiffuSeq [59]	Autoregressive models, VAE, GAN, Diffusion models
	Vision-to-Vision (Image-to-Image, Video-to-Video)		PaintMe.AI [64], Vizcom [65], Steve.AI [66]	PixelRNN [67], PixelCNN [68], IntroVAE [69], VQ-VAE-2 [70], CycleGAN [71], StyleGAN [72], CVAE-GAN [73], Zero-VAE-GAN [74], Glow [75], Real NVP [76], DDPM [77], DDIM [78], MoCoGAN[79], [80]	Autoregressive models, VAE, GAN, Flow-based models, Diffusion models
	Audio-to-Audio		Murf.AI [81], Resemble.AI [82], MetaVoice [83]	WaveNet [84], SampleRNN [85], GANSynth [86], WaveGAN [87], SpecGAN [88], VAE-VC [89], ArchiSound [90]	Autoregressive models, VAE, GAN, Diffusion models
Multi-modal	Text2X	Text-to-Image	DALL-E 2 [91], NightCafe [92], Dream Studio [93]	DALL-E [94], DALL-E 2 [95], Magic3D [96], DreamFusion [97], Imagic [98], Uni-ControlNet [99]	Autoregressive models, GAN, VAE, Flow-based models, Diffusion models
		Text-to-Video	Synthesia [100], Pictory [101], Make-A-Video [102]	Phenaki [103], CogVideo [104], Tune-A-Video [105]	
		Text-to-Audio	Murf AI [106], PlayHT [107]	Tacotron [108], AudioLM [109]	
	X2Text	Image-to-Text	Transkribus [110]	VisualGPT [111], ViT[112]	Transformer, VAE, GAN, CNN-RNN models
		Video-to-Text	Google Cloud Video Intelligence API	UniVL [113], VideoCLIP [114]	
		Audio-to-Text	Speak AI [115]	DeepSpeech [116], wav2vec 2.0 [117]	
	Voice Bots	Speech-to-Text and Text-to-Speech	Siri [118], XiaoIce [119], Google Assistant [120], Amazon Alexa [121]	Pipelines involving ASR, NLU and NLG models [122]	Autoregressive models, GAN, VAE, RL

Vision-to-Vision

Vision-to-vision GAI models, which consist of image-to-image and video-to-video models, are utilized in tasks like photo/video editing, medical imaging, and altering facial expressions. Most of these models are built on convolutional neural networks (CNNs) to capture spatial hierarchies and local patterns in images and videos.

In particular, image-to-image models can be broadly classified into five categories: autoregressive models, VAE-based models, GAN-based models, flow-based models and diffusion-based models. Autoregressive models, such as PixelRNN [67] and PixelCNN [68], treat an image as a sequence of pixels, capture the local dependencies and patterns, and predict the value of each pixel based on the previously generated pixels with a lower speed. VAE-based models, such as IntroVAEs [69] and VQ-VAE-2 [70], aim to make the latent vectors of image encoding follow a Gaussian distribution. They allow for parallel image generation, while the images they generated may be blurry and lack sharp details. GAN-based models, like CycleGAN [71] and StyleGAN [72], make the distribution of the generated images increasingly similar to that of the real images. While they have demonstrated the ability to generate clear and realistic images, they still face challenges, including a lack of diversity in the generated outputs and potential instability during the training process. Therefore, some works like CVAE-GAN [73] and Zero-VAE-GAN [74] combine VAE and GAN models for improved generation quality and stable training.

Flow-based models, such as Glow [75] and real NVP [76], generate images by learning an invertible transformation between the data distribution and a known distribution, typically a Gaussian distribution. These models have demonstrated improved quantitative performance as measured by logarithmic likelihood, while they may have limited expressiveness and a reduced ability to capture long-range dependencies in the data. Diffusion-based models, such as denoising diffusion probabilistic models (DDPM) [77] and denoising diffusion implicit models (DDIM) [78], learn to generate a clean image by reversing a gradual noising process. Despite their impressive results in image generation tasks, diffusion-based models have some disadvantages, including high computational complexity, slow inference speed, lack of explicit latent representation and significant memory requirements.

Video-to-video GAI models are designed to tackle various video processing tasks, such as video super-resolution, video inpainting and video denoising. Similar to image-to-image models, video-to-video models can employ autoregressive models, VAE-based, GAN-based, flow-based and diffusion-based models as well. MoCoGAN [79] and the video diffusion model proposed in [80] are two notable examples. MoCoGAN, a GAN-based model, generates videos by decomposing the video generation process into separate motion and content components, whereas the video diffusion model, a diffusion-based approach, produces real-

istic and diverse video clips from random noise through an iterative denoising process.

Audio-to-Audio

Audio-to-audio GAI models can generate new sounds based on input audio for the tasks of music generation, speech synthesis, audio editing, etc. Some of these models work directly with raw audio waveforms or other audio representations. Examples include autoregressive models, such as WaveNet [84] and SampleRNN [85], as well as GAN-based models, like GANSynth [86] and WaveGAN [87]. On the other hand, some audio-to-audio GAI models first transform the audio data into a visual representation, such as a spectrogram or mel-spectrogram. They allow the audio data to be treated as an image, enabling the use of well-established image-based generative models, including GAN, VAEs, and diffusion models. SpecGAN [88], VAE-VC [89], and ArchiSound [90] are examples of GAN, VAE, and diffusion models respectively. By leveraging the success of these image-based models, researchers can generate, manipulate, and analyze audio data in the visual domain before converting the results back into the audio domain.

2.3.2 Multimodal Models

The majority of current popular AIGC services like DALL-E 2 are empowered by multimodal GAI models. Compared with unimodal GAI models, multimodal ones are more complex and versatile to process multiple types of input and output. Text data, being simple and interpretable, are often used in multimodal GAI models as textual labels or descriptions to provide supervision for training image and audio models. By leveraging the associations between text and other modalities, multimodal models can learn to generate or manipulate data across different modalities. Thus, we categorize multimodal GAI models into three main categories: text input (text2X), text output (X2text), and voice conversation (voice bots) as shown in Table 3.2.

Text-to-X

Text-to-X GAI models transform textual input into diverse output formats, such as images, videos, and audio. Text-to-image models interpret the semantic content of the input text and generate corresponding visual representations. To achieve the complex task of translating text into images, most text-to-image models [94, 95, 98, 99] employ a two-stage architecture that integrates a text understanding model and an image generation model. Particularly, DALL-E [94] and DALL-E 2 [95] leverage Transformer-based architecture to process and

understand the input text. However, DALL-E uses a discrete VAE and DALL-E 2 uses a hierarchical vector quantized (VQ)-VAE to generate high-resolution images with fine-grained details. DALL-E 2 also incorporates a diffusion model which is conditioned on the input text using a contrastive language-image pre-training (CLIP)-like encoder. Besides, Magic3D [96] can create high quality 3D mesh models based on text prompts by improving the design of diffusion models in DreamFusion [97].

Text-to-video models generate dynamic multi-frame videos that include motion and temporal coherence. A popular approach is to generate videos from text sequence-to-sequence, like Phenaki [103], where video tokens are predicted from the paired text embeddings using a bidirectional Transformer architecture. While this approach can generate high-quality video, it requires substantial computing resources. Another approach is to generate sequential images using text-to-image models and connect or interpolate them to generate a video, such as CogVideo [104] and Tune-A-Video [105].

Text-to-audio models learn the mapping between text and audio, allowing them to synthesize natural-sounding speech for any given text input. A famous example is Tacotron [108], which is a Seq2Seq model that maps character embeddings to mel-scale spectrograms, which are then converted to audio using a vocoder like WaveNet or Griffin-Lim. Another example is AudioLM [109], which can produce speech extensions that are both syntactically and semantically coherent through a multi-stage Transformer-based language model.

X-to-Text

In contrast to the text-to-X models, X-to-text GAI models enable accurate interpretation and generation of descriptive textual content from diverse input modalities, such as images, videos and audio.

Image-to-text and video-to-text GAI models are trained on the relationship between visual content and its corresponding textual representation. For example, VisualGPT [111] and vision Transformer (ViT) [112] leverage Transformer-based architecture to generate textual descriptions from image features. Moreover, video-to-text models can capture temporal information from the sequential frames of a video. UniVL [113] and VideoCLIP [114] are two examples, both of which utilize Transformer-based architectures. UniVL leverages the self-attention mechanism to capture the relationships between video frames and text tokens, while VideoCLIP extends the contrastive learning approach of the CLIP model to the video domain.

Audio-to-text models, also known as speech recognition models, focus on transcribing spoken language into written text. These models can learn the relationship between acoustic features extracted from audio signal and textual description, e.g., DeepSpeech [116] and

wav2vec 2.0 [117]. To recognize speech, DeepSpeech employs an RNN-based system that directly maps input audio spectrograms to textual transcriptions, while wav2vec 2.0 utilizes a Transformer-based architecture that learns powerful representations from masked speech input in the latent space.

Voice Bots

Voice bots, also known as voice-based chatbots or voice assistants, enable voice conversation in human-computer interaction [123]. Several products have been widely used, like Amazon's Alexa [121], Apple's Siri [118], and Google Assistant [120]. Basically, voice bots interpret user's spoken input through automatic speech recognition (ASR) algorithms, and convert the audio into text. This text is then processed using natural language understanding (NLU) techniques to understand the intent and context behind the user's query. Once understood, voice bots generate a response which converts the text back into human-like speech through natural language generation (NLG) algorithms [122].

2.3.3 GAI, AIGC in SemCom systems

Considering the collaboration between GAI and SemCom, the authors in [124] propose a framework of GAI-assisted SemCom network that integrates global and local GAI with semantic coding models in a collaborative cloud-edge-mobile design. Moreover, the authors in [125] propose a GAI-aided SemCom framework without necessitating joint training with a reduction in both computational complexity and energy cost compared to conventional SemCom methods. Advancing this line of work, the authors in [126] investigate generative SemCom with foundation models, conducting perception-error analysis and developing a semantic-aware power allocation strategy that explicitly accounts for perceptual quality at the receiver. Complementing this, the authors in [127] address the practical challenge of transmission delay by proposing a latency-aware generative SemCom framework built upon pre-trained diffusion models, demonstrating that high-quality semantic reconstruction can be achieved while satisfying stringent latency constraints. Furthermore, the authors in [128] propose a token communication framework driven by large models, enabling cross-modal context-aware semantic communications that adaptively extract and transmit semantic tokens across different data modalities. These works delve into the detailed frameworks of SemCom networks assisted by GAI, but without extensive discussions on information effectiveness and knowledge management in wireless networks.

2.4 Information Theory for SemCom

Information theory plays a crucial role in understanding and modeling communication systems. For SemCom systems, semantic information theory (SIT) provides a mathematical framework to quantify the amount of information being exchanged, assess the capacity of communication channels, and determine the optimal coding schemes to minimize errors and maximize efficiency. The goal of SIT is to quantify, measure, and optimize the semantic content of messages, taking into account the context, purpose, and impact of the communicated information. Recently, SIT has evolved over offering a broader range of perspectives on the core nature of semantic information. The authors in [129] introduce a unique theory on semantic information, emphasizing its distinct position within the information trinity. From a physics standpoint, the work of [130] characterizes semantic information as the syntactic data between a system and its surroundings, which causally aids the system's ongoing operation. The work of [130] later provides a layered interpretation of semantic information across various strata of communication systems, and employs semantic entropy used in [131] for its assessment.

Particularly, central to this theory are the concepts of semantic entropy and semantic-aware channel capacity. Derived from information entropy developed by Shannon, semantic entropy quantifies the amount of semantic information contained in a message, considering factors such as the semantic similarity between symbols or the relevance of the information to the intended receiver [4]. The authors in [132] primarily measure the semantic information using the degree of confirmation. Then, the authors in [133] explain the theory of SemCom and the entropy-based quantification of semantic information. Semantic-aware channel capacity, on the other hand, characterizes the maximum rate at which semantic information can be reliably transmitted over a communication channel, taking into account the semantic noise and distortions that may affect the interpretation of the received message. Also in [133], the authors present the calculation of semantic channel capacity of a discrete memoryless channel.

As for GAI and AIGC, [134] provides a review of recent challenges and results in the field of GAI with application to mobile telecommunications networks. [135, 136] provide surveys of techniques, applications and challenges of AIGC as the exploration of GAI. Furthermore, [137] presents a comprehensive survey on the definition, lifecycle, models, and evaluation metrics of AIGC within mobile edge networks through the combined efforts of mobile-edge-cloud communication, computing, and storage infrastructures. It also bridgs GAI technology with SemCom in the discussion on AIGC transmission. Considering the collaboration between GAI and SemCom, the authors in [124] propose a framework of GAI-assisted SemCom

network that integrates global and local GAI with semantic coding models in a collaborative cloud-edge-mobile design. Moreover, the authors in [125] propose a GAI-aided SemCom framework without necessitating joint training with a reduction in both computational complexity and energy cost compared to conventional SemCom methods. These two works delve into the detailed frameworks of SemCom networks assisted by GAI, but without extensive discussions on information effectiveness and knowledge management in wireless networks.

2.5 Transceiver Design in SemCom

Transceivers within a SemCom system can be significantly enhanced by GAI models on optimizing the understanding, transmission, and management of information. Essentially, the purpose of semantic transceivers is to extract semantics at the sender's end and restore it at the receiver's end with minimum semantic errors over different channel conditions. Learned from KB, the semantic features are first distilled by semantic encoder, then compressed by channel encoder. After passing through a physical channel, these distorted semantic features are restored by channel decoder and semantic decoder. In recent years, the prevalent architecture of SemCom has undergone significant enhancements and refinements through numerous groundbreaking works that tackle a wide range of tasks. This section categorizes SemCom's transceiver designs based on the type of source data they handle, including text, image, speech, and video.

Text delivery

By leveraging advanced NLP techniques and DL models, SemCom systems can extract and transmit the essential semantic information from the textual tokens. Transformer, a widely adopted architecture in SemCom systems for text delivery, has gained significant popularity due to its ability to effectively capture contextual relationships through the attention mechanism. A notable milestone in the development of DL-based transceiver design for SemCom is called DeepSC [2], which is built on Transformer architecture. This groundbreaking work has inspired numerous variations and advancements and leaded to a proliferation of innovative approaches. The authors in [138] integrate semantic encoding with Reed-Solomon coding, a hybrid automatic repeat request mechanism and a similarity detection network to enhance the reliability of transmitting textual semantics. The authors in [139] also propose a Transformer-based SemCom system with a new loss function to quantify the impact of semantic distortion, allowing for a dynamic balance between semantic compression loss and semantic accuracy. In order to adapt the trained model for IoT devices with limited capability, the authors in [140] come up with a lite version of DeepSC (L-DeepSC) through pruning

and quantizing the fully trained DeepSC models to achieve as large as 40× compression ratio without performance degradation.

Different from that KBs in the aforementioned research merely acting as corpora with unprocessed text, the KGs consisting of interconnected entities and their relations, enhance reasoning ability and improve personalization of SemCom. The KG-based SemCom systems are capable of predicting words based on the relationships delineated by KG, rather than solely relying on context, hence enhancing the accuracy of prediction. For example, [141] introduces a KG-driven SemCom system and utilizes Text2KG and KG2Text networks in semantic encoder and semantic decoder. Additionally, [15] delivers a more reliable SemCom system by integrating extracted semantics and KG. Specifically, it aggregates context in KG extraction and semantic restoration, which shows great robustness especially when the channel quality is poor.

Furthermore, recent research has focused on semantic generative communication systems that integrate GAI techniques into SemCom systems for text delivery. In contrast to classical SemCom systems, these integrated approaches leverage pre-trained GAI models and employ prompt processing techniques to achieve superior performance in terms of accuracy and latency. For example, the authors in [142] propose a semantic importance-aware communication scheme using pre-trained language models (e.g., ChatGPT, BERT, etc.). In addition, [143] focuses on utilizing GAI techniques to assist knowledge construction in SemCom.

Image Delivery

SemCom systems for image delivery can capture the semantic information from images by learning from the relevance between adjacent pixels. Transformer, CNNs, GANs are usually used for transceiver design to process high-dimensional image data. For example, the authors in [144–146] leverage GANs and ViT in SemCom systems to enhance the efficiency and quality of image transmission. Moreover, researchers harness adaptive and task-oriented solutions, such as RL-based adaptive semantic coding [147] and unified transmission-classification systems [148]. These approaches aim to optimize the transmission and processing of images based on their semantic content and specific vision-related tasks. The application of SemCom in unmanned aerial vehicle (UAV) scenarios has also gained attention, with studies focusing on task-oriented scene classification [149] and personalized semantic encoding for energy-efficient transmission [150].

Furthermore, cooperating with multimodal GAI models is a promising approach for image SemCom systems. In [151], the authors address a GAN-based SemCom framework to reduce communication overhead and maintain the QoS of emerging applications. The authors in [124] propose a GAI-assisted SemCom network framework in a cloud-edge-mobile

design with a case study built on Stable Diffusion for image transmission service. The author in [152] also introduce a diffusion-based SemCom system with multimodal prompts for accurate content decoding.

Video Delivery

Video delivery in SemCom systems presents unique challenges compared to image delivery, as it requires maintaining temporal consistency between sequential frames to account for the time dimension. Furthermore, video content allows for reasoning based on action/trajectory logic and behavior patterns, enabling a deeper understanding of the video semantics. Researchers have proposed various approaches to address these challenges and optimize video transmission in SemCom systems. In [6], the authors design a joint source-channel coding strategy that optimizes the trade-off between transmission rate and distortion for over-the-air video transmission. In the context of video conferencing, [19] introduces a novel method that employs a semantic error detector and utilizes a still photo of the speaker as prior information. By leveraging this prior information, the system can reconstruct the speaker's facial expressions more effectively.

Moreover, researchers have explored novel approaches that go beyond traditional video transmission techniques, such as text-based video editing and transmission. In [153], the authors introduce a method for editing talking-head videos through text modification, allowing for flexible and efficient video manipulation. Similarly, the authors in [154] propose transmitting only text instead of video to substantially relieve network traffic, leveraging the power of NLP to convey video content.

Audio Delivery

Utilizing cutting-edge NLP techniques, spoken words can be efficiently converted to text, which can then be channeled into SemCom. However, compared to text, which is purely composed of characters, the intricacies of speech signals make them more challenging to handle. This complexity arises from factors beyond just the fidelity and volume of the speech, encompassing its frequency and tone as well. The authors in [155] introduce a speech-focused variant of DeepSC, called DeepSC-S. The authors in [156] extend DeepSC-S to DeepSC-ST which leverages RNNs to extract and transmit textual semantic content from speech signals. Also, the authors in [157] extend DeepSC-S to accommodate multiple users, deploying federated learning (FL) to collaboratively train a CNN-based encoder and decoder across various local devices and a central server. Additionally, in [158], the authors propose a novel audio SemCom system based on a diffusion model which can simultaneously restore the received

information from multiple degradations, including corruption noise and missing parts caused by transmission over the noisy channel.

2.5.1 AI Safety in SemCom

Recent advances in AI have brought safety concerns to the forefront of research attention. The fundamental properties of safe AI systems have been extensively explored in [159] and [160], both of which also investigate formal methods for verifying these properties. Building on this foundation, [161] introduces the concept of "guaranteed safe AI", proposing a safeguarded AI framework that envisions structured collaboration between humans and frontier AI systems while establishing a systematic family of approaches to AI safety. Beyond these theoretical foundations, [162] provides a focused analysis of safety challenges and mitigation strategies specific to large language models. The scope of AI safety research has also expanded into communication scenarios. [163] presents a comprehensive survey of AI-based safety solutions across various communication technologies and application domains, while [164] advances this line of work by enhancing AI model safety through conformal prediction techniques with formal calibration guarantees.

Chapter 3

Knowledge Graph-based SemCom frameworks

In this chapter, we investigate how to leverage knowledge in SemCom systems to enhance transmission reliability through the integration of contextual information and KGs. A KG-SemCom framework for text delivery is then presented, followed by detailed explanations of the KG fusion-based transceiver design methodology. Finally, simulation results are provided to validate the proposed framework's effectiveness.

3.1 KG-SemCom Framework

As shown in Fig. 3.1, the KG-SemCom framework consists of KG preprocessing, semantic representation and compression, as well as semantic recovery and data reconstruction modules. In this paper, we take the text message transmission as an example. The main notations are presented in Table 3.1.

3.1.1 Knowledge Graph Preprocessing

We first preprocess triplets in the transmitter's KG. We assume that both transmitter and receiver utilize identical KGs downloaded in advance from public platforms such as cloud servers. These KGs can be updated by incorporating new knowledge shared from other nodes. In this work, entity indexes within the KG remain fixed, ensuring that the update do not affect existing entity embeddings. Thus, transceivers can exchange only new triplets rather than the entire KG when updates occur. To maintain communication quality, these updates are strategically scheduled during off-peak periods when network traffic is minimal. Additionally, for devices with limited computational resources, the KG can be optimized by

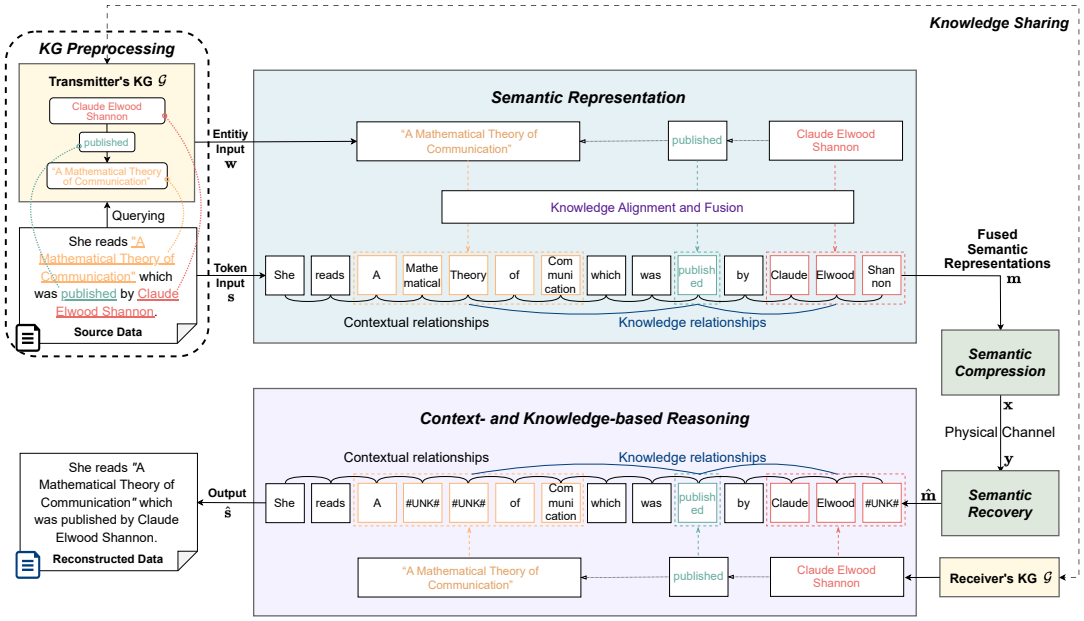


Figure 3.1: The framework of KG-SemCom, including KG preprocessing, semantic representation, data transmission, and context- and knowledge-based reasoning.

Table 3.1: Main notations with descriptions.

Notation	Description
$\mathcal{G}(\mathcal{E},\mathcal{R})$	The KG with entities $\mathcal{E}$ and relations $\mathcal{R}$.
(h,r,t)	An entity-relation-entity triplet in $\mathcal{G}$ with a head $h \in \mathcal{E}$, a relation $r \in \mathcal{R}$, and a tail $t \in \mathcal{E}$.
$\mathbf{s}$, $\mathbf{s}_w$	Tokens in a source sentence, and the named tokens in $\mathbf{s}$.
$\mathbf{w}$	The entities that can be matched with $\mathbf{s}$.
$\mathbf{e}_s$, $\mathbf{e}_w$	Token and entity embeddings respectively.
$\mathcal{S}_E, \alpha$	The semantic representation network with the parameter set α.
$\mathbf{m}$	Integrated semantic representations.
$\mathcal{C}_E$	The semantic compression network.
$\mathbf{x}$	Transmission symbols.
$h, \mathbf{n}$	Channel gain and noise in the channel.
$\mathbf{y}$	Received symbols.
$\mathcal{C}_D$	The semantic recovery network.
$\hat{\mathbf{m}}$	Recovered symbols.
$\mathcal{S}_D, \beta$	The data reconstruction network with the parameter set β.
$\hat{\mathbf{s}}$	Reconstructed tokens.

pruning less relevant nodes and edges, thereby reducing query time and memory consumption.

The KG in transmitter/receiver is denoted as $\mathcal{G}$, consisting of *entities* $\mathcal{E}$ and *relation* $\mathcal{R}$ [20]. In $\mathcal{G}$, an entity-relation-entity triplet (h,r,t), $h \in \mathcal{E}$, $r \in \mathcal{R}$, and $t \in \mathcal{E}$ represents head, relation and tail respectively. We denote the tokens in a sentence as $\mathbf{s} = \{s_1, \ldots, s_{l_s}\}$, where l_s is the number of tokens in the sentence. We then query these tokens in $\mathcal{E}$. The tokens successfully matched with entities, referred as named tokens, are denoted as $\mathbf{s}_w \in \mathcal{E}$. Here, a matrix $\mathbf{v} = \{v_1, \ldots, v_{l_s}\}$ is employed to indicate whether the tokens can be queried, i.e.,

$$v_i = \begin{cases} 1, \text{ if } s_i \in \mathcal{E}, \\ 0, \text{ otherwise.} \end{cases} \tag{3.1}$$

Particularly, each matched entity can be obtained from its corresponding tokens through a mapping function f, which can be expressed as $w_j = f(s_w)$. The details of this mapping function are explained in [165]. The set of all matched entities can be denoted as $\mathbf{w} = \{w_1, \ldots, w_{l_w}\}$, where l_w is the number of entities in the sentence.

3.1.2 KG Fusion based Semantic Communication Model

Based on the KG preprocessing, we describe a typical SemCom process. First, we create semantic representations integrating contextual and knowledge information at the transmitter end. The source tokens and matched entities are mapped into a low-dimensional vector space with normalization via a language representation approach [165]. These low-dimensional vectors are called token embeddings and entity embeddings, denoted as $\mathbf{e}_s = \{\tilde{s}_1, \ldots, \tilde{s}_{l_s}\}$ and $\mathbf{e}_w = \{\tilde{w}_1, \ldots, \tilde{w}_{l_w}\}$ respectively. These embeddings are subsequently fed into the semantic representation network for integration and fusion. In the semantic representation network, each entity embedding is aligned with its corresponding token embedding in sequence, thereby creating token-entity embedding fusions. These fusions are converted into the unified semantic representations $\mathbf{m} = \{m_1, \ldots, m_{l_s}\}$ through NLP networks such as Transformer layers [24]. The process of semantic representation can be formulated as

$$\mathbf{m} = \mathcal{S}_E(\mathbf{s}, \mathbf{w} \in \mathcal{G}; \alpha), \tag{3.2}$$

where $\mathcal{S}_E$ and α represent the network of the semantic representation module and its network parameter set respectively.

Subsequently, a semantic compression module condenses the heterogeneous semantic representations $\mathbf{m}$ into complex symbols $\mathbf{x}$. Let $\mathcal{C}_E$ represent the network of the semantic

compression module, and thus, the symbols $\mathbf{x}$ are generated as

$$\mathbf{x} = \mathcal{C}_E(\mathbf{m}). \tag{3.3}$$

Then, $\mathbf{x}$ is transmitted over a wireless channel. We consider an additive white Gaussian noise (AWGN) channel model with a channel gain h and a noise vector $\mathbf{n}$. Thus, the received symbols $\mathbf{y}$ are expressed as:

$$\mathbf{y} = h\mathbf{x} + \mathbf{n}. \tag{3.4}$$

At the receiver end, the reconstructed semantic representations $\hat{\mathbf{m}}$ are recovered from $\mathbf{y}$ via a semantic recovery module, *i.e.,*

$$\hat{\mathbf{m}} = \mathcal{C}_D(\mathbf{y}), \tag{3.5}$$

where $\mathcal{C}_D$ represents the semantic recovery module. Finally, the original tokens are reconstructed from these semantic representations through reasoning from context and the knowledge in the receiver's KG $\mathcal{G}$. The network of the data reconstruction module is constructed symmetrically to the semantic representation network, which is represented by $\mathcal{S}_D$ with its parameter set β. The reconstructed tokens, denoted as $\hat{\mathbf{s}}$, are written as

$$\hat{\mathbf{s}} = \mathcal{S}_D(\hat{\mathbf{m}}, \mathcal{G}; \beta). \tag{3.6}$$

Note that the KG-SemCom framework can be adapted to various formats of source data, including text, images, videos and audio. This adaptability stems from the fundamental nature of KGs, which can represent and interconnect diverse forms of information. For textual data, the framework can directly map words and phrases to KG entities and relationships. In the case of visual data like images and videos, the framework can leverage computer vision techniques to extract semantic concepts, objects, and scenes, which can then be aligned with KG entities. For instance, an image of a landmark could be associated with its matched entity in the KG, along with related historical and geographical information [20, 166–168]. In the realm of audio, speech recognition and natural language processing techniques can be utilized to convert spoken words into text, which can then be mapped to KG entities [169]. Additionally, non-speech audio elements like music or environmental sounds can be classified and linked to relevant KG concepts [170, 171].

3.1.3 Semantic Channel Capacity Modeling

We model semantic channel capacity based on Shannon's information theory [1] and semantic information theory [133]. First, in Shannon's information theory, the Shannon entropy of the tokens in a source sentence $\mathbf{s}$ is represented as

$$H(\mathbf{s}) = -\sum_{s \in \mathbf{s}} p(s) \log p(s), \tag{3.7}$$

where p(s) is the statistical probability of $\mathbf{s}$. We denote the semantic representation set of the sentence without incorporating the KG $\mathcal{G}$ as $\mathbf{m}$. For a token s and its generated semantic representation m, we use the notation $m \models s$ to indicate that m semantically represents s. This relationship implies that if m holds true, s must also be true. Consequently, we define the semantic representations of s as m_s, expressed as $m_s = \{m \in \mathbf{m} \mid m \models s\}$. In turn, we use $\theta(\cdot)$ to represent logical probabilities of a token, which quantifies the likelihood of encountering a background KB in which the statement is true [133]. Hence, the logical probability of token s is calculated as

$$\theta(s) = \frac{p(\mathbf{m}_s)}{p(\mathbf{m})} = \frac{\sum\limits_{m \in \mathbf{m}, m \models s} p(m)}{\sum\limits_{m \in \mathbf{m}} p(m)}. \tag{3.8}$$

According to semantic information theory [133], the semantic information content I_s of s can be obtained as

$$I_s(s) = -\log(\theta(s)). \tag{3.9}$$

Now, we consider a KG $\mathcal{G}$ integrated into semantic coding, then the semantic representations $\mathbf{m}$ should be restricted to the set compatible with $\mathcal{G}$. We use $\mathbf{m}'$ to represent the new semantic representation set given $\mathcal{G}$. Moreover, the statistical and logical probabilities of $\mathbf{m}$ need to be updated according to the knowledge in $\mathcal{G}$. We first denote the probability of the named tokens $\mathbf{s}_w$ in the source sentence as p_w. Then, let $p_{\mathcal{G}}(m')$ represent the probability of $m' \in \mathbf{m}'$ in $\mathcal{G}$. We denote the new probability of m' as $p(m' \mid \mathcal{G})$, i.e.,

$$p(m' \mid \mathcal{G}) = (1 - p_w) p(m') + p_w p(m') p_{\mathcal{G}}(m'). \tag{3.10}$$

Hence, the conditional logical probability of s given $\mathcal{G}$ is updated as

$$\begin{aligned}\theta(s \mid \mathcal{G}) = (1-p_w)&\left(\frac{\sum\limits_{m'\in\mathbf{m}', m'\models s} p(m')}{\sum\limits_{m'\in\mathbf{m}'} p(m')}\right)\\ &+ p_w\left(\frac{\sum\limits_{\substack{m'\in\mathbf{m}', m'\in\mathcal{G},\\ m\models s}} p(m')p_{\mathcal{G}}(m')}{\sum\limits_{m'\in\mathbf{m}', m'\in\mathcal{G}} p(m')p_{\mathcal{G}}(m')}\right).\end{aligned} \tag{3.11}$$

Therefore, the semantic information content of s given $\mathcal{G}$ is expressed as

$$I_s(s \mid \mathcal{G}) = -\log(\theta(s \mid \mathcal{G})). \tag{3.12}$$

Consequently, the semantic entropy of the tokens in the source sentence $\mathbf{s}$ is calculated as

$$H_s(\mathbf{s} \mid \mathcal{G}) = \sum_{s\in\mathbf{s}} p(s) I_s(s \mid \mathcal{G}). \tag{3.13}$$

Finally, we model the semantic channel capacity, consisting of syntactical and semantic aspects. First, based on Shannon's Theorem [1], we formulate the syntactical mutual information between the semantic representations of input tokens $\mathbf{m}$ and the received semantic representations $\hat{\mathbf{m}}$, which is calculated as $I(\mathbf{m};\hat{\mathbf{m}}) = H(\mathbf{m}) - H(\mathbf{m}|\hat{\mathbf{m}})$. Then, we incorporate the semantic ambiguities of semantic encoding and decoding. In particular, the semantic ambiguity refers to the degree of uncertainty for multiple interpretations when encoding/decoding semantic information. For convenience, we denote the semantic representations of $\mathbf{s}$ and $\hat{\mathbf{s}}$ as $\mathbf{m}$ and $\hat{\mathbf{m}}$ respectively, instead of $\mathbf{m}'$ and $\hat{\mathbf{m}}'$. In this context, given $\mathcal{G}$, let $H_s(\mathbf{m}|\mathbf{s},\mathcal{G})$ and $H_s(\hat{\mathbf{s}}|\hat{\mathbf{m}},\mathcal{G})$ represent the semantic ambiguities of encoding and decoding processes, respectively. Thus, the semantic channel capacity is expressed as

$$\begin{aligned}C_s &= \sup_{p(\mathbf{s}|\mathcal{G})} I(\mathbf{s};\hat{\mathbf{s}} \mid \mathcal{G})\\ &= \sup_{p(\mathbf{s}|\mathcal{G})} \left[I(\mathbf{m};\hat{\mathbf{m}}) + H_s(\mathbf{s}|\mathcal{G}) - H_s(\mathbf{m}|\mathbf{s},\mathcal{G}) - H_s(\hat{\mathbf{s}}|\hat{\mathbf{m}},\mathcal{G})\right].\end{aligned} \tag{3.14}$$

In this paper, we focus on reducing semantic ambiguities $H_s(\mathbf{m}|\mathbf{s},\mathcal{G})$ and $H_s(\hat{\mathbf{s}}|\hat{\mathbf{m}},\mathcal{G})$ by optimizing the semantic encoder/decoder network, rather than addressing syntactic channel capacity. Since the encoder and decoder networks are symmetric and use the same KG, our main goal is to minimize semantic ambiguity in the encoder network. The semantic ambiguity is directly influenced by the network parameter set and the KG $\mathcal{G}$. A higher semantic

ambiguity means the encoder produces multiple possible interpretations for the same input.

To optimize our model, we employ cross entropy (CE) as the loss function during model training, which effectively minimizes the discrepancy between the original data distribution and the reconstructed data distribution. If the CE is high, the encoded representations introduce distortion, which could lead to semantic ambiguity. In our model, the CE between $\mathbf{s}$ and $\hat{\mathbf{s}}$ can be extracted the Kullback-Leibler (KL) divergence between $\mathbf{s}$ and $\mathbf{m}$ for the encoder network. The CE between $\mathbf{s}$ and $\hat{\mathbf{s}}$ is expressed as

$$\mathcal{L}(\mathbf{s}, \hat{\mathbf{s}}, \mathcal{G}) = H(p(s_i)) + D_{\mathrm{KL}}(p(s_i) \parallel p(s_i|\mathcal{G})). \tag{3.15}$$

The second term can be expressed as

$$\begin{aligned} D_{\mathrm{KL}}(p(s_i) \parallel p(\hat{s}_i|\mathcal{G})) &= \sum_{i=1}^{l_s} p(s_i) \log\left(\frac{p(s_i)}{p(\hat{s}_i)}\right) \\ &= \sum_{i=1}^{l_s} p(s_i) \left[\log\left(\frac{p(s_i)}{p(m_i|\mathcal{G})}\right) + \log\left(\frac{p(m_i|\mathcal{G})}{p(\hat{m}_i)}\right) + \log\left(\frac{p(\hat{m}_i)}{p(\hat{s}_i|\mathcal{G})}\right)\right] \\ &= D_{\mathrm{KL}}(p(s_i) \parallel p(m_i|\mathcal{G})) + \sum_{i=1}^{l_s} p(s_i) \left[\log\left(\frac{p(m_i|\mathcal{G})}{p(\hat{m}_i)}\right) + \log\left(\frac{p(\hat{m}_i)}{p(\hat{s}_i|\mathcal{G})}\right)\right], \end{aligned} \tag{3.16}$$

where $p(s_i)$ is the real probability of the i-th token in $\mathbf{s}$, $p(m_i|\mathcal{G})$, $p(\hat{m}_i)$ and $p(\hat{s}_i|\mathcal{G})$ are the probability distribution of the i-th token in $\mathbf{m}$, $\hat{\mathbf{m}}$, and $\hat{\mathbf{s}}$ over $\mathcal{G}$ respectively. Approximately, the $D_{\mathrm{KL}}(p(s_i)||p(m_i|\mathcal{G}))$ is regarded as the semantic ambiguity of encoder network. As a result, in the process of reducing the CE in (3.15), we achieve minimized $D_{\mathrm{KL}}(p(s_i)||p(m_i|\mathcal{G}))$, which further reduces the semantic ambiguities and optimizes the channel capacity in (3.14). Moreover, we calculate CE for both predicted tokens and entities, with the detailed design elaborated in Section 3.2.3.

3.2 Transceiver Design in KG-SemCom

The neural network structure of the transceiver design in KG-SemCom is illustrated in Fig. 3.2. First, the tokens in a sentence $\mathbf{s}$ in the corpus $\mathcal{S}$, and the matched entities $\mathbf{w}$ in the KG $\mathcal{G}$ are fused and transformed into semantic representations $\mathbf{m}$ in the semantic representation network. $\mathbf{m}$ are encoded into symbols $\mathbf{x}$ in the semantic compression layers, and then transmitted over a physical channel. At the receiver end, the distorted symbols $\mathbf{y}$ are received and decoded into semantic representations $\hat{\mathbf{m}}$ in the semantic recovery layers. In turn, $\hat{\mathbf{m}}$ is

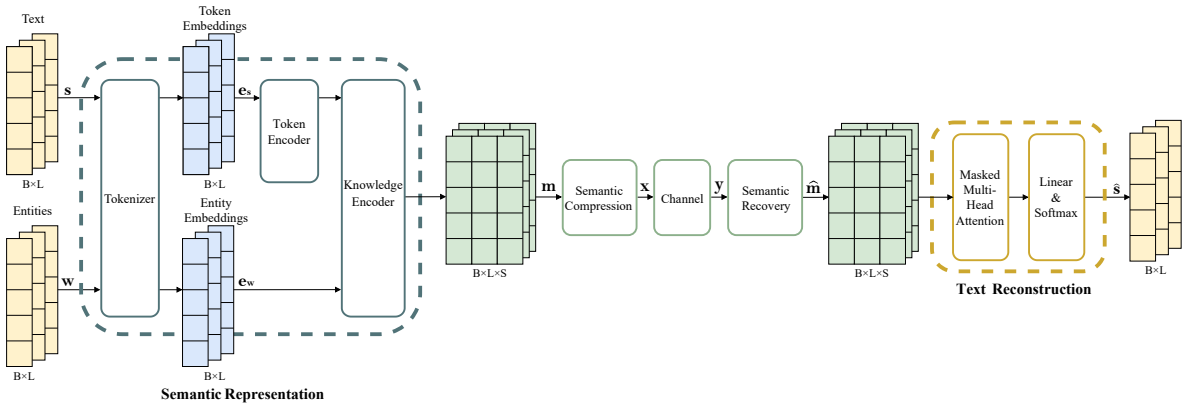


Figure 3.2: The transceiver design of KG-SemCom.

Algorithm 1 KG-SemCom network training algorithm

1: **Input**: The corpus $\mathcal{S}$ and the KG $\mathcal{G}$.
2: Query tokens in $\mathcal{G}$. Create the set of matched entities $\mathbf{w}$.
3: **Transmitter**:
4: BatchSource $(\mathcal{S}) \rightarrow \mathbf{s}$, $\mathcal{G} \rightarrow \mathbf{w}$.
5: $\mathcal{S}_E(\mathbf{s}, \mathbf{w}; \alpha) \rightarrow \mathbf{m}$. (18)-(22)
6: $\mathcal{C}_E(\mathbf{m}) \rightarrow \mathbf{x}$.
7: Transmit $\mathbf{x}$ over a wireless channel.
8: **Receiver**:
9: Receive $\mathbf{y}$.
10: $\mathcal{C}_D^{-1}(\mathbf{y}) \rightarrow \hat{\mathbf{m}}$.
11: $\mathcal{S}_D^{-1}(\mathbf{m}; \beta) \rightarrow \hat{\mathbf{s}}$. (23)-(26)
12: Compute the total loss function $\mathcal{L}$. (27)-(30)
13: Train $\alpha, \beta \rightarrow$ Gradient descent $(\alpha, \beta, \mathcal{L})$.
14: **Output**: The whole network $\mathcal{S}_E(\cdot)$, $\mathcal{C}_E(\cdot)$, $\mathcal{C}_D^{-1}(\cdot)$, $\mathcal{S}_D^{-1}(\cdot)$.

reconstructed into tokens $\hat{\mathbf{s}}$ by a text reconstruction network. The whole networks are optimized using stochastic gradient descent with CE as the loss function. Algorithm 1 shows the network training procedures, and the details of network structure are discussed as follows.

3.2.1 Semantic Representation

We first investigate the design of the semantic representation network, shown in Steps 3-5 in Algorithm 1. This network creates semantic representations from the source text and the transmitter's KG, and then encodes these representations. The process begins by extracting tokens at the subword level from the source data. For example, the word "communication" is tokenized into "com", "##mun", and "##ication". These tokens are queried in the KG using indexed lookups, where the relationships between tokens and entities have been predefined. In this mean, tokens are mapped directly to their corresponding entities with less

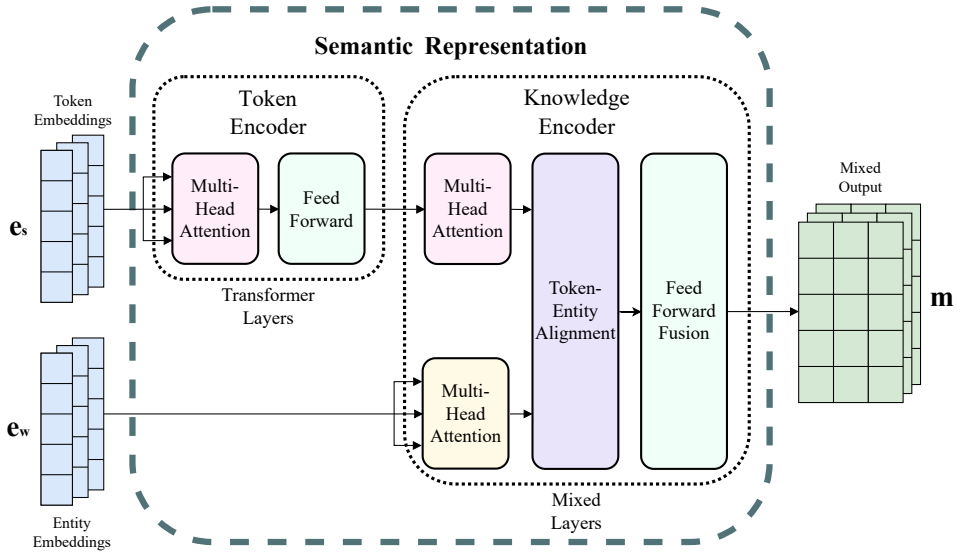


Figure 3.3: The neural network structure of the semantic representation network in KG-SemCom.

query time and computational complexity. Notably, when an entity is not found in the KG, it is recognized as a token rather than an entity. Furthermore, if a word does not exist in the training dataset, our model can understand unseen words by leveraging its comprehension of subword compositions. In cases where the unseen words cannot be split into recognizable subwords, these words are marked with "unknown" labels. The decoder will predict possible replacements based on contextual information.

Then, both tokens and their aligned entities are mapped into a low-dimensional vector space and transformed into token and entity embeddings. To enhance the model's contextual understanding, we incorporate next sentence prediction (NSP) through the addition of sentence position labels at the beginning of each sentence. As shown in Fig. 3.3, these embeddings are fed into the semantic representation network. Inspired by [172], the semantic representation network consists of a textual encoder (T-encoder) and a knowledge encoder (K-encoder). The T-encoder captures essential lexical and syntactic information from input tokens. Building upon this, the K-encoder integrates additional knowledge into the textual representation produced by the T-encoder. This hierarchical structure enables the semantic representation network to synthesize diverse information sources, effectively combining basic linguistic features with higher-level knowledge. The output is a unified feature that represents both tokens and entities, creating a integrated representation of the input text.

The semantic representation network, inspired by BERT's architecture, comprises twelve bidirectional Transformer encoder layers. The T-encoder consists of N layers, and the K-encoder contains the remaining $12-N$ layers. Each Transformer layer consists of a multi-head self-attention sub-layer followed by a position-wise fully connected feed-forward net-

work. Layer normalization and residual connections are applied after each sub-layer to stabilize training and improve gradient flow.

In particular, the T-encoder first combines three distinct types of embeddings for each token in a sentence $\mathbf{s}$ including the token embeddings, segment embeddings, and positional embeddings. The T-encoder uses these integrated embeddings to compute and extract the token's lexical and syntactic features $\mathbf{m}_s = \left\{\tilde{s}_1, \ldots, \tilde{s}_{l_s}\right\}$. Let $\mathcal{T}_e^{(i)}$ represent the i-th Transformer layer of the T-encoder, the output of this layer is denoted as

$$\mathbf{m}_s^{(i)} = \mathcal{T}_e^{(i)}(\mathbf{m}_s^{(i-1)}), \tag{3.17}$$

where $\mathbf{m}_s^{(0)}$ is equal to textual embeddings $\mathbf{e}_s$, and $\mathbf{m}_s^{(N)}$ is the final textual features.

Subsequently, both textual features $\mathbf{m}_s$ and entity embeddings $\mathbf{m}_w = \left\{\tilde{w}_1, \ldots, \tilde{w}_{l_w}\right\}$ are fed into the K-encoder. These features are first aligned in sequence according to the interaction matrix $\mathbf{v}$. The K-encoder is structured as a series of stacked aggregators, each designed to process and integrate diverse information. These aggregators aim to encode both tokens and entities, while simultaneously fusing their heterogeneous features. This architecture allows the K-encoder to create a unified representation that captures the rich interplay between linguistic elements (tokens) and semantic concepts (entities). First, the token embeddings and entity embeddings are processed by two multi-head self-attention layers (denoted as $\mathrm{Att}(\cdot)$) in aggregators. The outputs of the i-th aggregator are denoted as

$$\begin{aligned} \tilde{\mathbf{m}}_s^{(i)} &= \mathrm{Att}^{(i)}(\tilde{\mathbf{m}}_s^{(i-1)}), \\ \tilde{\mathbf{m}}_w^{(i)} &= \mathrm{Att}^{(i)}(\tilde{\mathbf{m}}_w^{(i-1)}), \end{aligned} \tag{3.18}$$

where $\tilde{\mathbf{m}}_s^{(0)}$ is equal to $\mathbf{m}_s$ and $\tilde{\mathbf{m}}_w^{(0)}$ is equal to $\mathbf{m}_w$.

Then, each aggregator employs an information fusion layer to integrate token and entity information mutually. However, for tokens without aligned entities, the information fusion layer calculates their output embeddings through a simplified process without the integration step. Let $\mathbf{m}$ represent the output of the K-encoder, the computation proceeds as follows:

$$\begin{aligned} \mathbf{h}^{(i)} &= \sigma\left(\tilde{\mathbf{W}}_s^{(i)}\tilde{\mathbf{m}}_s^{(i)} + \tilde{\mathbf{b}}^{(i)}\right), \\ \mathbf{m}^{(i)} &= \sigma\left(\mathbf{W}_s^{(i)}\mathbf{h}^{(i)} + \mathbf{b}_s^{(i)}\right). \end{aligned} \tag{3.19}$$

Here, $\mathbf{h}^{(i)}$ is the intermediate hidden state which integrates token and knowledge information. $\mathbf{W}_s^{(i)}$ and $\mathbf{b}_s^{(i)}$ are the weights and bias for token embeddings, respectively. $\sigma(\cdot)$ is a non-linear activation function named Gaussian error linear unit (GELU) [173].

For tokens with aligned entities, the information fusion layer computes updated embed-

dings for each token and its matched entity. To fuse token features $\tilde{\mathbf{m}}_s$ and their aligned entity $\tilde{\mathbf{m}}_w$, we employ a denoising entity auto-encoder (dEA) [174]. The k-th aligned entity distribution for the i-th token is defined as

$$q\left(\tilde{w}_k \mid \tilde{s}_i\right) = \frac{\exp\left(\text{linear}\left(\tilde{s}_i\right) \cdot \tilde{w}_k\right)}{\sum_{j=1}^{l_w} \exp\left(\text{linear}\left(\tilde{s}_i\right) \cdot \tilde{w}_j\right)}, \tag{3.20}$$

where linear$(\cdot)$ represent a linear layer. Here, the number of aligned entities may differ from the number of the input entities because some entities, particularly those represented by multi-token phrases, need to be decomposed. Each component of a multi-token entity is individually aligned to its corresponding token in the input sequence. The fusion process is expressed as

$$\begin{aligned}
\mathbf{h}^{(i)} &= \sigma\left(\tilde{\mathbf{W}}_s^{(i)} \tilde{\mathbf{m}}_s^{(i)} + \tilde{\mathbf{W}}_e^{(i)} \tilde{\mathbf{m}}_w^{(i)} + \tilde{\mathbf{b}}^{(i)}\right), \\
\mathbf{m}_s^{(i)} &= \sigma\left(\mathbf{W}_s^{(i)} \mathbf{h}^{(i)} + \mathbf{b}_s^{(i)}\right), \\
\mathbf{m}_w^{(i)} &= \sigma\left(\mathbf{W}_w^{(i)} \mathbf{h}^{(i)} + \mathbf{b}_w^{(i)}\right), \\
\mathbf{m}^{(i)} &= \mathbf{m}_s^{(i)} + \mathbf{m}_w^{(i)},
\end{aligned} \tag{3.21}$$

where $\mathbf{W}_w^{(i)}$ and $\mathbf{b}_w^{(i)}$ are the weights and bias on the entity embeddings, respectively. The semantic representations $\mathbf{m}^{(12-N)}$ are the final output of the semantic representation network, which is written as $\mathbf{m}$ in the following.

3.2.2 Conventional Wireless Transmission Module

The semantic representations are then processed through a conventional wireless transmission module to emulate the data transmission process of the transceiver. Initially, these representations pass through the semantic compression network as shown in Step 6 in Algorithm 1. The semantic compression network consists of dense layers, pooling layers, and linear layers. The dense and pooling layers compress the semantic representations, while the linear layers transform these compressed representations into binary symbols $\mathbf{x}$. These symbols are subsequently transmitted over physical channels corresponding to Step 7 in Algorithm 1. At the receiver, the received symbols, denoted as $\mathbf{y}$, are fed into the semantic recovery network that is structurally symmetrical to the semantic compression network, which is formulated in Step 9-10 in Algorithm 1. Through this process, the original semantic representations are restored at the receiver end.

3.2.3 Data Reconstruction

The recovered semantic representations $\hat{\mathbf{m}}$ are then processed through the text reconstruction network. This network is expressed as in Steps 11-13 in Algorithm 1. Due to the additive noise and interference in physical channels, transmission errors may occur. The decoder employs a dual error correction method. When token errors arise, the textual decoder utilizes contextual information to perform reasoning-based error correction. Similarly, when entity errors occur, the knowledge decoder draws upon both contextual information and knowledge stored in the KG to implement appropriate corrections. Particularly, the text reconstruction network is similar to the semantic representation network, which comprises multiple Transformer layers that utilize multi-head self-attention mechanisms. Initially, the distorted semantic representations are processed through these Transformer layers, which analyze intact tokens within the context to predict the distorted ones. Subsequently, for each distorted token position, the text reconstruction network generates a probability distribution over the entire vocabulary. This distribution is derived from the output of the final Transformer layer, denoted as $\mathbf{h}^t$. The output is then passed through a linear projection layer followed by a softmax function $\text{softmax}(\cdot)$, converting raw scores into probabilities which can be computed as

$$p(\hat{s}_i^t \mid \hat{m}_i^t) = \text{softmax}(W_i^t h_i^t + b_i^t), \tag{3.22}$$

where $\hat{s}_i^t$ and $\hat{m}_i$ represent the i-th predicted token and the semantic representations of the token that do not have aligned entities, respectively. W_i^t and b_i^t are learnable parameters. These parameters are typically initialized randomly and then updated iteratively through stochastic gradient descent algorithms. The optimal values for W_i^t and b_i^t are not predefined but learned from the training data to minimize the CE value.

Named tokens are predicted with higher accuracy compared to other tokens, as they receive more "attention" focused on related named tokens. This is achieved through a modified attention mechanism that incorporates knowledge from the KG. When predicting a named token, the data reconstruction network considers:

- Contextual tokens from both directions (left and right) through the standard self-attention mechanism.
- Other named tokens with relationships defined in the KG. The attention mechanism is applied based on their relationships.

Let $\hat{\mathbf{m}}^e$ and $\hat{\mathbf{s}}^e$ represent the recovered semantic representations of named tokens and the reconstructed named tokens respectively. The k-th predicted named token combining both the contextual information from Transformer layers and the knowledge from $\mathcal{G}$ is calculated

as

$$p(\hat{s}_k^e|\hat{m}_k^e, \mathcal{G}) = softmax(W_k^t h_k^t + W_k^e h_k^e + b_k), \tag{3.23}$$

where h_j^e is the output of the Transformer layer for the k-th named token. W_k^e, and b_k are learnable parameters. Hence, the i-th token and the k-th named token with the highest probability are calculated as:

$$\begin{aligned} \hat{s}_i^t &= \arg\max_w p(s_i^t|\hat{m}_i^t), \\ \hat{s}_k^e &= \arg\max_w p(s_k^e|\hat{m}_k^e, \mathcal{G}). \end{aligned} \tag{3.24}$$

The final reconstructed tokens $\hat{\mathbf{s}} = \hat{\mathbf{s}}^t \oplus \hat{\mathbf{s}}^e$ are created by integrating predicted tokens based on $\mathbf{v}$. The i-th reconstructed token is computed as

$$\hat{s}_i = \begin{cases} \hat{s}_k^e, \text{ if } v_i = 1, \\ \hat{s}_i^t, \text{ if } v_i = 0. \end{cases} \tag{3.25}$$

The loss function of KG-SemCom is provided as

$$\mathcal{L} = \mathcal{L}_T + \mathcal{L}_N + \mathcal{L}_E, \tag{3.26}$$

where $\mathcal{L}_T$, $\mathcal{L}_N$ and $\mathcal{L}_E$ denote the loss of predicted tokens, NSP and predicted entities, respectively. First, let $p(\tilde{s}_i)$ represent the probability distribution of the i-th predicted token and $q(s_i)$ represent the real probability distribution of the i-th original token. The CE (*i.e.,* loss) for predicted tokens, i.e.,

$$\mathcal{L}_T = -\sum_{i=1}^{l_s} \left[q(s_i) \log p(\tilde{s}_i) + (1 - q(s_i)) \log(1 - p(\tilde{s}_i)) \right]. \tag{3.27}$$

Next, the CE for NSP is computed as

$$\mathcal{L}_N = -\sum_{l=1}^{B} \left[q_l \log(p_l) + p_l \log(q_l) \right], \tag{3.28}$$

where p_l is the predicted probability of the negative class (i.e., the l-th sentence is not adjacent), q_l the probability of the positive class (i.e., the l-th sentence is adjacent), and B is the number of sentences need to be transmitted. Subsequently, the CE calculation of dEA using

the aligned entity distribution shown in (3.20) is

$$\begin{aligned}\mathcal{L}_E = -\sum_{k=1}^{l_w} [&q(w_k|s_i, \mathcal{G}) \log(p(\tilde{w}_k|s_i, \mathcal{G})) \\ &+ (1 - q(w_k|s_i, \mathcal{G})) \log(1 - p(\tilde{w}_k|s_i, \mathcal{G}))].\end{aligned} \tag{3.29}$$

The KG-SemCom network, through iterative loss minimization, learns to comprehend language at multiple levels including the syntactic structure of sentences, the semantic meanings of words, and their contextual and knowledge-based relationships.

3.3 Simulation results and Discussion

In this section, we conduct simulations to evaluate the performance of the proposed KG-SemCom framework compared with two benchmarks: 1) Deep-SC [2], a DL-based SemCom scheme that uses Transformer encoder-decoder architecture without KG integration; 2) a traditional communication scheme combining Huffman coding [175] for source coding and 1/2 and 2/3-rate low-density parity-check (LDPC) codes [176] for channel coding, denoted as Huffman + LDPC throughout this paper, which relies purely on mathematical algorithms without ML techniques. Additionally, the simulations are performed under the AWGN channels and Rayleigh fading channels with varying SNRs from -3 dB to 24 dB.

3.3.1 Dataset and Simulation Settings

Our pre-training procedure for KG-SemCom follows established practices in language model pre-training, with some adaptations. To mitigate the substantial computational cost of training from scratch, we initialize the BERT-base-uncased model (https://huggingface.co/google-bert/bert-base-uncased). that includes Transformer blocks for token encoding. We use English Wikipedia (https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2) as our primary corpus, aligning text with Wikidata entities. We also use a subset of TACRED [177] test dataset, consisting of 15509 sentences, as an additional dataset to test our model. In addition, we select sentences with the length of 15-35 words and symbols, as well as exclude those with fewer than three entities. For entity representation, knowledge embedding is pre-trained on Wikidata using the TransE [178] method. Specifically, we sample a subset of Wikidata, using 21,841,016 triplets involving 4,536,887 entities for pre-training. Additionally, we leverage 2,426,780 triplets with 504,099 entities for evaluation purposes.

In our model, we employ distinct parameters for T-encoder and K-encoder. The token-related components use a hidden dimension of 768 and 12 self-attention heads, while the

Table 3.2: The comparison of the reconstructed sentences from KG-SemCom, DeepSC and Huffman + LDPC schemes (SNR = −3 dB).

Received Sentence	True Value	KG-SemCom's Results	DeepSC's Results	Huffman + LDPC's Results
The ladder __ attracts young honeybees with its scent, derived from a variety of aromatic compounds.	orchid	orchid	flower	alopk#
Travelers on the newtown route between __ (which was then only the area known today as downtown city) and Yorktown passed through the area .	Philadelphia	Philadelphia	Washington	phs5iuwnbsah
__ was appointed as chief executive at Liverpool in July 1998 by then chairman David Moores.	Rich Parry	Rich Parry	Richard Parry	sahh P;isq

entity-related components use a hidden dimension of 100 and 4 self-attention heads. Both T-encoder and K-encoder utilize 6 encoding layers. The parameters of the data reconstruction network are the same as the semantic representation network. These configuration results in a total of approximately 114 million parameters. Particularly, BERT-based model has about 110 million parameters, indicating that KG-SemCom's additional knowledge module is relatively small and has minimal impact on runtime performance. Our pre-training largely follows BERT's hyperparameters, with the exception of setting the learning rate to 5×10^{-5}. All simulations are implemented in a computer with an NVIDIA 3090Ti GPU and an Inter Core i9-12900 processor, where the main software environment is Python 3.9.

3.3.2 Performance Metrics

Conventional performance metrics such as BER may not adequately reflect a system's language understanding capability in SemCom models. In these models, a high BER does not necessarily indicate poor performance, because the semantic content may still be accurately conveyed despite bit-level errors. To better measure the network performance, we use two performance metrics in this paper, i.e., the named token similarity [179] and the BERT-based sentence similarity [180].

First, we use the named token similarity to measure the semantic similarity between predicted tokens and original named tokens in the KG. This metric focuses on comparing concepts provided by the KG in which similar concepts being more closely connected. In this way, synonyms are typically grouped together, while polysemous words with different meanings are placed farther apart, even if they are written identically. Specifically, this metric combines path length and information content (IC) to measure semantic similarity between concepts by using a weighted path length approach. It employs the IC of the least common subsumer (LCS) [181] of two concepts to weight their shortest path length. Consequently, concept pairs with identical path lengths can have different semantic similarity scores if their LCSs differ in IC. Thus, the named token similarity between a named token s_e and a predicted

token $\hat{s}_e$ is calculated as

$$\mathrm{sim_t}(\hat{s}_e, s_e) = \frac{1}{1 + length(\hat{s}_e, s_e) \times k^{IC(c_{lcs})}}, \tag{3.30}$$

where $k \in (0, 1]$ represents the influence of IC on the shortest path length. When $k = 1$, IC has no effect, while lower values of k increase the contribution of the LCS's IC, which represents the shared information between two concepts. The range of this named token similarity score is $(0, 1]$. In this means, the path length represents the difference between concepts, while common information represents their commonality. For identical concepts, the path length is 0, resulting in maximum semantic similarity (a score of 1). As the path length between concepts in the taxonomy increases, their semantic similarity decreases. Compared to conventional word-based similarity metrics, this concept-based approach leverages the knowledge embedded in the KG to measure predictions more accurately. It is particularly well-suited for our KG-SemCom model, as it captures semantic relationships more effectively than surface-level word comparisons.

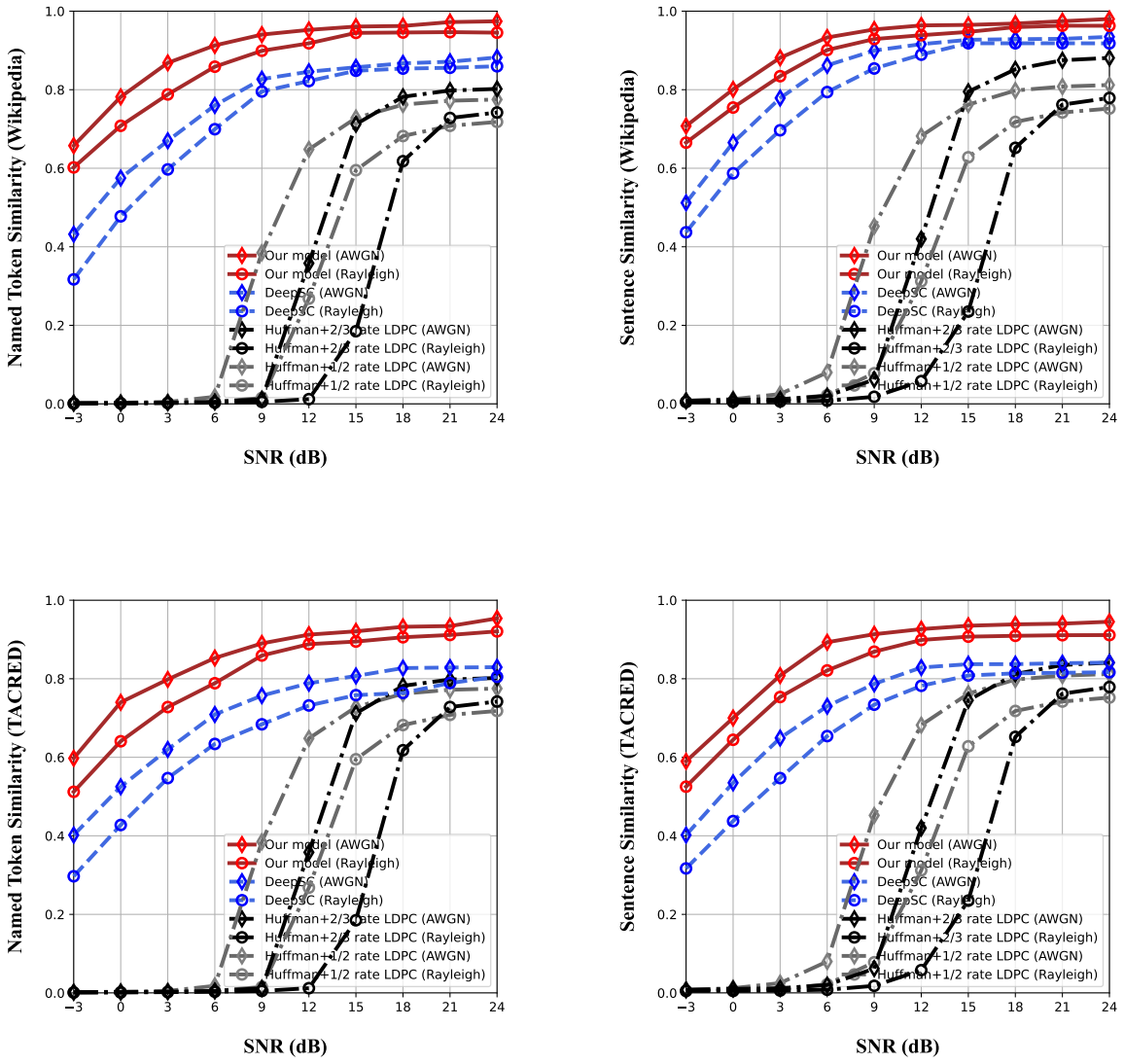
Next, we are going to measure the semantic similarity between original and predicted sentences. Previous SemCom systems often use bilingual evaluation understudy (BLEU) score to measure the predicted sentence. However, BLEU's limitation lies in its focus on word-level comparisons, which fails to capture the deeper semantic information conveyed in the sentences. Thus, we calculate the BERT-based sentence similarity between the source tokens $\mathbf{s}$ and the reconstructed tokens $\hat{\mathbf{s}}$, which can be formulated as

$$\mathrm{sim_s}(\hat{\mathbf{s}}, \mathbf{s}) = \frac{\mathbf{B}_\Phi(\mathbf{s}) \cdot \mathbf{B}_\Phi(\hat{\mathbf{s}})}{\| \mathbf{B}_\Phi(\mathbf{s}) \| \| \mathbf{B}_\Phi(\hat{\mathbf{s}}) \|}, \tag{3.31}$$

where $\mathbf{B}_\Phi(\mathbf{s})$ and $\mathbf{B}_\Phi(\hat{\mathbf{s}})$ represents the embeddings of $\mathbf{s}$ and $\hat{\mathbf{s}}$ obtained from the BERT model. The BERT-based sentence similarity is limited between 0 to 1, with a higher score indicating better performance. Unlike BLEU and other word-based approaches [182], this metric can capture deep contextual information and sentence-level semantics, since BERT allows it to better understand nuanced meanings and handle polysemy.

3.3.3 Simulation Results

First, Table 3.2 compares the performance of KG-SemCom, DeepSC and Huffman + LDPC schemes in predicting missing words or phrases in three examples at an SNR of −3 dB, alongside the true values. The predicted results reflect that KG-SemCom can accurately predict missing named tokens in diverse contexts, outperforming the other two schemes in these examples. Particularly, Huffman + LDPC demonstrates the poorest performance since it op-

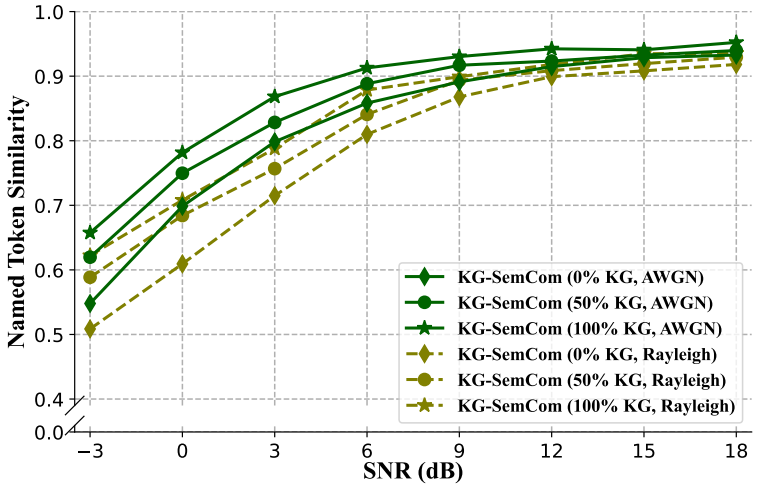


(a) (b)

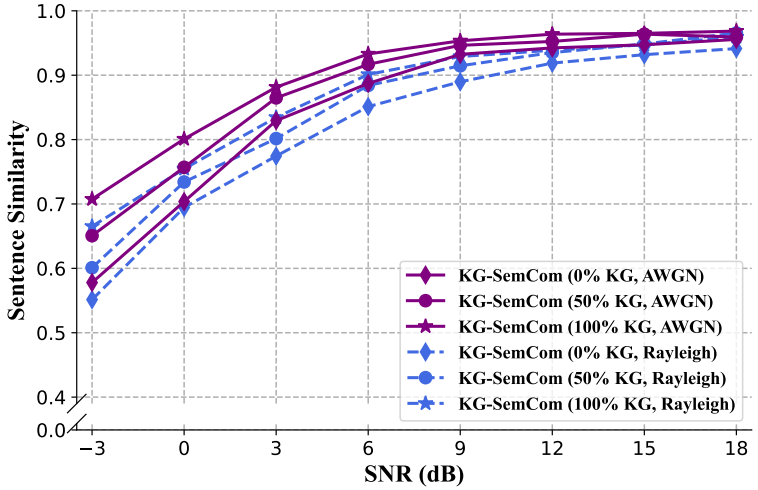
Figure 3.4: (a) Named token similarity and (b) BERT-based sentence similarity of predictions utilizing Wikipedia (top) and TACRED (bottom) over AWGN channels and Rayleigh fading channels versus varying SNRs.

erates exclusively at the symbol level without comprehending content meaning. When noise corrupts symbols, the message can not be recovered accurately as this method lacks reasoning capabilities to reconstruct distorted data. DeepSC shows improved performance but predict incorrect entities in poor communication environment because it reasons solely based on contextual information without leveraging external knowledge. In contrast, KG-SemCom incorporates both contextual information and knowledge provided by the KG, resulting in superior prediction accuracy, especially for entities.

Figures 3.4(a) and 3.4(b) illustrate the performance of named token similarity and BERT-



(a) Named Token Similarity



(b) Sentence Similarity

Figure 3.5: Named token similarity (a) and sentence similarity (b) of recovered text with 0%, 50% and 100% KGs versus varying SNRs.

based sentence similarity, respectively, for Wikipedia (top) and TACRED (bottom). The results compare three schemes trained on Wikipedia across varying SNRs from -3 to 18 dB under both AWGN and Rayleigh channel conditions. We find that for all schemes achieve better the named token similarity and the BERT-based sentence similarity in AWGN channels compared with that in Rayleigh fading channels under all SNR conditions. This disparity is expected due to the additional challenges posed by multipath fading in Rayleigh channels, including fast signal strength fluctuations, time-varying distortions, and deep fades. Unlike the uniform and predictable nature of AWGN, Rayleigh fading introduces complex amplitude and phase variations that are more difficult for NLP-based models to adapt and compensate. For both Wikipedia and TACRED, KG-SemCom consistently outperforms other

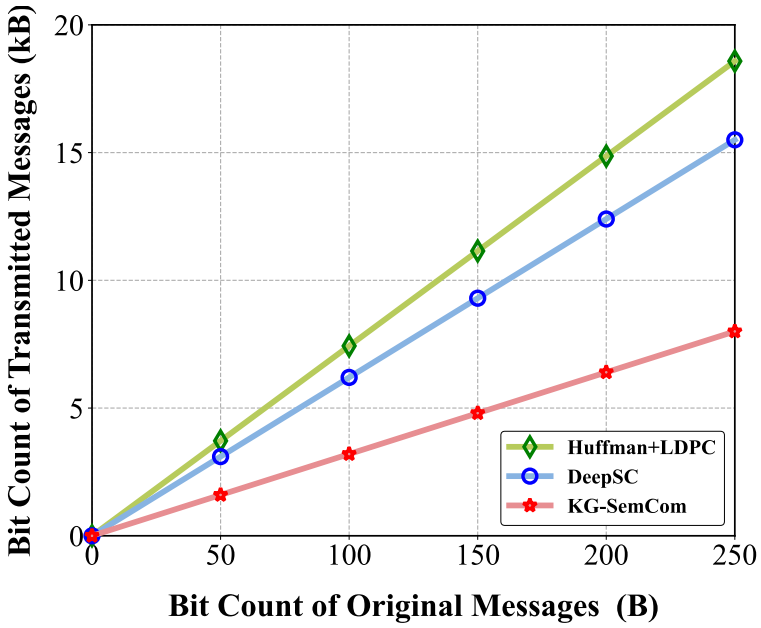


Figure 3.6: The bit counts of transmitted messages generated by three schemes versus original messages.

schemes across the entire SNR range in both channel conditions, particularly when SNR is lower than 12 dB. Moreover, the performance difference between Wikipedia and TACRED is marginal since we leverage the same KG, i.e., Wikidata, to process these datasets. These results demonstrates KG-SemCom's superior ability and generalization to leverage KG information for robust and reliable SemCom.

Comparing the two similarity metrics, the named token similarity focuses on specific and important tokens (e.g., names, key terms), while the BERT-based sentence similarity considers overall meaning. Consequently, the BERT-based sentence similarity demonstrates higher overall scores and appears more robust to small errors, as the overall meaning can be preserved even if some words are incorrect. This difference in focus explains why the performance gap between KG-SemCom and other methods is more pronounced in the named token similarity compared to the BERT-based sentence similarity.

Next, Fig. 3.6 compares the bit counts of transmitted messages generated by three schemes versus original messages. Specifically, Huffman + LDPC requires the highest number of transmitted bits since it uses conventional coding methods. DeepSC reduces the transmitted bits, leveraging semantic understanding to improve communication efficiency. As expected, KG-SemCom achieves the best performance, transmitting the fewest bits for the same message. Despite incorporating KG information, we design a complex encoder network consisting of multiple Transformer and dense layers to generate optimized transmission data. Consequently, KG-SemCom demonstrates superior data compression capability.

Table 3.3: The training details of KG-SemCom, DeepSC and Huffman+LDPC.

	KG-SemCom	DeepSC	Huffman+LDPC
Number of trainable parameters	115754812	10582762	0
Optimizer	BERTAdam	Adam	None
Batch size	32	128	None
Number of epochs	283854	80	None

To further examine the effectiveness of KG, we compare three levels of KG utilization in KG-SemCom, as shown in Fig. 3.5(a) and Fig. 3.5(b). The KG utilization is categorized into 0%, 50%, and 100% three groups, reflecting the proportion of triplets used in knowledge fusion. Fig. 3.5(a) shows the named token similarity of different KG utilization levels under AWGN channels and Rayleigh fading channels. It can be seen that higher KG utilization levels lead to improved text recovery. The performance generally improves as SNR increases, with the gap between KG utilization levels narrowing at higher SNRs. In turn, Fig. 3.5(b) shows the BERT-based sentence similarity of different KG utilization levels. The sentence similarity scores tend to be higher than named token similarity scores. Although both metrics show improvement with higher KG utilization levels, the gap between 0%, 50%, and 100% KG is generally wider for named token similarity than for the sentence similarity. This suggests that KG integration may have a particularly strong impact on preserving specific named entities or key tokens in the recovered text, even when overall sentence semantics are challenged by poor channel conditions. These two figures illustrate the trade-off between KG utilization and model accuracy. The size of KG can be adjusted to accommodate various devices, particularly those with resource constraints, though this adaptation comes at the cost of reduced accuracy.

Finally, we list the number of trainable parameters, optimizers, batch size and number of epochs of KG-SemCom, DeepSC, and Huffman + LDPC coding schemes in Table 3.3. Notably, KG-SemCom employs twelve Transformer layers in its encoder/decoder structure, resulting in the highest number of trainable parameters. With six Transformer layers in its encoder-decoder architecture, DeepSC contains fewer trainable parameters. The Huffman + LDPC, implementing purely mathematical algorithms, contains no trainable parameters. We also compare the average runtime (including processing and transmission) per sentence of three schemes in Fig. 3.7. KG-SemCom requires the highest computational runtime, followed by DeepSC, while the Huffman + LDPC achieves the fastest execution. Moreover, KG-SemCom is evaluated with three variations based on KG utilization (0%, 50%, and 100%), showing a progressive increase in runtime from 1.216 s to 1.662 s to 1.998 s as more of the KG is utilized. The increased computational cost is due to the complexities

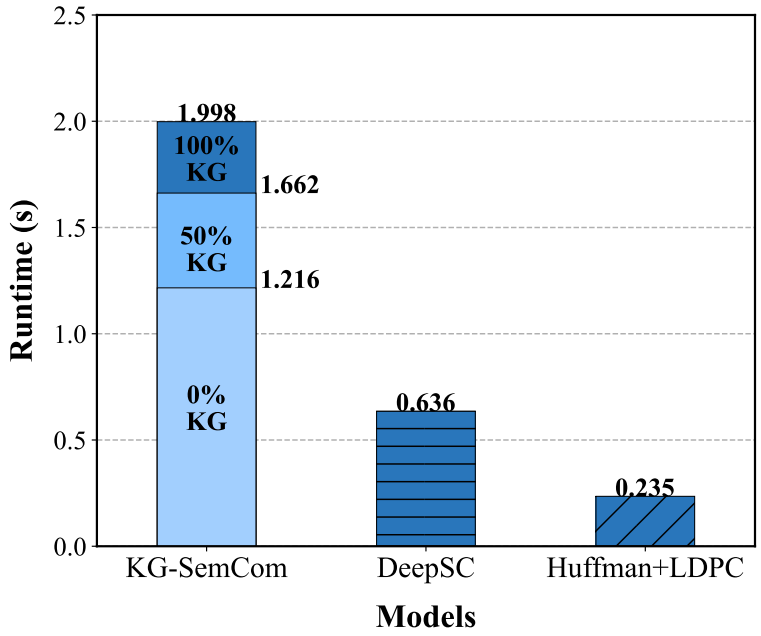


Figure 3.7: The comparison of runtimes between KG-SemCom, DeepSC, and Huffman+LDPC coding schemes.

introduced by our knowledge alignment and fusing algorithm, as well as the knowledge- and context-based reasoning process, both of which scale with the size of the KG. Although our framework requires more computational time and resources compared to the baselines, it demonstrates significant benefits when operating at lower SNR levels. This performance-complexity trade-off highlights the importance of selecting communication systems based on specific operational requirements and environmental constraints. However, it is important to note that these runtime values are based on our own laptop, which has limited computational resources compared to practical communication infrastructure. In real-world applications, the hardware is expected to be much more powerful, which can significantly reduce the runtime cost.

Chapter 4

KG-based SemCom framework for Video Delivery

This chapter presents the architecture of an SLG-based SemCom system, named VISTA, and describes the corresponding challenges addressed. The implementation details are subsequently provided, including the design of key modules and their operational mechanisms. Evaluation results are then presented to demonstrate the performance and effectiveness of the proposed model.

4.1 Video Transmission Framework in VISTA

In this section, we delve into the video transmission framework underpinning VISTA.

We consider a source video composed of T sequential frames: $\mathbf{s} = \left\{s^1, \ldots, s^T\right\} \in \mathbb{R}^{H \times W \times T}$, where H and W respectively denote the height and width of a frame. These frames are first fed in the convolutional semantic-encoder to distill the textual semantic information $\mathbf{g}$. In addition, the semantic-encoder divides the source video into two parts: environment (static background) $\mathbf{s}_e$ and behavior segments (dynamic objects) $\mathbf{s}_b$ individually. Thus, the encoded frames can be written as $\hat{\mathbf{s}} = \{\mathbf{s}_e, \mathbf{s}_b, \mathbf{g}\}$ under the semantic-encoder network $\mathcal{S}(\cdot)$ with parameter set α_s, i.e.,

$$\hat{\mathbf{s}} = \{\mathbf{s}_e, \mathbf{s}_b, \mathbf{g}\} = \mathcal{S}\left(\mathbf{s}; \alpha_s\right). \tag{4.1}$$

The encoded frames $\hat{\mathbf{s}}$ then flow into JSCC module for SNR-adaptive wireless transmission. In this module, source-encoder $\mathcal{E}$ and channel-encoder $\mathcal{C}$ with parameter sets α_ϵ and α_c generate the symbols $\mathbf{x}$ to be transmitted,

$$\mathbf{x} = \mathcal{C}\left(\mathcal{E}\left(\hat{\mathbf{s}}; \alpha_\epsilon\right); \alpha_c\right). \tag{4.2}$$

At the receiver side, $\mathbf{y}$ is denoted as the received symbols for the input $\mathbf{x}$ over the wireless channel with additive noise w, i.e.,

$$\mathbf{y} = h * \mathbf{x} + w, \tag{4.3}$$

where h denotes the channel gain. $\mathbf{y}$ is then fed to the channel-decoder $\mathcal{C}^{-1}$ and source-decoder $\mathcal{E}^{-1}$ sequentially to reconstruct the environment $\tilde{\mathbf{s}}_e$ and behavior segments $\tilde{\mathbf{s}}_b$ with the help of the semantics. The decoded frames $\tilde{\mathbf{x}}$ is presented as

$$\tilde{\mathbf{x}} = \{\tilde{\mathbf{s}}_e, \tilde{\mathbf{s}}_b\} = \mathcal{E}^{-1}\left(\mathcal{C}^{-1}\left(\mathbf{y}; \beta_c\right); \beta_\epsilon\right). \tag{4.4}$$

where β_c and β_ϵ denote the parameters of channel-decoder and source-decoder networks, respectively.

Finally, the recovered video $\tilde{\mathbf{s}}$ should be constructed as per the two parts of segments $\tilde{\mathbf{s}}_e$ and $\tilde{\mathbf{s}}_b$. The semantic-decoder network and its parameters are given as $\mathcal{S}^{-1}$ and β_s. Thus, the final recovered video is expressed as

$$\tilde{\mathbf{s}} = \mathcal{S}^{-1}\left(\tilde{\mathbf{x}}; \beta_s\right). \tag{4.5}$$

In this work, the ultimate goal is minimizing the semantic ambiguity of the recovered video. We use average peak signal-to-noise ratio (PSNR) ([183]), a popular video quality metric, to measure the differences between the recovered and original video frames. In detail, for the t-th frame with the size $m \times n$, the mean squared error (MSE) between the original frame s^t and the recovered one $\tilde{s}^t$ is calculated as

$$MSE^t = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[s^t(i,j) - \tilde{s}^t(i,j)\right]. \tag{4.6}$$

Thus, the average of PSNR of the original and recovered video is expressed as

$$PSNR = \frac{1}{T} \sum_{t=1}^{T} 10 \cdot log_{10}\left(\frac{{I_{max}}^2}{MSE^t}\right), \tag{4.7}$$

where ${I_{max}}^2$ represents the maximum pixel value of the frame.

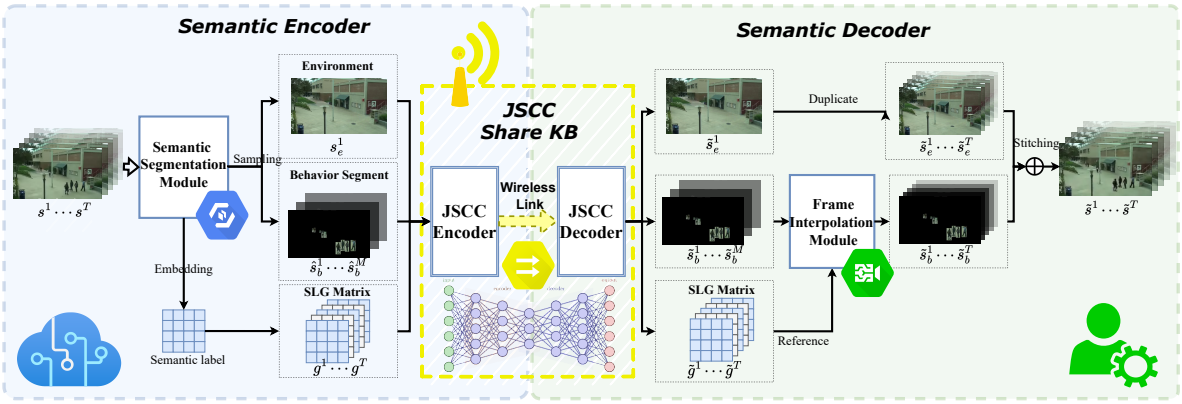


Figure 4.1: The diagram of transceiver in VISTA.

4.2 SLG-based Transceiver Design in VISTA

Building on the previously described video transmission framework, we now focus on the design of the transceiver in the VISTA system. As shown in Fig. 4.1, VISTA consists of three key modules: semantic segmentation, JSCC, and frame interpolation. The semantic segmentation and JSCC encoder modules are located at the transmitter side, while the receiver side features the frame interpolation module along with the JSCC decoder. Importantly, SLGs are created during the semantic segmentation process and later utilized in frame interpolation. These modules are individually trained with distinct loss functions, aiming to minimize the PSNR for the restored video. Next, we will explore the construction and training of the DL networks in these three modules.

4.2.1 Semantic Segmentation Module

Semantic segmentation module is deployed at the transmitter to recognize and distill the dynamic objects from video. Four tasks should be performed in this module, object detection, trajectory prediction, SLG construction and frame sampling. Generally, we first bound all objects using rectangular boxes in each frame and differ the dynamic objects from static background via velocity testing. The semantic information of each dynamic object is able to be extracted by the means of category recognition. However, the occlusion caused by overlapping objects in the video will affect the accuracy of position detection and semantics extraction. Thus, we predict the trajectory of each object for the continuous frames in the second task. After trajectory prediction, an SLG is designed to assist in reserving the estimated positions and semantics of all dynamic objects in each frame. Finally, we sample the frames and send them along with SLGs to JSCC module. Let us below illustrate the design of the four tasks separately.

Object detection: Borrowing the idea from [184], we apply a conventional network to

outline bounding boxes using features of the entire frame. Specifically, we initialize several bounding boxes and they are projected to enlarge and shift dynamically until all the objects are bounded with the optimized confidence scores. In this way, each bounding box is associated with six predictions: 2D-coordinates (u, v) of the center for the object, the width and height (w, h) of the box containing relative to the whole image of the object, object category l and the associated confidence score c.

Trajectory prediction: After getting these objective detections, we should guarantee that every dynamic objects can be captured completely. Therefore, we deploy the trajectory prediction module to track the dynamic objects when they are occluded. The input of the network of trajectory prediction is the images of dynamic objects and the five predictions of the corresponding bounding boxes. We first define an "observation" of a bounding box as $z = [u, v, w, h, c]^\top$. Moreover, we employ the Kalman filter (KF) to generate the state $q = [u, v, a, r, \dot{u}, \dot{v}, \dot{a}]^\top$, where a is the bounding box scale (area), r is the width-to-height ratio of the bounding box, and the other three variables ($\dot{u}$, $\dot{v}$ and $\dot{a}$) are the related time derivatives.

Next, we utilize an observation-centric tracker [185] with the object movement. Specifically, since a non-linear motion can be regarded as a synthesis of many small-scale linear motions in a reasonably short time, we calculate the velocity consistency (momentum) to gain the accurate velocity value and direction. Then, for an untracked object, an observation-centric online smoothing strategy is performed through a virtual trajectory $\hat{\mathbf{z}}^t$ starting from its last occurrence and ending at the re-associated observation, which is denoted as

$$\hat{\mathbf{z}}^t = \mathcal{T}_v\left(\mathbf{z}^{t_1}, \mathbf{z}^{t_2}, t\right), t_1 < t < t_2, \tag{4.8}$$

where $\mathbf{z}^{t_1}$ is the the last observation before being untracked, $\mathbf{z}^{t_2}$ is observation triggering the re-association, and $\mathcal{T}_v()$ represents the network of virtual trajectory. Along this virtual trajectory, the status at t_1 is recalled back to check the filter parameters. Thus, the refreshed state $\hat{q}^t$ is estimated as

$$\hat{q}^t = \mathbf{F}^t \hat{q}^{t-1} + \mathbf{K}^t\left(\hat{z}^t - \mathbf{H}^t \mathbf{F}^t \hat{q}^{t-1}\right), \tag{4.9}$$

where $\mathbf{K}^t$ denotes the KF matrix, $\mathbf{F}^t$ and $\mathbf{H}^t$ denote the state transition and observation model respectively. With the instruction of $\hat{q}^t$, for the t-th frame in the video, we update the bounding boxes of dynamic objects and use the behavior segments $\hat{s}_b^t$ to represent the images of all estimated boxes covering. Moreover, the rest of this frame is represented by the environment $\hat{s}_e^t$.

SLG construction: Aiming at locating the dynamic objects and illustrating the association between their location and semantics, we deliver an SLG to concatenate the classes and

location from the refreshed states $\hat{q}$. With respect to a frame containing B boxes, the set of object categories is $\mathbf{l} = \{l_1, \ldots, l_B\}$, and the 2D-coordinates set are $\hat{\mathbf{u}} = \{\hat{u}_1, \ldots, \hat{u}_B\}$ and $\hat{\mathbf{v}} = \{\hat{v}_1, \ldots, \hat{v}_B\}$. Thus, the SLG $g^t \in \left\{g^1, \ldots, g^T\right\}$ of the t-th frame can be represented as

$$g^t = \{\mathbf{l}, \hat{\mathbf{u}}, \hat{\mathbf{v}}\}. \tag{4.10}$$

Frame sampling: According to the results of trajectory prediction, we split the whole video into environment and behavior segments and transmit them separately. Since the environment is fixed, it is supposed that only the environment of the first frame s_e^1 needs to be transmitted. It is also thriftless for encoder to cope with behavior segments in the whole video, so that we sample them every T_s frames and denote $M = \lceil T/T_s \rceil$ samples as $\hat{\mathbf{s}}_b' = \left\{\hat{s}_b^1, \hat{s}_b^{T_s+1}, \ldots, \hat{s}_b^T\right\}$. The output $\hat{s}^t$ of the t-th frame is illustrated as

$$\hat{s}^t = \begin{cases} \left\{s_e^1, \hat{s}_b^1, g^1\right\}, & t = 1, \\ \left\{\hat{s}_b^t, g^t\right\}, & t = nT_s + 1, n = \{1, \ldots, M-1\}, \\ g^t, & otherwise. \end{cases} \tag{4.11}$$

Generally, the overall output for all frames after semantic-encoder is composed of environment of the first frame, behavior segments from the sample frames and SLGs of all frames.

4.2.2 JSCC Module

As illustrated, all the extracted semantic segments along with an SLG should be transmitted through a wireless channel. In VISTA, we employ an SNR-adaptive JSCC module which can configure its parameters depending on the SNR of the channel [186]. Its overall structure can be described as source-encoder, channel-encoder, channel-decoder and source-decoder. In more detail, the features $\mathbf{f} = \left\{f_e^1, \mathbf{f}_b\right\}$ are first extracted from the input of environment and behavior segments (s_e^1 and $\hat{\mathbf{s}}_b'$) via several conventional layers and some of them are activated to be transmitted first. After getting $\mathbf{f}$, the channel-encoder produces two groups of length-L features. The first group with the length G_s contains either active or inactive features selected by a policy network $\mathcal{P}$, while the following G_n groups are always active without selection. The selection for each input is conducted by a binary mask W_i, where can only be 0 or 1. The total number of active groups is demonstrated as $\tilde{G} = G_n + \sum_{i=1}^{G_s} W_i$. All the active features are passed through the power normalization network to generate complex-valued transmission symbols $\left\{x_e^0, \hat{\mathbf{x}}_b'\right\} \in \mathbb{C}^{G \times L/2}$ with unit average power using the first half of features as the real part and the other half as the imaginary part. Moreover, the textual SLGs $\mathbf{g} = \left\{g^1, \ldots, g^T\right\}$ are encoded to bits $\mathbf{x}_g$ and transmitted directly. In a word, the total encoded symbols are

represented by $\mathbf{x} = \left\{x_e^0, \hat{\mathbf{x}}_b', \mathbf{x}_g\right\}$.

Next, $\mathbf{y} = \left\{y_e^1, \hat{\mathbf{y}}_b', \mathbf{y}_g\right\}$ is received as $\mathbf{x}$ should be transmitted over the wireless channel model in (4.3), where y_e^1, $\hat{\mathbf{y}}_b'$ and $\mathbf{y}_g$ denote the transmitted symbols of environment, behavior segments, and SLG, respectively. Then, $\mathbf{y}$ is fed to the channel-decoder and source-decoder sequentially to reconstruct the environment $\tilde{s}_e^1$, behavior segments $\tilde{\mathbf{s}}_b'$ and SLGs $\tilde{\mathbf{g}}$. Specifically, $\tilde{s}_e^1$ and $\tilde{\mathbf{s}}_b'$ are recovered through several convolutional layers while $\tilde{\mathbf{g}}$ are decoded to text directly.

It is worth noting that the SNR value is the part of input fed to the policy network and the SNR adaptive network leveraged in channel-encoder, channel-decoder [187]. Particularly, for the SNR adaptive network, the features in one frame are first pooled averagely across diverse feature channels (different from the wireless channels) of a neural network and then concatenated with the SNR value. Next, the results are received by two multi-layer perceptrons to produce the factors for channel-wise scaling and addition. In this way, we adjust the network of transceiver in JSCC module depending on SNR value.

4.2.3 Frame Interpolation Module

After receiving the environment and behavior segments of sample frames, we complement them and combine the results to rebuild the video with the help of SLGs $\tilde{\mathbf{g}}$ in the semantic-decoder. In more detail, we make T copies of the one-frame environment and generate the sequence of environment as $\tilde{\mathbf{s}}_e = \left\{\tilde{s}_e^1, \ldots, \tilde{s}_e^1\right\}$ at first. Then, according to the behavior segments $\tilde{\mathbf{s}}_b' = \left\{\tilde{s}_b^1, \ldots, \tilde{s}_b^M\right\}$ of M sample frames, we utilize Transformer for frame interpolation with the inspiration of Video frame interpolation with Transformer (VFIformer) ([188]), aiming at predicting the behavior segments for all the remaining frames. Consider the behavior segments $\tilde{s}_b^1$ and $\tilde{s}_b^2$ of the two adjacent sample frames, and the intermediate frame is denoted as $\tilde{s}_b^t$.

A convolutional network called flow estimator is utilized to obtain the optical flows $O^{t\to1}$ and $O^{t\to2}$. Additionally, the images w_b^1 and w_b^2 are restored as per the features $\tilde{f}_i^1$ and $\tilde{f}_i^2$ which are warped by $O^{t\to1}$ and $O^{t\to2}$ respectively. Further, the semantic decoder includes Transformer blocks (TFB) and each TFB consists of convolutional layers and several Transformer layers (TFL) with Cross-Scale Window-based Attention (CSWA) network which is a state-of-the-art attention mechanism. For the i-th TFB, its output feature f_i^t is formulated as

$$f_i^t = TFB_i\left(f_{i-1}^t, \tilde{f}_i^1, \tilde{f}_i^2\right), \tag{4.12}$$

where f_{i-1}^t is the output of (i-1)-th TFB.

Then, the intermediate frame $\tilde{s}_b^t$ is generated by a soft mask H and an image residual $\Delta\tilde{s}_b^t$

(from flow errors and occlusion) in the decoder as follows:

$$\tilde{s}_b^t = H \odot w_b^1 + (1 - H) \odot w_b^2 + \Delta\tilde{s}_b^t, \tag{4.13}$$

where $\odot$ signifies the Hadamard product. It is worth noting that the interpolation is under the guidance of SLG. In other word, the prediction of behavior segments in the intermediate frames is limited in the bounding boxes provided by SLG.

In terms of model training, the loss should be evaluated from three aspects. The first is reconstruction loss, which compares the recovered behavior segments $\tilde{s}_b^t$ and its ground-truth s_{gt}^t in t-th frame as

$$\mathcal{L}_{rec} = \left\| s_{gt}^t - \tilde{s}_b^t \right\|_1 . \tag{4.14}$$

Next, the census loss [189] is robust to illumination changes, which is defined as the soft Hamming distance between census-transformed ([190]) image patches of s_{gt}^t and $\tilde{s}_b^t$. The last one is distillation loss for supervising the estimated flows explicitly,

$$\mathcal{L}_{dis} = \left\| \tilde{O}^{t\to 1} - O^{t\to 1} \right\|_1 + \left\| \tilde{O}^{t\to 2} - O^{t\to 2} \right\|_1 , \tag{4.15}$$

where $\tilde{O}^{t\to 1}$ and $\tilde{O}^{t\to 2}$ are derived from a pretrained flow estimation network presented by [191].

As a result, the total loss is presented as

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{css}\mathcal{L}_{css} + \lambda_{dis}\mathcal{L}_{dis}, \tag{4.16}$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{css}$ and $\mathcal{L}_{dis}$ correspond to the reconstruction loss, census loss and distillation loss with their weights λ_{rec}, λ_{css} and λ_{dis} respectively.

After frame interpolation, the behavior segments are estimated as the combination of the sample frames and intermediate frames. The recovered video $\tilde{\mathbf{s}}$ is the synthesis of the behavior segments $\tilde{\mathbf{s}}_b$ and the copies of the environment $\tilde{\mathbf{s}}_e$, which can be expressed by

$$\tilde{\mathbf{s}} = \tilde{\mathbf{s}}_b \oplus \tilde{\mathbf{s}}_e . \tag{4.17}$$

Herein, we stitch $\tilde{\mathbf{s}}_e$ and $\tilde{\mathbf{s}}_b$ via $\oplus$ and maintain $\tilde{\mathbf{s}}_b$ as their overlapping parts.

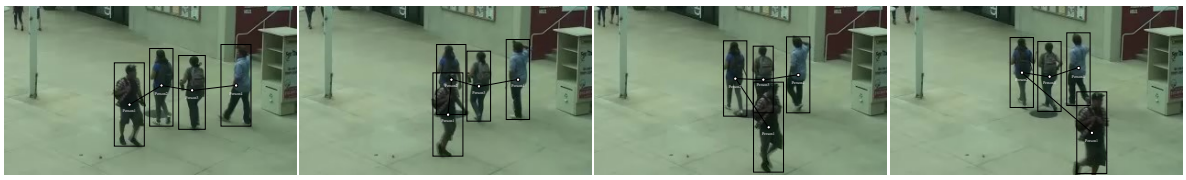


Figure 4.2: The examples of SLGs in original video.

4.3 Simulation results and discussions

4.3.1 Simulation Setting

The pre-training dataset is DanceTrack ([192]) which includes 100 videos of group dance, 105k frames and 877k high-quality bounding boxes. The test dataset is from a open dataset named VIRAT ([193]). Our team conduct simulations to evaluate the performance of the proposed VISTA framework in comparison with two different benchmarks: 1) A JSCC integrated with VFIformer scheme (JSCC-VFI), which first employs a single deep neural network to transmit video frames over wireless channels without any awareness of semantics and then uses the powerful Transformer model for behavior segments interpolation; 2) A conventional separation-based video transmission scheme employing H.265 (HEVC) [194] for source compression and LDPC codes [195] for channel error protection, which inherently suffers from the cliff effect near the channel coding threshold.

For the simulation settings, the OC-SORT structure is first leveraged for object segmentation of video frames, which keeps consistent with the setup offered by [185]. Besides, the parameters in JSCC-related channel encoding and decoding networks are proceeding as those presented by [186], where the wireless channel model is simulated as an AWGN channel with SNR values varying from -3 to 18 dB. Moreover, the architecture details of VFIformer-related networks can refer to [188]. Note that the Adam optimizer is adopted to train the VISTA with an initial learning rate of 5×10^{-4}, and all subsequent simulations are implemented in a computer with six CPU cores and Inter Core i7 processor, where the main software environment is Python 3.9.

4.3.2 VISTA Framework Performance Evaluation

Here, Fig. 4.2 shows four examples of SLGs with the labels of dynamic objects in original frames. Guided by them, the videos are recovered by VISTA under varying SNRs from -9 to 6 dB. Some examples are shown in Fig. 4.3. Obviously, VISTA can restore relatively clear frames even at an SNR of -3 dB. Since it outperforms in low-SNR scenarios, we further exhibit some specific frames of original video and the videos recovered by VISTA, JSCC-VFI

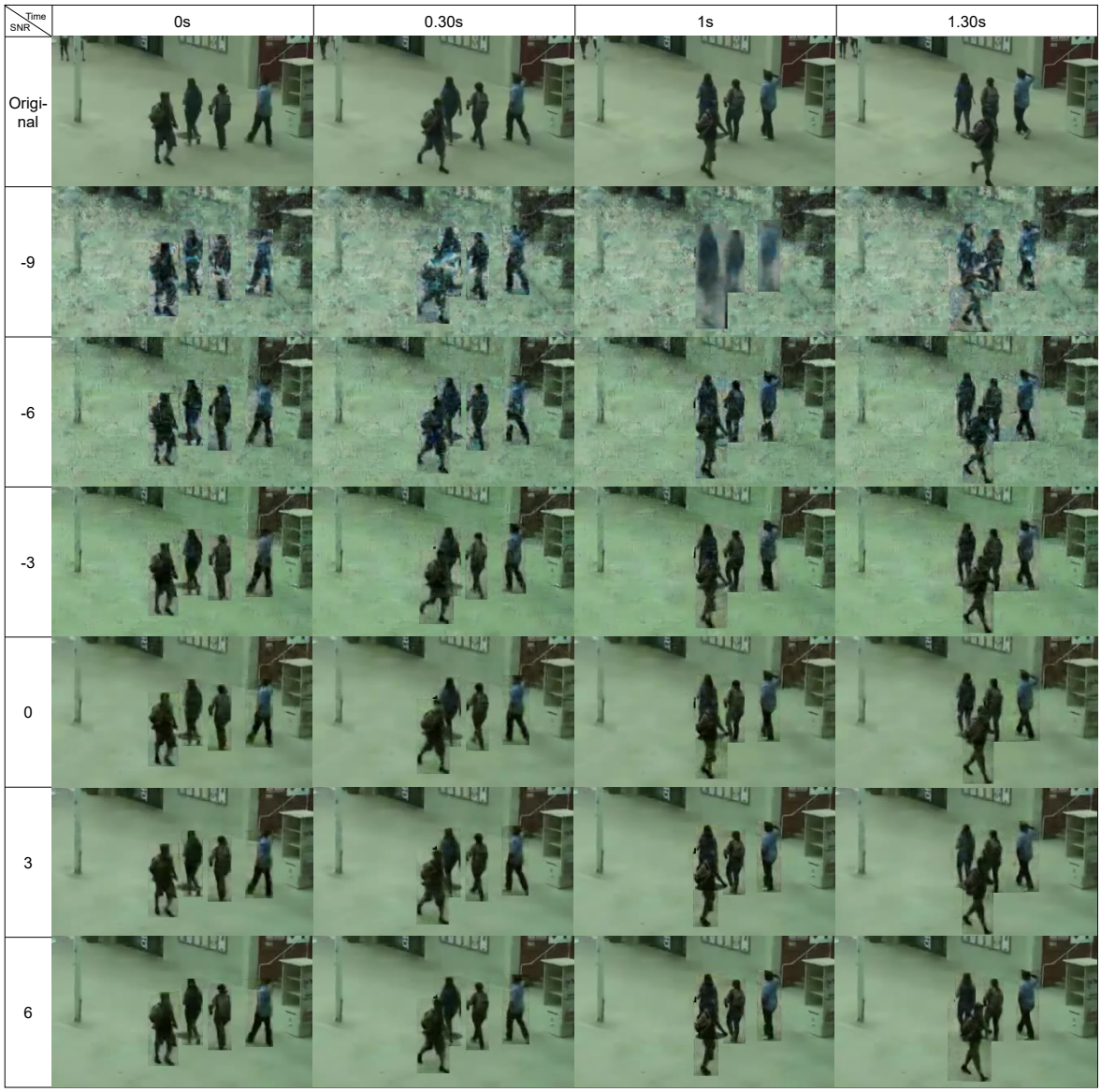


Figure 4.3: The frame samples recovered by VISTA under varying SNRs from −9 to 6 dB.

and LDPC codes at an SNR of 0 dB considering three differing interpolation proportions 0, 50%, and 75% Fig. 4.4. Importantly, it should be noted that a 50% interpolation corresponds to a 50% sampling ratio, while a 75% interpolation corresponds to a 25% sampling ratio. To clarify, a 50% sampling ratio indicates that only one out of every two sequential frames is used. All objects in the frames of VISTA and JSCC-VFI are well recovered in the contrast of the conventional scheme, and the recovered videos will be more blurred and inconsistent as the ratio of interpolation increases. Specifically, compared with JSCC-VFI, VISTA may not perfectly align moving objects with the environment. However, it offers superior frame quality and more precise detail in object rendering.

To evidence it, we present the PSNR performance under varying SNRs from −3 to 18 dB

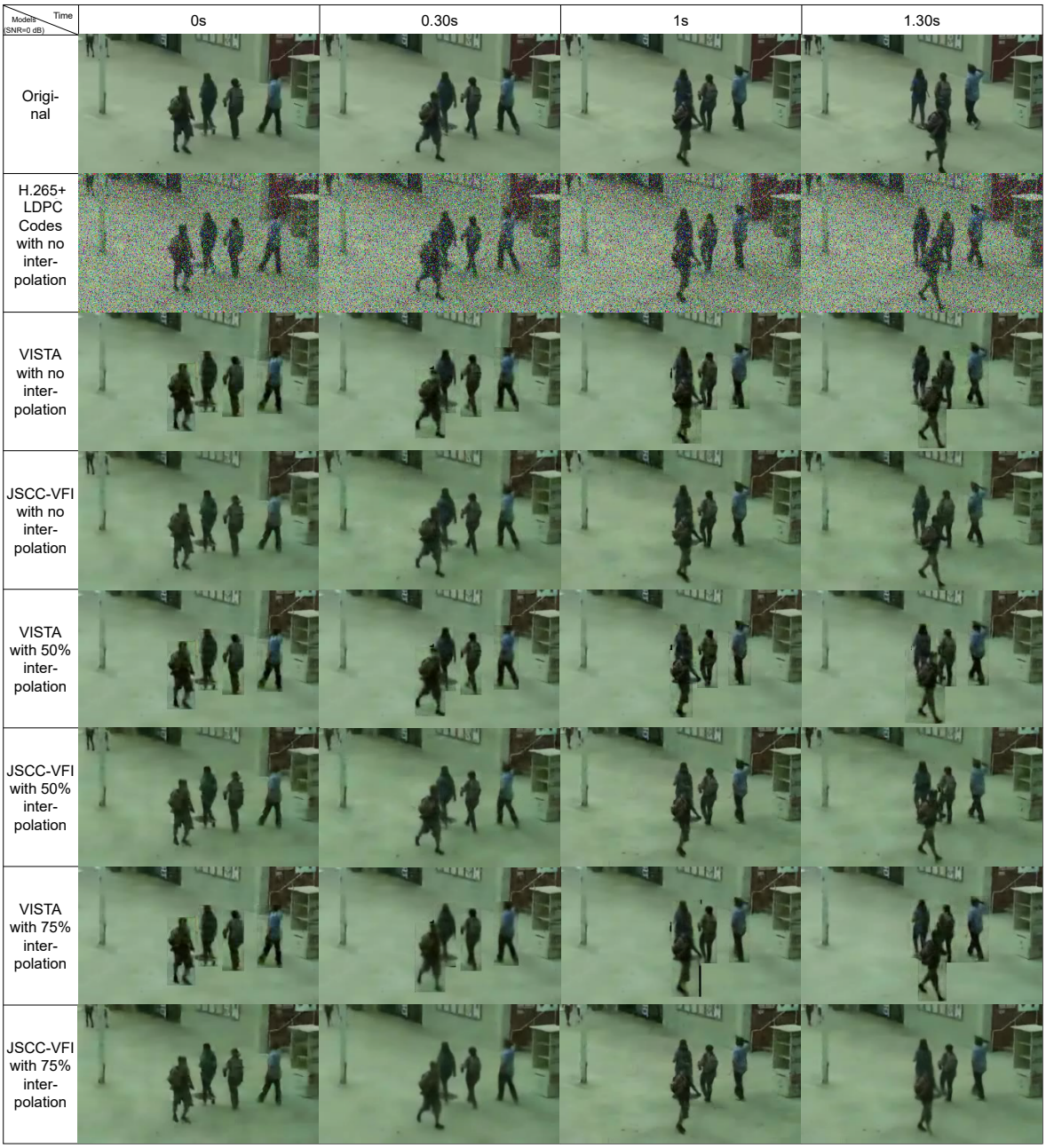


Figure 4.4: Visual comparison on frame samples for original video, recovered video by LDPC codes, VISTA and JSCC-VFI with 0%, 50%, and 75% interpolation at a SNR of 0 dB.

with all situations in Fig. 4.5. It can be seen that the PSNR of all schemes increases with SNR, which is because higher SNR leads to less impairment of transmitted semantic features, thereby enabling more accurate frame recovery. Additionally, VISTA achieves better PSNR at lower interpolation proportions. This trend is attributed to the fact that transmitting fewer behavior frames means more compressed features are lost between consecutive dynamic objects, resulting in degraded PSNR performance.

Notably, VISTA with no interpolation consistently exhibits strong performance across

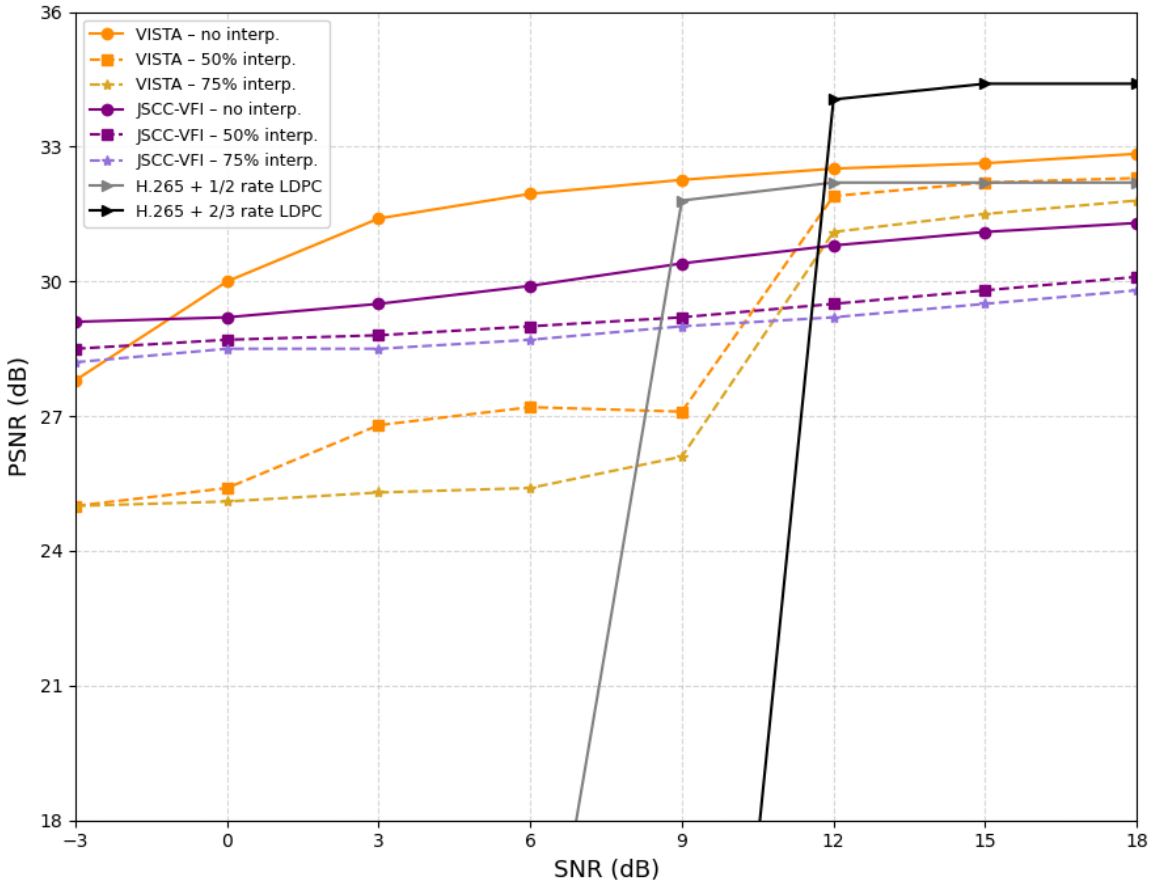


Figure 4.5: PSNR performance of recovered video frames versus varying SNRs from −3 to 18 dB.

the entire SNR range, starting at a PSNR of 28.5 dB at −3 dB and reaching 32.5 dB at 9 dB, which outperforms all other schemes at SNRs from −3 dB to 12 dB. Such a performance gain of VISTA can be credited to its accurate semantic calibration function provided by the SLG, which sufficiently guarantees high reliability of video transmission even under low-SNR conditions.

However, H.265 + 2/3 rate LDPC exceeds VISTA with no interpolation beyond SNR = 15 dB, reaching approximately 33.0 dB at SNR = 18 dB, because semantics extraction inevitably causes information loss. Consequently, the video recovered via semantic reasoning will differ from the original even at high SNR. In contrast, H.265 + LDPC conveys the complete information content of each frame, yielding superior video quality once the SNR surpasses the cliff threshold. Nonetheless, such a conventional approach suffers from catastrophic cliff-effect failure below the threshold—H.265 + 1/2 rate LDPC collapses to the noise floor below ∼9 dB, and H.265 + 2/3 rate LDPC below ∼12 dB—and necessitates a substantially higher bandwidth to transmit the full video information. Additionally, JSCC-VFi lacks the classification of environmental and dynamic objects as well as the utilization of SLGs. As a result,

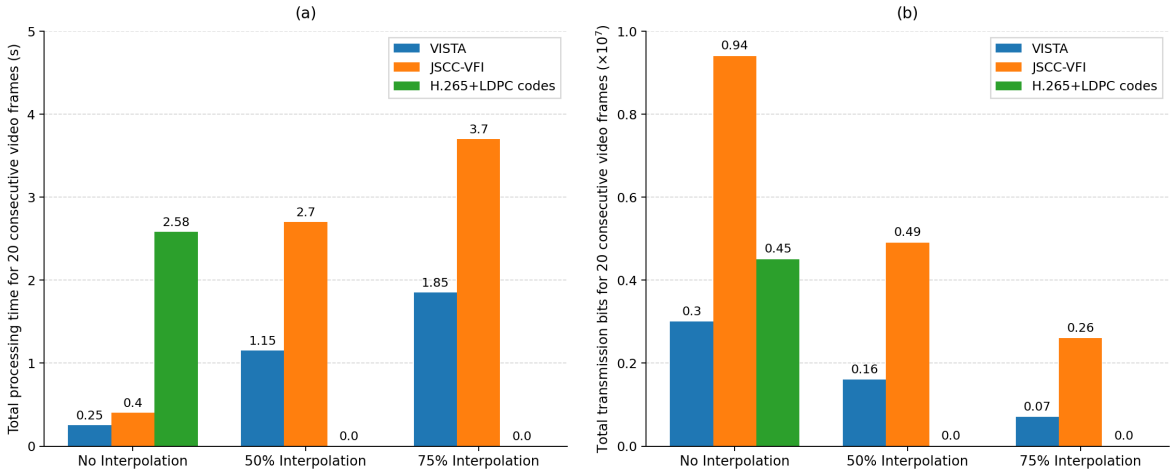


Figure 4.6: Total processing time (a) and transmission bits (b) for 20 consecutive video frames under different interpolation proportions.

its performance is inferior to VISTA across the entire SNR range from −3 dB to 18 dB.

Next, we evaluate the total processing time (including encoding, transmission, and decoding) for 20 consecutive video frames under different interpolation proportions in Fig. 4.6 (a), measured on the aforementioned hardware configuration. Note that H.265+LDPC does not involve a frame sampling process, so interpolation is not applicable; consequently, results at 50% and 75% interpolation are unavailable for this baseline. It can be observed that without interpolation, the proposed VISTA requires only 0.25 s of total processing time for 20 frames, which is approximately 10.3× faster than H.265+LDPC (2.58 s) and reduces the per-frame latency from ~ 0.129 s to ~ 0.013 s. Furthermore, VISTA consistently consumes less than half the processing time of JSCC-VFI across all interpolation proportions (37.5%, 57.4%, and 50.0% reduction at 0%, 50%, and 75% interpolation, respectively).This efficiency stems from the SLG mechanism in VISTA, which confines processing to semantically significant behavior segments, thereby reducing the number of pixels that must be encoded and decoded.In addition, a higher interpolation proportion leads to a longer processing time, since more intermediate frames must be sampled and synthesized.

Fig. 4.6 (b) presents the total number of transmitted bits for the same 20 consecutive video frames. As with Fig. 4.6 (a), results for H.265+LDPC at 50% and 75% interpolation are not applicable.Without interpolation, VISTA transmits 0.30×10^7 bits, which corresponds to approximately 66.7% of the H.265+LDPC baseline (0.45×10^7 bits) and only 31.9% of JSCC-VFI (0.94×10^7 bits), demonstrating a meaningful reduction in communication overhead.Moreover, the bit consumption of VISTA decreases monotonically as the interpolation proportion increases, since the SLG mechanism extracts more compact semantic representations at higher interpolation ratios, allowing the same video content to be conveyed with

fewer transmitted bits.

Chapter 5

GAI-driven SemCom Networks

This chapter presents a GAI-driven SemCom framework specifically designed for AIGC delivery. The chapter conducts a thorough investigation of AIGC information effectiveness from three critical perspectives: task-oriented systems, AoI, VoI, and causal control mechanisms. Furthermore, a novel architecture and associated algorithms are introduced for optimizing communication and computing resource allocation alongside knowledge management strategies, encompassing knowledge construction, updating, and sharing protocols necessary for operating and maintaining GAI-driven SemCom networks.

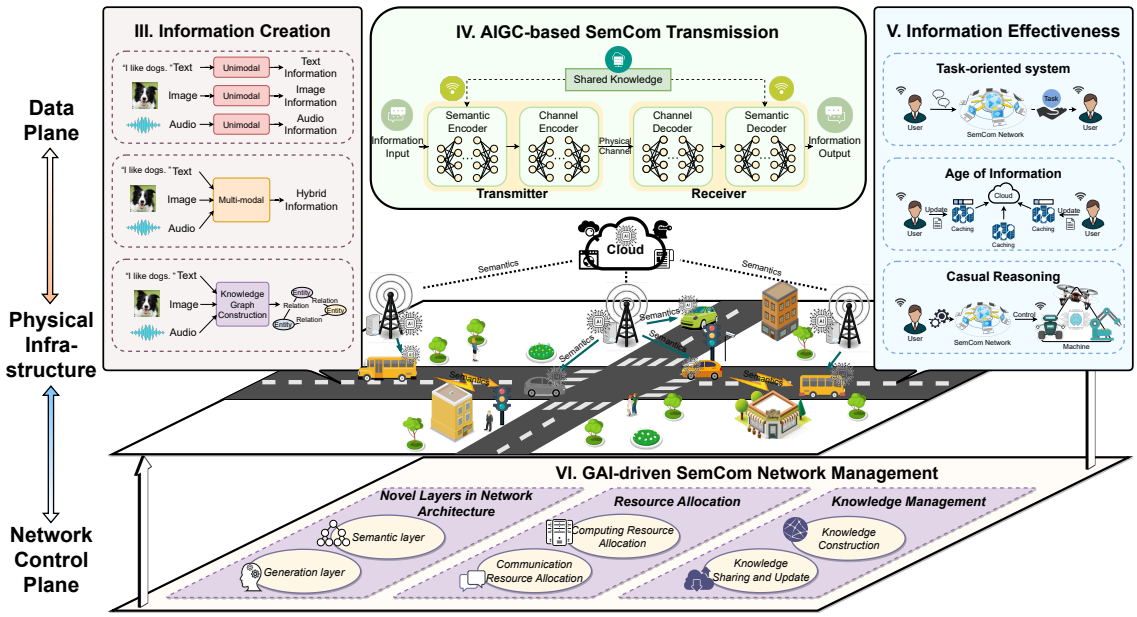


Figure 5.1: The architecture of the GAI-driven SemCom networks involving three perspectives: data plane, physical infrastructure and network control plane. The data plane layer includes information creation, AIGC-based SemCom transmission and information effectiveness. The network control plane layer includes GAI-driven SemCom network management.

5.1 GAI-driven SemCom Network Architecture

Given the basics and features of GAI and SemCom, next, we present a synergistic interaction between GAI algorithms and SemCom networks. Thus, in this section, we present the vision of GAI-driven SemCom network architecture, as depicted in Fig. 5.1. We illustrate this architecture from the perspectives of physical infrastructure, data plane and network control plane.

Physical Infrastructure

Similar to conventional communication networks, the physical infrastructure in GAI-driven SemCom networks consists of multiple wireless terminal devices (TDs), access points (APs), BSs, edge servers, and central cloud servers [196]. Besides performing conventional functions in communication systems, these entities are armed with additional intelligent techniques to support novel AIGC services. To be specific, TDs, such as smartphones, tablets, and laptops, are equipped with KBs and well-trained GAI models including encoder and decoder modules in SemCom system. Before transmission, TDs upload sensing data, as well as download knowledge and well-trained models through APs and BSs, thus integrating knowledge and updating KBs.

In GAI-driven SemCom networks, the edge nodes, including mobile edge computing (MEC) servers and BSs, enable to pre-train and fine-tune GAI models with the knowledge from themselves, connected TDs and central cloud servers. Then, edge nodes will offload the well-trained models to TDs corresponding to their tasks and environments. Additionally, the edge servers account for managing knowledge sharing and update with optimization of resource consumption (energy, bandwidth, etc.).

Due to the large storage and computing resource of central cloud servers, the large-scale GAI models can be employed and pre-trained. Virtually, most global AIGC services (e.g., ChatGPT) are trained in cloud utilizing the data from many data suppliers. Meanwhile, the centralized model will be updated, absorbing new knowledge to refresh models and adjusting resource allocation strategies.

Data Plane

The AIGC data is generated, transmitted and evaluated on the data plane of this network. First, the AIGC information is created through GAI models, including unimodal and multimodal models, which will be discussed in Section III. Then, AIGC data is transmitted through a wireless channel in the approach of SemCom. To be concrete, the source messages are fed into semantic encoder and channel encoder at the transmitter to extract and compress their

semantic information. Subsequently, the compressed semantic information passes through a wireless channel. At the other end, the distorted data are recovered by the channel decoder and semantic decoder based on the shared knowledge beforehand. Through this approach, SemCom could enhance the efficiency of AIGC transmission and resource utilization by transmitting only essential semantic information of AIGC data are transmitted.

Another function achieved by the data plane is to measure the AIGC information effectiveness from the perspectives of task completion, data freshness and relevance, as well as causal reasoning. First, some performance metrics on evaluating the task implementation for task-oriented systems are delivered [197]. Next, the AoI [198] is regarded as an important metric to measure how fresh the information is, which is significant in real-time supervision system and update system. If the information is expired, it may reduce the accuracy and reliability of system decision. Moreover, the VoI focusing on the importance and relevance of the information being transmitted is also an practicable metric for information effectiveness measurement in SemCom [199]. Finally, due to the dynamics in wireless communication environments, new measurements related to causal reasoning are envisioned considering the state of SemCom networks.

Network Control Plane

Unlike conventional communication network, in GAI-driven SemCom networks, the network management should be more intelligent, knowledgeable and adaptive to GAI. Consequently, the network control plane encompasses network architecture, knowledge management, and resource allocation. First, the novel layers in the proposed networks are discussed including semantic level and generation level. Next, the knowledge management features the utilization of KB which are essential in the processes of training both GAI and SemCom models containing public and private knowledge, especially for the personalized function. In this network, the key procedures consist of KB construction, sharing and update. To create a KB, the raw data, such as users' history and channel status, are collected, classified, and encoded. In turn, KBs are continuously monitored by GAI automatically, updated based on new knowledge and users' feedback, ensuring knowledge remains dynamic and reliable over time. Also, KBs in transceivers need to be aligned since the inconsistent KBs would lead to content misunderstanding.

Additionally, since the limited resource restricts the implementation of AIGC services with extensive data, new resource allocation methods for GAI-driven SemCom networks are urgently required. Beyond the conventionally utilized communication resources such as energy and bandwidth, some unprecedented issues are explored for SemCom networks, e.g., the matching degree of physical channel and KB. Furthermore, GAI acts as an add-on

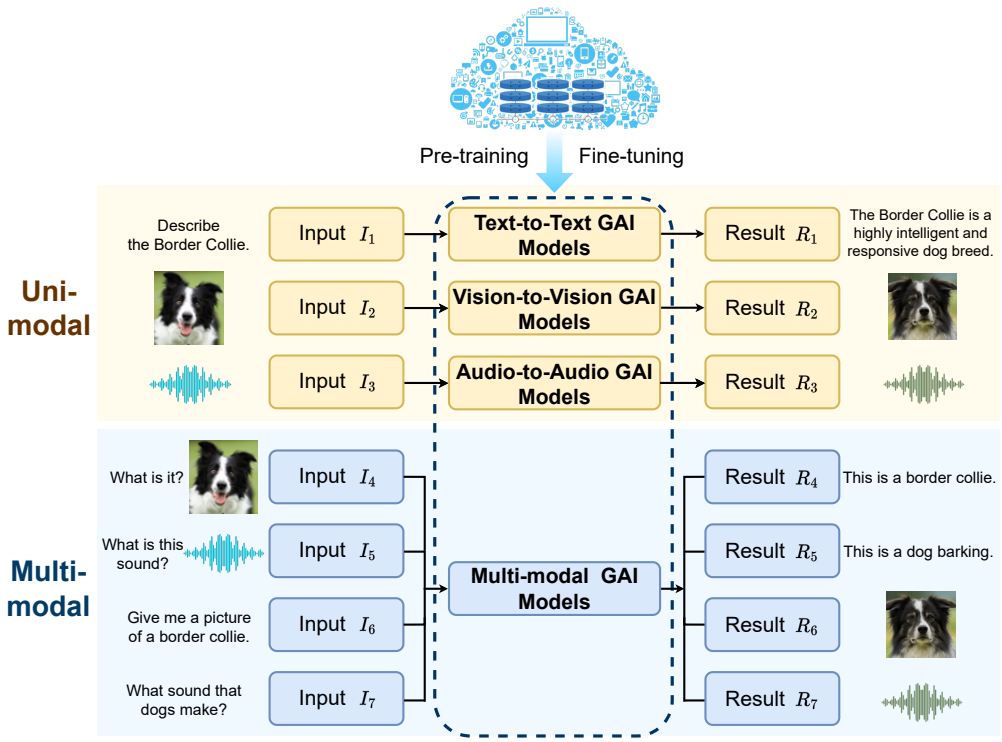


Figure 5.2: Two types of GAI models for information creation: unimodal and multimodal. Unimodal GAI models specialize in processing a single type of data, while multimodal GAI models integrate and interpret multiple data types.

module to boost network performance. The strategies for resource allocation are decided by GAI automatically and they can be adjusted dynamically according to new network status.

Specially, we focus on the network control plane which delves into the management of GAI-driven SemCom networks. We first illustrate the novel layers introduced in the network to manage resources from an architecture perspective. Then, we discuss the knowledge management, including knowledge construction and knowledge sharing and update. Finally, we investigate the computing and communication resource allocation strategies.

5.1.1 Novel Layers in GAI-driven SemCom Architecture

To manage the GAI-driven SemCom networks, some research works propose novel layers for dedicatedly dealing with semantic message. A novel GAI-assisted SemCom network framework in a cloud-edge-mobile design is proposed in [124], which enables multimodal semantic content provisioning, semantic-level joint-source-channel coding, and AIGC acquisition. The authors in [200] come up with a two-tier architecture primarily including physical and semantic levels in the semantic-aware network management and communications realm.

Drawing inspiration from previous research, the proposed network architecture, as depicted in Fig. 5.3, comprises three distinct layers: the *physical layer*, *the semantic layer*, and

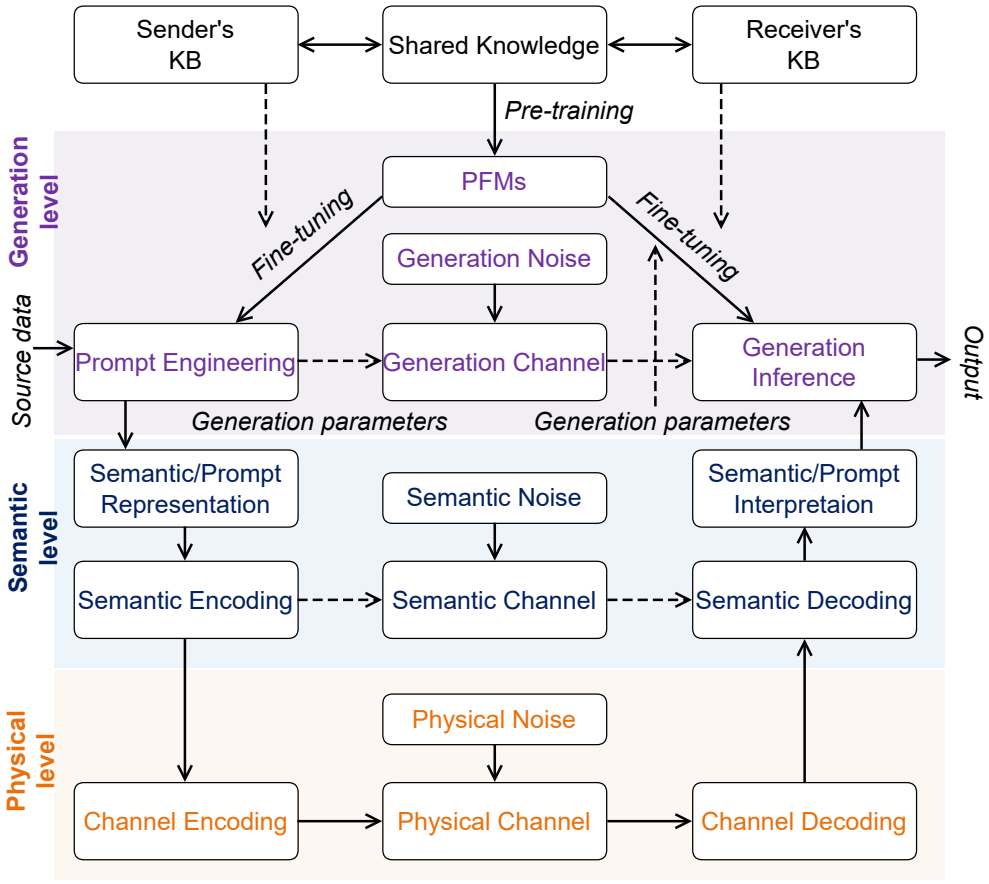


Figure 5.3: GAI-driven SemCom networks with physical, semantic, and generation levels.

the newly introduced *generation layer*. The *physical layer* handles the actual data transmission, encompassing the encoding and decoding of signals. The *semantic layer* focuses on the meaning or context of the transmitted information. The innovative *generation layer* utilizes semantic information and algorithmic parameters to guide GAI models, producing content that aligns with specific communication goals. Crucially, knowledge is shared between the sender and receiver beforehand to pre-train the GAI foundation models. The sender employs prompt engineering, based on the sender's KB, to create generation parameters from source data. Once transmitted through a channel, these signals are processed into generation inferences at the receiver end, guided by the receiver's KB, and then interpreted into the final output. This architecture represents a cohesive integration of physical transmission, semantic understanding, and content generation, tailored to enhance communication efficacy for AIGC services.

Moreover, the pre-trained foundation models (PFMs) can be fine-tuned for the specific tasks such as feature extraction and parameter optimization, to provide personalized services and meet the unique demands of various applications [201]. The introduced novel layers also restrict the exposure of sensitive information as only semantic instructions and prompts are

transmitted. Hence, the integration of GAI into SemCom models is envisioned to herald a new era of unparalleled personalization, adaptability, and security.

5.1.2 Knowledge Management in GAI-driven SemCom Networks

Knowledge, regarded as the foundation of GAI, comprises two categories as follow:

- **Background Knowledge:** Task-specific parameters at the transmitter end and required expertise for model fine-tuning at the receiver end form the components of this tier.
- **Common Knowledge:** A shared database enables both the transmitter and receiver to pull relevant data, facilitating the use of PFMs for further refinements.

In this sense, knowledge management is significant in the proposed networks since GAI relies on accumulative knowledge for learning, while the network hinges on constant knowledge sharing and updating for seamless operation [202, 203]. To be concrete, knowledge is first constructed from raw data, then shared and updated between each communication entity.

Knowledge Construction

In terms of knowledge construction in GAI-driven SemCom networks, KBs are compiled from an amalgamation of public and proprietary data sources, processed through GAI algorithms. These sources are diverse, ranging from crowdsourced content to data marketplaces, from the input of IoT sensors to passive collections, as well as encompassing user histories and records [204]. This expansive data assimilation is vital for the effective operation of the proposed networks. Besides, the KG creation is more complex, which has three fundamental processes: knowledge extraction, KRL, and KGC [205]. The initial stage in the knowledge management process involves leveraging algorithms for NER and relation extraction to distill valuable entities and their connections from unstructured data, forming a network of triples. GAI algorithms then employ KRL to convert these triples into compact, low-dimensional vectors, rendering complex knowledge into a machine-interpretable format. Finally, KGC algorithms are responsible for inferring and inserting the missing pieces within these triples via triple and relation based reasoning to ensure data integrity and completeness [206].

Knowledge Sharing and Update

After collecting the knowledge in various communication nodes, knowledge sharing and update are crucial for maintaining high accuracy and relevance of knowledge in SemCom systems. The processes of knowledge sharing and updating ensure that GAI's decisions are

created from the latest data, fostering efficiency and innovation. Regularly refreshing KBs is vital to enable quick adaptation to new market trends, technologies, and user demands. Especially in customer-centric services, it enhances personalization and dedicated user experience.

- **Knowledge Sharing:** Sharing among edge nodes enables collective learning and cooperative knowledge creation, often facilitated by methods like FL [207–209]. Additionally, a specific application for knowledge sharing in the context of the Industrial IoT is presented via edge GAI platforms [207].
- **Knowledge Update:** To sustain the accuracy and relevance of the KB community, periodic audits are employed to identify and excise outdated or incorrect data, while concurrently integrating new research and insights [210–212]. Tracking KB versions periodically is advisable to streamline the management of these updates. Such a system allows for the archiving of significant updates as separate versions, providing users the flexibility to compare changes and revert to prior versions if needed [207].

Through these multifaceted strategies, the KB community maintains high integrity, adaptability, and utility, thereby serving as a robust asset in GAI-driven SemCom networks.

5.2 Problem Formulations

Based on the framework presented shown in Fig. 5.3, we propose a detailed framework which delivers images with multi-model prompts for accurate content decoding [152]. Particularly, this system utilizes multi-modal prompts which incorporate visual prompts to restore images' structural fidelity, and textual prompts to capture the semantic information of images. The semantic encoder and decoder are design as follow:

5.2.1 Semantic Encoder

The semantic encoder in this framework extracts textual and visual prompts from a source image. To generate textual prompts, a capsulate vision-language pre-training (VLP) model for vision-language understanding and generation tasks. This model is the multi-modal mixture of encoder-decoder (MED), including three distinct operational modes:
(1) A Transformer-based unimodal encoder generates and aligns visual and textual representations through image and text contrastive learning.
(2) Image-grounded text encoder incorporates additional cross attention mechanisms to capture dependencies between vision and language, employing image text matching objectives

to differentiate matching from non matching combinations.
(3) Image-grounded text decoder modifies the architecture by substituting bidirectional self attention with causal self attention while retaining the cross attention and feedforward components from the encoder. This decoder uses language modeling objectives to produce image captions.

We consider a source image $\mathbf{s}_0$ whose textual description $\mathbf{d}_{tex}$ is extracted using the image-grounded text decoder component of MED, denoted as E_{MED}. This extraction can be expressed as $\mathbf{d}_{tex} = E_{\text{MED}}(\mathbf{s}_0; \omega_t)$, where ω_t represents the model parameters.

Next, the visual prompts are created by pre-trained generative diffusion model (GDM). GDMs have demonstrated their capability to model target distributions through learning a denoising process across various noise scales [77]. Beginning with Gaussian noise drawn from $\mathcal{N}(0, I)$, where I represents the covariance matrix, an effective GDM progressively refines this noisy input into a synthesized image over N denoising iterations. Inspired by [77], a function $\varepsilon_\theta^n(\mathbf{s}_n)$, which ingests a noisy image $\mathbf{s}_n$ and predicts the corresponding noise, is presented.

The GDM optimization involves the loss function $|\varepsilon_\theta^n(\mathbf{s}_n) - \varepsilon_a|$, where ε_a symbolizes the actual noise that was added to $\mathbf{s}_0$ to produce $\mathbf{s}_t$. A significant stride in the realm of denoising is the DDIM model [78], which stands out due to its deterministic generative process:

$$\mathbf{s}_{n-1} = \sqrt{\alpha_{n-1}} \left(\frac{\mathbf{s}_n - \sqrt{1-\alpha_n}\varepsilon_\theta^n(\mathbf{s}_n)}{\sqrt{\alpha_n}} \right) + \sqrt{1-\alpha_{n-1}}\varepsilon_\theta^n(\mathbf{s}_n), \tag{4}$$

and

$$q(\mathbf{s}_{n-1} \mid \mathbf{s}_n, \mathbf{s}_0) = \mathcal{N}\left(\sqrt{\alpha_{n-1}}\mathbf{s}_0 + \sqrt{1-\alpha_{n-1}} \frac{\mathbf{s}_n - \sqrt{\alpha_n}\mathbf{s}_0}{\sqrt{1-\alpha_n}}, \mathbf{0} \right), \tag{5}$$

where α_n represents a noise schedule parameter that determines how much of the original signal remains ($\sqrt{\alpha_n}$) and how much noise has been added ($\sqrt{1-\alpha_n}$). An intriguing aspect of DDIM is the capacity to run its generative procedure in reverse, deterministically retrieving the noise map $\mathbf{s}_N$ [78]. This map can be perceived as the latent encoding for the image $\mathbf{s}_0$. Though the reconstruction accuracy is commendable, the resultant $\mathbf{s}_N$ lacks higher-level semantics expected of a meaningful representation.

Thus, with the $\mathbf{d}_{tex}$ to catch the high-level semantic information, the conditional DDIM can be employed to encode an image $\mathbf{s}_0$ into the visual prompt $\mathbf{d}_{vis}$ to catch the image structure information as demonstrated in

$$\mathbf{s}_{n+1} = \sqrt{\alpha_{n+1}}\mathbf{f}_\theta\left(\mathbf{s}_n, n, \mathbf{d}_{tex}\right) + \sqrt{1-\alpha_{n+1}}\varepsilon_\theta\left(\mathbf{s}_n, n, \mathbf{d}_{tex}\right), \tag{5.1}$$

where the denoised observation is denoted as

$$\mathbf{f}_\theta\left(\mathbf{s}_n, n, \mathbf{d}_{tex}\right) = \frac{1}{\sqrt{\alpha_n}}\left(\mathbf{s}_n - \sqrt{1-\alpha_n}\varepsilon_\theta\left(\mathbf{s}_n, n, \mathbf{d}_{tex}\right)\right). \tag{5.2}$$

Then, an image can be regenerated accurately by using both textual and visual prompts. The visual prompt $\mathbf{v}_{\text{sem}}$ is defined as $\mathbf{v} = \mathcal{V}_E\left\{\mathbf{d}_{tex}, \mathbf{s}_0; \omega_{vis}\right\}$, where $\mathcal{V}_E$ denotes the diffusion process with N steps, and ω_{vis} represents its parameter set.

5.2.2 Semantic Decoder

The purpose of the GDM-based semantic decoder is to use the textual and visual prompts, i.e., $\mathbf{d}_{tex}$ and $\mathbf{d}_{vis}$, to generate the source image $\mathbf{s}_0$. This decoder is a conditional DDIM that models $p_\theta\left(\mathbf{s}_{n-1} \mid \mathbf{s}_n, \mathbf{d}_{tex}\right)$ to match the noising distribution $q\left(\mathbf{s}_{n-1} \mid \mathbf{s}_n, \mathbf{s}_0\right)$ defined in (5), with the following reverse (generative) process as:

$$p_\theta\left(\mathbf{s}_{0:N} \mid \mathbf{d}_{tex}\right) = p\left(\mathbf{s}_N\right)\prod_{n=1}^{N} p_\theta\left(\mathbf{s}_{n-1} \mid \mathbf{s}_n, \mathbf{d}_{tex}\right), \tag{8}$$

which can be further expressed as

$$p_\theta\left(\mathbf{s}_{n-1} \mid \mathbf{s}_n, \mathbf{d}_{tex}\right) = \begin{cases} \mathcal{N}\left(\mathbf{f}_\theta\left(\mathbf{s}_1, 1, \mathbf{d}_{tex}\right), \mathbf{0}\right) & \text{if } n = 1, \\ q\left(\mathbf{s}_{n-1} \mid \mathbf{s}_n, \mathbf{f}_\theta\left(\mathbf{s}_n, n, \mathbf{d}_{tex}\right)\right) & \text{otherwise.} \end{cases} \tag{9}$$

Training is done by optimizing

$$L_{\text{sim}} = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{s}_0, \varepsilon_n}\left[\left\|\varepsilon_\theta\left(\mathbf{s}_n, n, \mathbf{d}_{tex}\right) - \varepsilon_n\right\|_2^2\right], \tag{10}$$

where $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{s}_n = \sqrt{\alpha_t}\mathbf{s}_0 + \sqrt{1-\alpha_n}\varepsilon_n$.

5.3 Experiments and Results

We conduct numerical evaluations to demonstrate the performance of GAI-driven SemCom systems, and we implement all subsequent simulations in a computer with an Intel Core i9 CPU and NVIDIA Geforce RTX 3090 Ti GPU processors where the main software environment is Python 3.9. Moreover, this system employs well-trained diffusion model and does not require joint training in pre-training process, which offers a reduction in both computational complexity and energy cost compared to conventional SemCom methods.

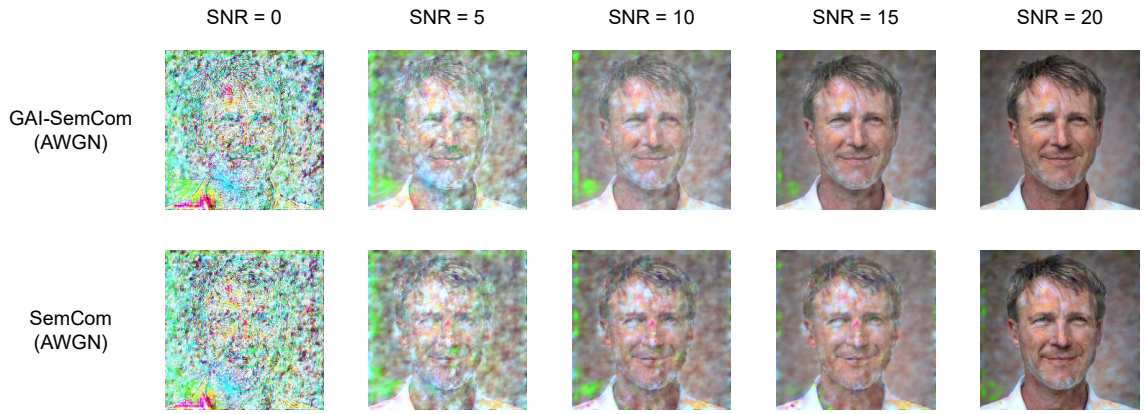


Figure 5.4: Performance comparison of GAI-SemCom and SemCom for image transmission over AWGN channel at different SNR levels (0 – 20 dB)

As illustrated in Fig. 5.4, the GAI-SemCom framework demonstrates superior performance in image reconstruction quality compared to the standard SemCom approach across different SNR levels in the AWGN channel. At extremely low SNR conditions (SNR = 0), GAI-SemCom produces images with discernible facial features and preserved color information, albeit with significant noise, while maintaining better structural integrity. As the SNR increases to 5 and 10 dB, GAI-SemCom achieves notably clearer facial details, more accurate skin tones, and reduced artifacts compared to SemCom, which still exhibits substantial color distortion and blurriness at these levels. At higher SNR values (15 and 20 dB), both methods converge toward high-quality reconstruction, but GAI-SemCom consistently maintains sharper edges, more natural color reproduction, and better preservation of fine details such as facial hair and background elements.

As shown in Fig. 5.5, the GAI-driven SemCom system shows great performance of SSIM over a wide range of SNRs spanning from 0 to 35 dB, comparing with classical SemCom system [144] and traditional wireless image transmission system using LDPC codes [213]. The channel types are AWGN channel and Rayleigh fading channel, and the channel coding approach adopted is binary phase-shift keying (BPSK). We can observe that the superior performance of the GAI-driven SemCom system is particularly pronounced under varying SNRs from 10 to 35 dB, where the SSIM value rapidly improves, showcasing its exceptional capabilities in favorable conditions. In lower SNR conditions, the generation of prompts is adversely impacted by the presence of noise, leading to a degradation in the quality and accuracy of the reconstructed images. As illustrated in Fig. 5.6, the GAI-driven SemCom system demonstrates a significant reduction in processing time compared to traditional wireless communication systems. This improvement can be attributed to the utilization of a well-trained diffusion model, which efficiently processes and generates the transmitted data. Moreover,

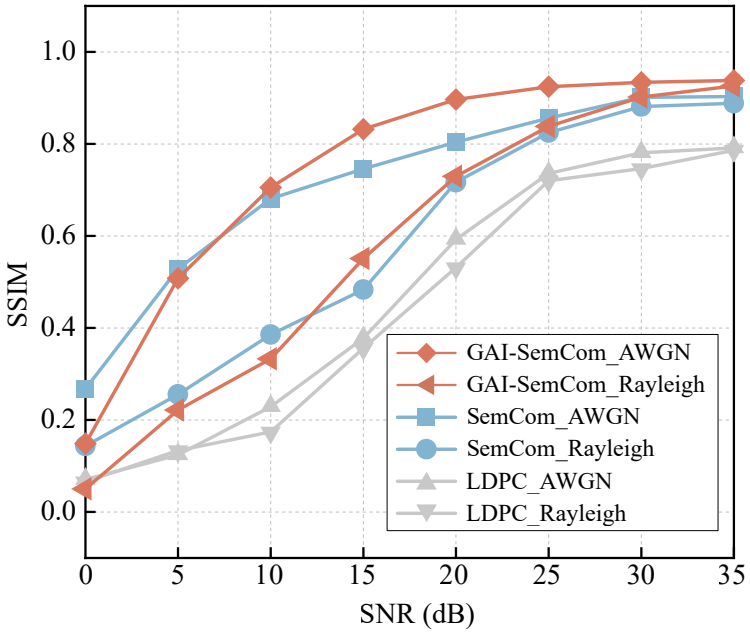


Figure 5.5: SSIM performance of images reconstructed by a GAI-driven SemCom system, a classical SemCom system and a traditional wireless communication system versus varying SNRs.

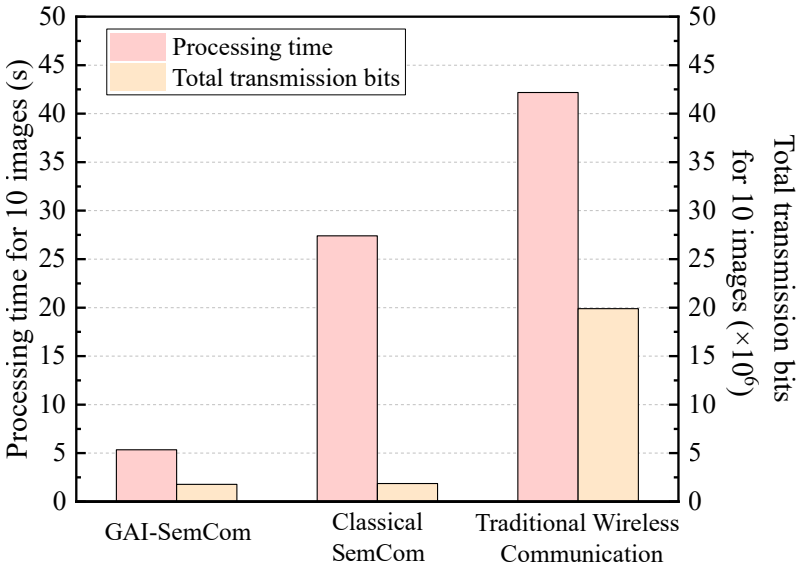


Figure 5.6: Processing time and total transmission bits for 10 images of a GAI-driven SemCom system, a classical SemCom system and a traditional wireless communication system.

the GAI-driven SemCom system requires the transmission of fewer bits compared to traditional wireless communication systems, highlighting its enhanced efficiency in data com-

pression and transmission. However, this system may transmit slightly more bits compared to classical SemCom systems due to its capability to handle multimodal data transmission, which involves the integration and processing of various data types.

5.4 Use Cases

In this section, we conceive important cases for GAI-driven SemCom networks, including autonomous driving, smart cities, and the Metaverse.

5.4.1 Autonomous Driving

In the realm of autonomous driving, AVs need to actively gather sensing data and swiftly analyze the data to form a perception of their surrounding environment. However, the data collection and transmission processes for AVs are often cumbersome and expensive [214–216]. To this end, several prominent studies [217, 218] have been conducted on SemCom systems for autonomous driving. [217] focuses on knowledge sharing strategy to improve driving decisions in AV systems. [218] presents a high altitude platform-supported fully connected AV network where the traffic infrastructure transmits its semantic information to the macro BS whenever it observes a connected AV.

Looking ahead, GAI-driven SemCom systems have the potential to revolutionize autonomous driving by enabling AVs to exchange semantic information with other nodes swiftly and efficiently. By generating semantic data between communication nodes, the latency of data transmission can be significantly reduced, enabling real-time decision-making in dynamic driving environments. However, the limited communication bandwidth and on-board processing capacity in connected and autonomous vehicles (CAVs) pose a critical challenge in terms of resource management and competition. To address this issue, advanced AI algorithms can be employed to intelligently allocate resources based on the specific requirements of each AV and the overall traffic situation. By considering factors such as the criticality and time-sensitivity of the semantic data being exchanged, these networks can ensure that safety-critical information, such as collision warnings or sudden changes in road conditions, is given the highest priority and allocated the necessary resources to guarantee minimal latency and maximum reliability in transmission. One potential solution to tackle the resource management challenge is the development of a collaborative multi-objective optimization framework that takes into account various performance metrics related to driving safety, vehicle string stability, and road traffic throughput.

5.4.2 Smart City

Smart cities represent intricate socio-technical networks made up of various interrelated components like IoT devices, mobile phones, other portable devices, physical infrastructures, services, applications, and the data shared among these elements [219]. The high complexity of smart city networks comes from dealing with numerous data, diverse content types, distributed control systems, and the intricate interconnections between various urban subsystems spanning physical, digital, organizational, and societal spheres [220]. Some existing works [221, 222] utilize ML algorithms and semantic models to handle smart city issues. The authors in [222] focus on the application of semantic technologies that can enhance interoperability among Internet-of-Everything components in smart cities. In [221], a smart city digital twin architecture is introduced, which facilitates the representation and reasoning of semantic knowledge.

Compared with these works, GAI-driven SemCom networks can develop smart cities by embedding multimodal sensing data into semantic space, connecting them with the semantics. The scalability of these networks allows them to be applied to a wide range of tasks, benefiting multiple domains within the smart city ecosystem. These network can reduce the cost of training different systems while meeting the diverse requirements for different tasks. For instance, in the domain of transportation, GAI-driven SemCom networks can significantly enhance efficiency by optimizing traffic flow, reducing congestion, and improving public transit services. By analyzing real-time traffic data, these networks can dynamically adjust traffic signal timings, recommend optimal routes to drivers, and predict and mitigate potential bottlenecks. Energy conservation is another domain where GAI-driven SemCom networks can be employed. By leveraging semantic data from smart meters, weather sensors, and building management systems, these networks can optimize energy consumption at both the individual building and city-wide levels.

5.4.3 Metaverse

Metaverse is a collective virtual shared space, emerging from the fusion of enhanced virtual depictions of our physical world and persistent digital realms. Advancements in GAI technologies have led to a significant increase in Metaverse applications, notably in AR, VR, and extended reality (XR) [20, 223–225]. Both academia and industry are exploring the Metaverse to create immersive, dynamic virtual landscapes that can adapt in real-time, reflecting user interactions and inclinations. However, to authentically mirror our physical world within these virtual domains, vast amounts of data, spanning text, images, and videos are essential. Recent studies [226–228] have started to explore potential solutions to this problem through

GAI-driven SemCom systems. These frameworks utilize diffusion models [226], Magic3D [227], and GANs [228] to generate digital content, render graphics, and exchange semantic information between transceivers' local semantic multiverses. Additionally, a trustworthy SemCom system using FL and intelligent radio is conceived for privacy protection [228].

Furthermore, user experience quality is of utmost importance in Metaverse applications. Future research can focus on designing and employing GAI-driven SemCom systems to generate and transmit personalized content, enabling tailored services for individual users. By analyzing user preferences, behavior patterns, and contextual information, these networks can infer the content that users are most interested in and adapt their models accordingly to meet users' specific requirements.

Chapter 6

Safeguarded AI-driven Semantic Communication

6.1 Safeguarded AI Design Principles

In this section, we introduce the design principles in safeguarded AI framework. Fig. 6.1 illustrates the architecture of safeguarded AI framework. An AI model operates within a world model that simulates or represents the operational environment. The model generates outputs based on its inputs, which are then evaluated by a gatekeeper according to safety specifications derived from the world model. Only outputs deemed safe proceed to the operator, who is responsible for executing these approved actions in the real-world system. Finally, users, such as domain experts, monitor the outcomes and behaviors post-execution and provide feedback to gatekeepers, allowing future safety evaluations and updates. Specifically, the world model, safety specifications, and gatekeeper are three main components:

- *World Model*: A world model serves as a mathematical and logical representation of real-world dynamics, establishing the formal foundation necessary for rigorous safety verification in safeguarded AI framework. The world model transforms abstract safety concepts into verifiable constraints by converting vague directives into precise mathematical inequalities and logical predicates over defined model states. By systematically capturing safety-relevant environmental aspects, the model provides structured observations that enable the AI model's predictive reasoning about potential actions and safety implications.

- *Safety Specifications*: Safety specifications contain the safety properties of AI systems, defining the conditions under which the system is considered to operate safely. These specifications typically include constraints on system behavior, acceptable operating

ranges, failure modes to avoid, and responses to uncertain or adversarial inputs. They serve as formal criteria for verifying and validating the system's reliability, robustness, and compliance with safety standards throughout its deployment lifecycle.

- *Gatekeepers*: The central concept of safeguarded AI is gatekeepers, which function as an oversight mechanism for autonomous AI systems. A gatekeeper defines the safety boundaries and ensures that the outputs and actions of AI systems strictly adhere to predefined and verifiable safety boundaries. According to this model, an AI output is deemed safe only if it remains within these boundaries. Advanced methodologies, including mathematical analysis and ML techniques, are employed to monitor, assess, and certify the safety of AI decisions in real-time.

However, narrowing down to SemCom, the design of these three components presents substantial challenges.

How to build world models for SemCom networks? The world model for SemCom networks is hybrid, integrating both syntactic and semantic levels. The syntactic level encompasses traditional communication models, such as channel models. In contrast, the semantic level involves models for semantic encoding and decoding, as well as sematic aware adaption. The hybrid model introduces new aspects to traditional problem formulations, such as semantic-aware resource allocation strategies. Consequently, building an world model for SemCom networks is a complex challenge, requiring the seamless integration of syntactic and semantic components while addressing the dynamic nature of communication environments.

How to define semantic safety specifications? Safety specifications in SemCom networks involve abstract and context-dependent factors. In this context, the definition of "safety" is extended beyond conventional reliability measures, including the accuracy and consistency of semantic understanding. For instance, a mismatch in interpretation between different models can be considered "unsafe", potentially leading to miscommunication and unintended actions. However, defining and measuring semantic safety remains ambiguous.

How to design gatekeepers in SemCom networks? In SemCom networks, gatekeepers should operate in real-time while monitoring complex semantic interpretations, requiring sophisticated algorithms that balance computational efficiency with monitoring accuracy. Moreover, gatekeepers need to adapt to varying network conditions and semantic contexts without compromising its monitoring reliability.

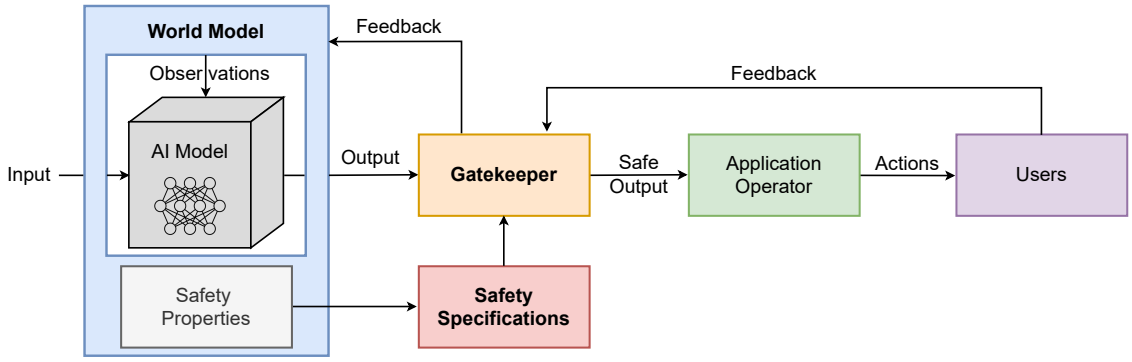


Figure 6.1: The architecture of safeguarded AI framework.

6.2 Safeguarded-AI in Semantic Communication Networks

To solve the challenges discussed in Section 6.1, in this section, we investigate potential approaches to world model construction, safety specification formulation, and gatekeeper design in SemCom networks.

6.2.1 World Models

In order to explore the safety boundaries for SemCom networks, we should first develop a world model to describe the environment in which SemCom models operate. In particular, the tasks in SemCom networks are conceptualized at two levels: end-to-end level and network level. First, at the end-to-end level, we consider the encoding and decoding processes of semantic representations as target AI model. The external changing factors, i.e., semantic representation methods, KBs, and channel conditions, constitute the world model. Second, the tasks at the network level encompass bandwidth allocation, knowledge updates, power control, etc. Here, decision-making AI models serve as the target models which are responsible for analyzing network conditions and implementing optimal decision strategies. The environment elements of these models are related to time-varying channel conditions, user allocation, and multidimensional resource constraints. Notably, knowledge resources are essential in SemCom networks, as they directly impact the accuracy of semantic encoding and decoding processes. Meanwhile, knowledge processing need significant memory and computing capabilities, which requires robust models and efficient knowledge allocation strategies under resource constraints.

To generally describe the environment of SemCom networks, we investigate various approaches that leverage both mathematical tools and ML techniques.

Mathematical Tools

For diverse goals and environment in SemCom networks, there are four categories of mathematical tools:

- Random spatial patterns: *stochastic geometry* are commonly used to model the randomness SemCom networks in both temporal and spatial domains. For example, Poisson-based models are usually utilized to derive the coverage probability of heterogeneous cellular network. In SemCom networks, tractable analysis of semantic-aware metrics can be derived by exploiting these models, such as KB synchronization rates between distributed nodes and maximum semantic information rate achievable under given channel conditions.

- Probabilistic inference: *probabilistic models*, such as Bayesian networks [229, 230], are capable to capture conditional dependencies among multiple variables. A Bayesian network is a directed acyclic graph $G = (V, E)$, where each node $v \in V$ represents a random variable, each directed edge $(u \rightarrow v) \in E$ represents a dependency between variables, and the joint probability distribution $P(V)$ factorizes according to G as $P(V) = \prod_{v \in V} P(v \mid \mathrm{Pa}(v))$, where $\mathrm{Pa}(v)$ denotes the parent nodes of v in G. Specifically, the intricate probabilistic relationships between semantic content, knowledge states, and SemCom network conditions can be captured by such probabilistic models.

- Random process: *Markov models* [231] and stochastic processes emphasize sequential structure, focusing on transitions between states to efficiently capture temporal dynamics in modeling. A Markov chain is defined by a set of states $S = \{s_0, s_1, \ldots, s_n\}$ and a set of directed transitions between these states. The process begins at an initial state s_{ini}, and the evolution of the model is governed by transition probabilities p_{ij}, where p_{ij} denotes the probability of transitioning from state s_i to state s_j. In Markov models, the next state depends only on the current state and not on the sequence of states preceding it. Particularly, these models are particularly well-suited for dynamic semantic channel modeling, SemCom network state prediction, semantic encoder and decoder evolution, and KB updates.

- Stochastic decisions: for scenarios involving multiple agents, *game theory* model how agents choose strategies to maximize their utility, considering the actions of others. Equilibrium concepts (e.g., Nash equilibrium) are used to predict how players act. For instance, a possible strategy for player i is denoted as s_i, where $i = 1, \ldots, N$. A strategy profile, a set consisting of one strategy for each player, is represented as $s^* = (s_i^*, s_{-i}^*)$, where s_{-i}^* denotes the $N - 1$ strategies of all the players except i. Additionally, the

player i's payoff is $u_i(s_i, s^*_{-i})$ as a function of the strategies. In SemCom networks, it can be utilized in competitive or cooperative tasks, e.g., resource allocation, multiple access control, network security, and knowledge sharing.

Machine Learning Tools

Compared to mathematical tools, ML approaches offer greater flexibility and intelligence for more complex environment, though they show less interpretability. ML tools for SemCom can be classified into three main categories: DL, RL, and causal learning (CL). Notably, while DL provides the foundation, RL and CL are particularly valuable for modeling the evolution and adaptation of SemCom systems over time.

- DL serves as a fundamental modeling approach, particularly for end-to-end SemCom systems. These deep neural network models excel at extracting and encoding source data and knowledge into semantic representations, as well as interpreting semantic representations back into source data. Also, instead of manually defining physical or communication dynamics, these networks are trained on observed data to predict environmental changes, channel variations, or user behaviors over time.

- RL is widely used to learn the environment's dynamics, specifically how the world model state evolves in response to actions. During interactions with the environment, the agent collects experience tuples (s_t, a_t, s_{t+1}, r_t), consisting of the current network state, action taken, resulting next network state, and received reward. These experience samples are then used to train a RL network that models the environment's behavior. The learned model includes a state transition function $s_{t+1} = f(s_t, a_t)$, which predicts the next state given the current state and action, and a reward function $r_t = g(s_t, a_t)$, which predicts the immediate reward associated with the state-action pair. In SemCom networks, the model state specifically encompasses semantic encoder/decoder parameters, KB updates and consistency metrics, contextual understanding levels, and semantic channel conditions. The reward design is tailored to semantic-aware metrics for various tasks, such as semantic fidelity preservation, KB alignment across nodes, and computational efficiency.

- CL can capture cause-and-effect relationships between semantic concepts, KBs and network states though stable causal mechanisms. It improves the interpretability of SemCom model outputs and generalization across different contexts, KBs and network conditions. As one of the core CL methods, a structural causal model (SCM) can be formally represented as a tuple $M(\mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}))$, where $\mathbf{U}, \mathbf{V}$ represent the sets of

external and model variables, respectively. The set $\mathbf{F} = \{f_1, f_2, \ldots, f_n\}$ defines deterministic functions such that each $V_i \in \mathbf{V}$ is determined by $V_i = f_i(\mathrm{Pa}(V_i), U_i)$. These causal models can be integrated into Bayesian networks for semantic inference and RL frameworks for adaptive SemCom strategies.

6.2.2 Safety Specifications

Safety specifications identify safety properties derived from the world model, serving as objective functions and/or constraints in AI models for SemCom networks. These specifications are classified into four categories.

- The scale of *knowledge domains* in KBs and the *KB alignment coefficient* between transceivers represent critical safety specifications. These metrics must exceed defined thresholds to ensure the accuracy of semantic encoders and decoders. Specifically, insufficient knowledge domain coverage will result in inaccurate semantic interpretation. When the KB alignment coefficient falls below the established minimum threshold, knowledge sharing protocols between transceivers should be triggered to achieve proper alignment.

- *Semantic effectiveness* is fundamentally connected to AI safety because it measures whether SemCom models accurately preserve and convey intended meaning. This effectiveness can be assessed through metrics such as semantic similarity calculated at the semantic decoder output.

- The *semantic quality of service (S-QoS)* focuses on the meaning preservation, in contrast to traditional QoS. For example, it quantifies how well semantic information is preserved relative to bit consumption and semantic transmission latency [232]. Degraded S-QoS may signal that SemCom models cannot process semantic information quickly enough for safe real-time decisions.

- *AoI/Age of Incorrect Information (AoII)* in SemCom measures how timely information and knowledge reaches the receiver compared to when it was generated, or how long incorrect information and knowledge persists before correction. When AoI or AoII values are high, SemCom models risk making unsafe decisions based on outdated or incorrect semantic information. Thus, low AoI and AoII are critical safety requirements for time-sensitive applications.

6.2.3 Gatekeeper Design

According to safety specifications, gatekeepers are implemented to develop and maintain decision-support tools for the output generated by AI models in SemCom networks. As illustrated in Fig. 6.1, the gatekeeper is positioned between the AI model and application operator. It analyzes feedback received from users to determine the appropriate safety boundaries. It then filters out any outputs that fall outside the defined safe region and forwards only the reliable outputs to the operator.

Let us discuss four techniques to design and implement gatekeepers in SemCom systems.

- Simple *filter* based on predefined safety specifications is a fundamental mechanism in gatekeepers, which strictly maintain AI behavior within established safety boundaries. These safety boundaries are established through rigorous empirical testing and calibrated to meet diverse user requirements. Moreover, gatekeepers can evolute by dynamically adjusting these boundaries based on ongoing performance analysis and evolving user needs.

- *Random smoothing* can be developed to analyze the robustness of AI models in SemCom networks. This method works by creating a smoothed classifier for semantic encoding/decoding around a its model by adding calibrated Gaussian noise to inputs before classification. By aggregating predictions across multiple noise-perturbed versions of the same input, random smoothing establishes probabilistic guarantees of consistent classification for semantic information within a specified radius around any input point.

- *Conformal prediction* [233] is a method that transforms the output of SemCom model into a prediction set instead of a single prediction. Therefore, it can be used to predict the original data in semantic decoding process. In particular, this method offers formal guarantees of reliability that satisfies $P(\text{true output} \in \text{predicted set}) \geq 1 - \alpha$, for any user-defined miscoverage level α. Moreover, conformal prediction adaptively adjusts its thresholds based on past prediction errors in order to minimize regret over time. When the semantic decoder model is uncertain, the prediction set can be intentionally large, or in some cases, may even indicate uncertainty by returning no prediction at all. This behavior prevents SemCom models from blindly producing unreliable outputs in uncertain conditions. In addition, conformal prediction provides calibrated confidence estimates that are valid regardless of the underlying model or data distribution. By avoiding unjustified certainty, it effectively addresses the issue of overconfidence.

- *Monte Carlo dropout* [234] is utilized to estimate epistemic uncertainty in neural networks, can also be applied to enhance the robustness of AI models in SemCom net-

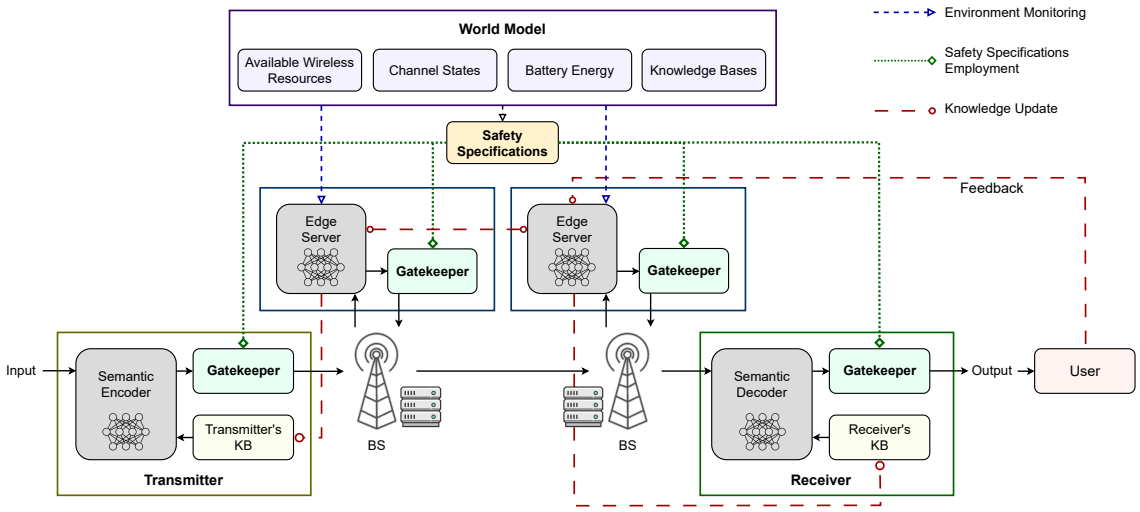


Figure 6.2: The framework of the safeguarded-AI driven SemCom framework.

works. By enabling stochastic forward passes during semantic decoding, a Monte Carlo dropout mechanism can generate multiple reconstructed data for the same received signals, forming a distribution of possible meanings. Analyzing the variation across these reconstructions allows the system to quantify its uncertainty in semantic interpretation. This not only prevents overconfident but incorrect semantic reconstructions under uncertain channel conditions but also enables the system to adjust its communication strategy dynamically based on the estimated reliability of the reconstructed meaning.

6.3 Safeguarded AI-driven Semantic Communication Framework

In this section, we present a uniform safeguarded AI-driven SemCom framework as shown in Fig. 6.2. SemCom networks consist of end devices, BSs, edge servers, and cloud servers. In common SemCom systems, a semantic encoder extracts and encodes semantic information from source data based on its KB. The information is transmitted wirelessly via edge servers to the receiver. At the receiver, the semantic information are recovered by a semantic decoder. Building upon classical SemCom system architecture, our framework employs gatekeepers to monitor AI model behavior and ensure compliance with predefined safety boundaries. This theoretical foundation allows our framework to be universally deployed across any layers, tasks, and modules utilizing AI models.

Particularly, as multi-modal large language models (MLLMs) and foundation models are becoming state-of-the-art in SemCom systems, they present unique safeguarded AI chal-

lenges [**?**]. These large-scale models offer superior semantic understanding and cross-modal reasoning capabilities [235]. However, they simultaneously introduce significant safety concerns including complex internal representations, potential for cross-modal misalignment, and substantial computational resource consumption. Our framework can address these challenges by implementing specialized gatekeepers that monitor foundation model outputs for cross-modal semantic consistency and resource costs. Crucially, our framework focuses on controlling and validating the outputs of AI models rather than modifying the models' internal architectures or training procedures which is ideally suited for encapsulated MLLMs and foundation models.

Typically, in our proposed framework, the world model is first established as the foundation. By considering the safety-specific properties in the world model, the task-specific safety specifications are defined to evaluate the performance of AI model in SemCom networks. Gatekeepers are strategically positioned after AI models, including semantic encoders, edge servers, and semantic decoders. At each distinct deployment point, these gatekeepers serve different roles:

- After semantic encoders, gatekeepers monitor semantic representation quality and ensure KBs consistency before transmission, verifying that encoded information preserves essential semantic meaning within acceptable fidelity thresholds.
- At edge servers, gatekeepers oversee resource allocation decisions and network management operations, ensuring that semantic processing tasks are distributed safely without compromising system performance or creating bottlenecks that could lead to semantic information loss.
- Following semantic decoders, gatekeepers validate the reconstructed semantic content against original intent, checking for meaning distortion or context misinterpretation that could result in unsafe operational decisions.

Before communication begins, edge servers distribute pre-trained SemCom models with KBs to end devices. The transmitter first extracts semantic information from the source. The gatekeeper after semantic encoder evaluates the generated semantic information against established safety specifications. If they are in the safety boundaries, the gatekeeper forwards the validated output to the BS. The edge server at the BS then determines optimal network strategies (e.g., resource allocation) for transmitting the semantic information, with these decisions also subject to gatekeeper verification. Subsequently, the receiver decodes the received messages and reconstructs the original data from semantic features using its KB. The

gatekeeper after semantic decoder monitors the reconstructed data , evaluates the entire process, and filter out unsafe output for users who will provide feedback based on gatekeepers' behavior. This feedback, along with real-time sensing data, serves as new knowledge shared among communication nodes. To maintain communication quality, knowledge updates are strategically scheduled during off-peak periods when network traffic is minimal.

Moreover, considering real-world scenarios, implementation details include:

- *Unbiased Datasets Collection*: Unbiased data collection and calibration ensures that training data for SemCom models accurately reflects real-world operational conditions. Data can be systematically gathered from controlled laboratory testbeds and open-field semantic network deployments. The controlled environment establishes baseline performance metrics by systematically varying antenna configurations, beamforming settings, and KB synchronization parameters under reproducible conditions. Conversely, real-world deployments capture the full complexity of operational scenarios, including dynamic context changes, user mobility patterns, and environmental factors.

- *Empirical Testing*: Rigorous expert testing is conducted based on comprehensive empirical understanding of both SemCom systems and practical networking environments. This testing involves multiple stages, including controlled laboratory evaluations, limited field trials, and full-scale deployments with continuous monitoring. Domain experts from SemCom systems, networking, and AI safety fields collaborate to design test scenarios that specifically target potential failure modes. The real-world evaluations incorporate adversarial testing methodologies, deliberately challenging the system with difficult edge cases to establish robust safety bounds.

- *Interfaces Deployment*: The mechanisms and protocols of interfaces should be designed to facilitate seamless, secure, and efficient data exchange and interaction between the safeguarded AI algorithms and the SemCom network infrastructure. These interfaces implement standardized application programming interfaces (APIs) with well-defined input/output specifications, error handling protocols, and performance monitoring capabilities. Additionally, they incorporate version compatibility layers to accommodate gradual updates of SemCom models and KBs across the network without disrupting ongoing communications.

6.4 Case Study

In our case study, we evaluate the performance of our proposed framework through simulations by targeting DeepSC [2], a conventional SemCom AI model. We compare its perfor-

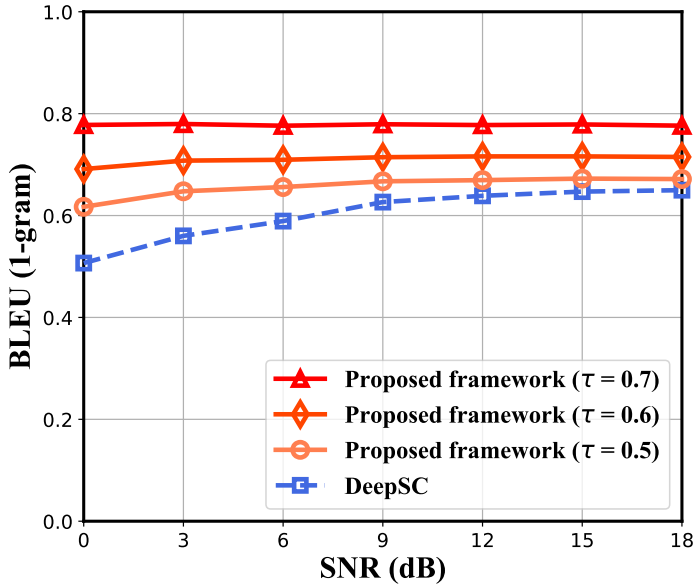


Figure 6.3: BLEU (1-gram) scores for the proposed framework with varying thresholds ($\tau = 0.5, 0.6, 0.7$) vs. DeepSC.

mance with and without safeguarded AI implementation. To demonstrate the improvement in output's accuracy, we employ the BLEU score, which is the same metric utilized in DeepSC. The BLEU score measures semantic similarity between transmitted and received messages by evaluating n-gram overlap and word order preservation. The gatekeeper filters out outputs with BLEU scores below predefined safety thresholds. We define threshold parameters τ at 0.5, 0.6, and 0.7 to assess diverse performances.

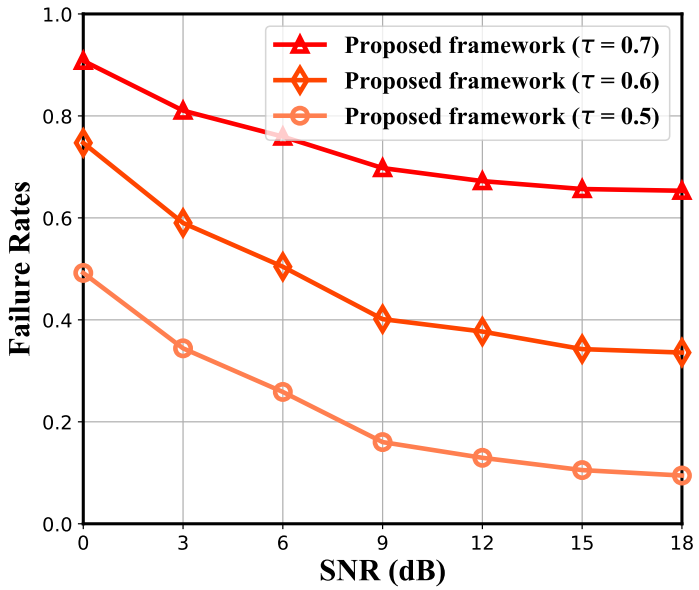


Figure 6.4: Failure rates of the proposed framework with varying thresholds ($\tau = 0.5, 0.6, 0.7$).

Fig. 6.3 presents the 1-gram BLEU scores of the proposed framework under different thresholds ($\tau = 0.5, 0.6, 0.7$) compared to the benchmark DeepSC versus varying SNRs (from 0 dB to 18 dB). Our proposed framework consistently outperforms DeepSC over the entire SNR range, especially in low-SNR conditions. As the threshold τ increases, the BLEU score of the proposed method improves, and it obtains the highest BLEU scores when $\tau = 0.7$. Particularly, at SNR = 0 dB, the proposed framework with $\tau = 0.7$ achieves a 54% improvement in BLEU score compared to DeepSC.

Moreover, Fig. 6.4 shows the the failure rates of the proposed framework under different thresholds ($\tau = 0.5, 0.6, 0.7$) versus varying SNR. The failure rate is calculated as the ratio of the number of filtered-out reconstructed sentences to the total number of reconstructed sentences. Our proposed framework exhibits higher failure rates at low SNRs due to poor channel conditions, particularly when $\tau = 0.7$. As the SNR increases, the failure rate of the proposed framework gradually decreases. In contrast, as the semantic threshold increases, the failure rate rises significantly. Notably, this figure highlights a key trade-off: although the proposed framework achieves more accurate semantic reconstruction at higher thresholds, it naturally incurs a higher risk of data reconstruction failure. To balance semantic accuracy and reconstruction robustness, the gatekeeper can dynamically adjust the threshold based on channel conditions and user requirements.

Chapter 7

Conclusion and Future Work

In conclusion, this thesis advances the field of SemCom by systematically addressing the fundamental challenges of integrating KGs into multi-modal transmission systems. Through five interconnected contributions spanning video delivery, generative AI networks, audio transmission, comprehensive KG fusion methodologies, and AI-safeguarded frameworks, this research establishes both theoretical foundations and practical implementations for KG-empowered SemCom systems. By tackling the critical issues of semantic alignment, data reconstruction under adverse conditions, and transmission overhead management, this work demonstrates that KGs can serve as an effective bridge between abstract semantic meaning and concrete communication systems. The proposed frameworks and methodologies represent a significant step toward realizing efficient, robust, and secure SemCom across diverse data modalities, laying the groundwork for future research in next-generation communication systems where meaning, rather than bits, becomes the primary currency of information exchange.

7.1 Conclusion

Chapter 1 provides a comprehensive introduction to SemCom systems, covering KG-assisted, GAI-driven and safeguarded AI-driven framework. Subsequently, we identify key gaps in existing SemCom research that form the foundation for our investigation.

Chapter 2 presents a comprehensive literature review of semantic representation, KG processing, GAI models, information theory, and transceiver design in SemCom domains is provided.

In Chapter 3, we present a novel KG-SemCom framework with semantic channel capacity modeling involving classical and semantic information theory. Moreover, a transceiver design for KG-SemCom is proposed, leveraging the power of context and public knowledge

in KGs to perform semantic representation and data reconstruction. Our numerical results demonstrate the performance of the proposed framework in terms of the named token similarity and the BERT-based sentence similarity, highlighting its effectiveness across various channel conditions.

In Chapter 4, we propose a SemCom-enabled wireless video transmission framework, named VISTA. In VISTA, a unique transceiver is developed for semantic encoding and decoding, incorporating a semantic location graph that operates alongside various neural networks for the extraction and restoration of video semantics. Simulations demonstrate a substantial decrease in the number of bits transmitted, maintaining (and even enhancing under SNR below 3 dB) video quality and transmission efficiency.

In Chapter 5, we propose a GAI-driven SemCom framework, delving into network management, covering aspects of novel layers, knowledge management, and resource allocation. We also deliver a detailed case study, which delivers images with multi-model prompts for accurate content decoding. Simulation result show its superior performance against a traditional image transmission scheme and a SemCom without multi-modal prompts.

In Chapter 6, we explore enhancing AI safety for SemCom networks through safeguarded AI implementation. We begin by introducing the safeguarded AI architecture, focusing on its three core components: world model, safety specifications, and gatekeeper. Next, we have investigated detailed designs for each component to provide practical implementation guidance. Then, We have proposed our novel safeguarded AI-driven SemCom framework with a comprehensive architectural overview and implementation details. Also, we have demonstrated simulation results in a case study to evaluate the performance of our framework.

7.2 Future Work

7.2.1 KG-empowered SemCom Systems

One important future direction for KG-empowered SemCom systems is a more rigorous complexity and latency analysis. Specifically, the computational complexity of key modules, such as Big-O complexity, will be measured with attention to resource consumption in real-world wireless networks. Beyond theoretical complexity, how practical inference latency scales with the number of entities, relations, and sentence length will also be quantified, offering clearer insights into framework scalability in large and dynamic semantic networks.

A further planned extension is the design of adaptive and stronger digital baselines for fairer comparison. Multiple digital curves spanning different source and channel coding rates will be incorporated, following the evaluation philosophy commonly adopted in prior JSCC

studies, to rigorously verify whether the proposed semantic framework genuinely outperforms well-optimized digital systems rather than a single fixed-rate baseline.

Another planned improvement concerns loss function design. Learnable or manually tuned weighting factors for different semantic objectives will be introduced, and a dedicated sensitivity study on these weights will be conducted to reveal how the trade-off among semantic preservation, reconstruction fidelity, and reasoning consistency affects overall performance, potentially motivating task-dependent optimization strategies for different SemCom scenarios.

For VISTA specifically, a comprehensive investigation of real-time capability and practical deployment scenarios is planned. The end-to-end frame rate at the receiver will be explicitly measured to verify whether it meets practical real-time thresholds, and model compression, pipeline parallelization, and hardware-aware optimization techniques will be explored to further improve runtime efficiency.

From the modeling perspective, the semantic expressiveness of the SLG will be further examined. While the current framework primarily captures spatial semantic relations, which prove effective for reconstruction, whether spatial relations alone suffice for more complex dynamic scenes remains an open question to be addressed. Accordingly, the SLG will be extended toward richer graph formulations that also encode temporal dependency, object interaction, event-level semantics, and causal structure. A particular focus will be placed on investigating whether SLG can evolve beyond spatial layout modeling to support causal semantic reasoning for video communication, especially in scenes characterized by strong temporal evolution or interaction-driven changes.

7.2.2 GAI-assisted SemCom Systems

In the future research, the optimization objective underlying the simulation results will be made explicit. Alongside this, a dedicated mechanism to detect and suppress hallucinations in GAI-based SemCom will be incorporated, such as confidence-aware generation constraints, semantic consistency verification, or retrieval-augmented grounding.

Furthermore, the tradeoff between communication overhead and computational cost will be systematically analyzed, particularly examining how larger generative models increase computation at the receiver while reducing the volume of transmitted data, with the aim of identifying optimal operating points under different resource constraints.

Finally, to address the instability and complexity associated with diffusion models, strategies such as adaptive noise scheduling, early stopping criteria, and lightweight distilled diffusion variants will be investigated to improve training stability and inference reliability within

the GAI-assisted SemCom framework.

7.2.3 Safeguarded AI-assisted SemCom Systems

For the safeguarded AI-assisted SemCom systems proposed in Chapter 6, there are several unavoidable and complex challenges that can be addressed to fully unlock its potential.

Lack of Standardized SemCom Frameworks: For various users, SemCom lacks standardized representations of meaning, creating difficulties in defining universal safety properties. The diversity of knowledge representation approaches from neural embeddings to symbolic graphs further complicates the development of consistent verification methodologies. In these cases, safety verification should address complex properties like meaning preservation and contextual appropriateness that resist straightforward mathematical formalization. These gaps collectively hinder the development of interoperable safeguarded mechanisms that can be consistently applied across different SemCom implementations.

Resource Costs: When the gatekeepers monitor the resource costs of AI models, their own resource costs are also essential considerations for the whole SemCom network. The safeguarded AI framework introduces computational overhead that creates a recursive challenge where safety mechanisms require substantial resources to monitor resource-constrained AI models. Gatekeepers consume significant processing power for continuous behavior monitoring, real-time safety boundary evaluation, and semantic validation processes including deep contextual analysis and KB alignment and updates checks. Therefore, this framework must carefully balance the trade-off between enhanced safety assurance and computational resource consumption.

Bibliography

[1] C. E. Shannon, A Mathematical Theory of Communication, The Bell system technical journal 27 (3) (1948) 379–423.

[2] H. Xie, Z. Qin, G. Y. Li, B.-H. Juang, Deep Learning Enabled Semantic Communication Systems, IEEE Transactions on Signal Processing 69 (2021) 2663–2675. doi:10.1109/TSP.2021.3071210.

[3] X. Luo, H.-H. Chen, Q. Guo, Semantic Communications: Overview, Open Issues, and Future Research Directions, IEEE Wireless Communications 29 (1) (2022) 210–219.

[4] Z. Qin, X. Tao, J. Lu, W. Tong, G. Y. Li, Semantic Communications: Principles and Challenges, arXiv preprint arXiv:2201.01389 (2021).

[5] S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I. Horrocks, The Semantic Web: The Roles of XML and RDF, IEEE Internet Computing 4 (5) (2000) 63–73. doi:10.1109/4236.877487.

[6] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, P. Zhang, Wireless Deep Video Semantic Transmission, IEEE Journal on Selected Areas in Communications 41 (1) (2023) 214–229. doi:10.1109/JSAC.2022.3221977.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[8] A. Ahmad, A. B. Mansoor, A. A. Barakabitze, A. Hines, L. Atzori, R. Walshe, Supervised-learning-Based QoE Prediction of Video Streaming in Future Networks: A Tutorial with Comparative Study, IEEE Communications Magazine 59 (11) (2021) 88–94.

[9] Z. Weng, Z. Qin, Semantic Communication Systems for Speech Transmission, IEEE Journal on Selected Areas in Communications 39 (8) (2021) 2434–2444.

[10] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, K. Huang, What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence, Journal of Communications and Information Networks 6 (4) (2021) 336–371.

[11] H. Xie, Z. Qin, A Lite Distributed Semantic Communication System for Internet of Things, IEEE Journal on Selected Areas in Communications 39 (1) (2020) 142–153.

[12] Anthropic PBC, Video AI, accessed: Jul. 19, 2023.
URL `https://www.promptengineering.org/claudes-100k-context-window-how-this-`

[13] OpenAI, ChatGPT Can Now See, Hear, and Speak, accessed: Dec. 9, 2023.
URL `https://openai.com/blog/chatgpt-can-now-see-hear-and-speak`

[14] D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, et al., Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems, arXiv preprint arXiv:2405.06624 (2024).

[15] S. Jiang, Y. Liu, Y. Zhang, P. Luo, K. Cao, J. Xiong, H. Zhao, J. Wei, Reliable Semantic Communication System Enabled by Knowledge Graph, Entropy 24 (6) (2022) 846.

[16] L. Hu, Y. Li, H. Zhang, L. Yuan, F. Zhou, Q. Wu, Robust Semantic Communication Driven by Knowledge Graph, in: 2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS), IEEE, 2022, pp. 1–5.

[17] F. Zhou, Y. Li, M. Xu, L. Yuan, Q. Wu, R. Q. Hu, N. Al-Dhahir, Cognitive Semantic Communication Systems Driven by Knowledge Graph: Principle, Implementation, and Performance Evaluation, IEEE Transactions on Communications (2023).

[18] Z. Jin, Y. Chen, F. Gonzalez, J. Liu, J. Zhang, J. Michael, B. Schölkopf, M. Diab, Analyzing the Role of Semantic Representations in the Era of Large Language Models, arXiv preprint arXiv:2405.01502 (2024).

[19] P. Jiang, C.-K. Wen, S. Jin, G. Y. Li, Wireless Semantic Communications for Video Conferencing, IEEE Journal on Selected Areas in Communications 41 (1) (2023) 230–244. doi:10.1109/JSAC.2022.3221968.

[20] L. Xia, Y. Sun, C. Liang, D. Feng, R. Cheng, Y. Yang, M. A. Imran, WiserVR: Semantic Communication Enabled Wireless Virtual Reality Delivery, IEEE Wireless Communications 30 (2) (2023) 32–39.

[21] L. Xia, Y. Sun, D. Niyato, X. Li, M. A. Imran, Joint User Association and Bandwidth Allocation in Semantic Communication Networks, IEEE Transactions on Vehicular Technology (2023).

[22] R. Thiagarajan, G. Manjunath, M. Stumptner, Computing Semantic Similarity Using Ontologies, HP Laboratories). Technical report HPL-2008-87 (2008).

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv preprint arXiv:1301.3781 (2013).

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Advances in neural information processing systems 30 (2017).

[25] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, A Survey on Knowledge Graph-based Recommender Systems, IEEE Transactions on Knowledge and Data Engineering (2020).

[26] G. A. Miller, WordNet: An Electronic Lexical Database, MIT press, 1998.

[27] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 1247–1250.

[28] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data, in: international semantic web conference, Springer, 2007, pp. 722–735.

[29] D. Vrandečić, M. Krötzsch, Wikidata: A Free Collaborative Knowledgebase, Communications of the ACM 57 (10) (2014) 78–85.

[30] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang, Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 601–610.

[31] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge Graph Embedding: A Survey of Approaches and Applications, IEEE transactions on knowledge and data engineering 29 (12) (2017) 2724–2743.

[32] Y. Lin, X. Han, R. Xie, Z. Liu, M. Sun, Knowledge Representation Learning: A Quantitative Review, arXiv preprint arXiv:1812.10901 (2018).

[33] D. B. Kurka, D. Gündüz, DeepJSCC-f: Deep Joint Source-Channel Coding of Images With Feedback, IEEE Journal on Selected Areas in Information Theory 1 (1) (2020) 178–193. doi:10.1109/JSAIT.2020.2987203.

[34] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based Learning Applied to Document Recognition, Proceedings of the IEEE 86 (11) (2002) 2278–2324.

[35] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, S. Khudanpur, Recurrent Neural Network Based Language Model., in: Interspeech, Vol. 2, Makuhari, 2010, pp. 1045–1048.

[36] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[37] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The Graph Neural Network Model, IEEE transactions on neural networks 20 (1) (2008) 61–80.

[38] R. S. Sutton, A. G. Barto, et al., Reinforcement Learning: An Introduction, Vol. 1, MIT press Cambridge, 1998.

[39] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, arXiv preprint arXiv:1409.0473 (2014).

[40] A. See, P. J. Liu, C. D. Manning, Get to the Point: Summarization with Pointer-Generator Networks, arXiv preprint arXiv:1704.04368 (2017).

[41] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to Sequence Learning with Neural Networks, Advances in Neural Information Processing Systems 27 (2014).

[42] O. Vinyals, Q. Le, A Neural Conversational Model, arXiv preprint arXiv:1506.05869 (2015).

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language Models are Unsupervised Multitask Learners, OpenAI Blog 1 (8) (2019) 9.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-shot Learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[46] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with A Unified Text-To-Text Transformer, The Journal of Machine Learning Research 21 (1) (2020) 5485–5551.

[47] R. Dey, F. M. Salem, Gate-variants of Gated Recurrent Unit (GRU) neural networks, in: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017, pp. 1597–1600. doi:10.1109/MWSCAS.2017.8053243.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All You Need, Advances In Neural Information Processing Systems 30 (2017).

[49] L. Fang, C. Li, J. Gao, W. Dong, C. Chen, Implicit Deep Latent Variable Models for Text Generation, arXiv preprint arXiv:1908.11527 (2019).

[50] W. Wang, Z. Gan, H. Xu, R. Zhang, G. Wang, D. Shen, C. Chen, L. Carin, Topic-guided Variational Autoencoders for Text Generation, arXiv preprint arXiv:1903.07137 (2019).

[51] X. Zhang, Y. Yang, S. Yuan, D. Shen, L. Carin, Syntax-Infused Variational Autoencoder for Text Generation, arXiv preprint arXiv:1906.02181 (2019).

[52] S. Dai, Z. Gan, Y. Cheng, C. Tao, L. Carin, J. Liu, APo-VAE: Text Generation in Hyperbolic Space, arXiv preprint arXiv:2005.00054 (2020).

[53] L. Yu, W. Zhang, J. Wang, Y. Yu, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 31, 2017.

[54] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, L. Carin, Adversarial Feature Matching for Text Generation, in: International conference on machine learning, PMLR, 2017, pp. 4006–4015.

[55] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, J. Wang, Long Text Generation via Adversarial Training with Leaked Information, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[56] H. Wang, Z. Qin, T. Wan, Text Generation Based on Generative Adversarial Nets with Latent Variable, in: Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part II 22, Springer, 2018, pp. 92–103.

[57] C. Zhang, C. Xiong, L. Wang, A research on generative adversarial networks applied to text generation, in: 2019 14th International Conference on Computer Science & Education (ICCSE), 2019, pp. 913–917. doi:10.1109/ICCSE.2019.8845453.

[58] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, T. B. Hashimoto, Diffusion-LM Improves Controllable Text Generation, Advances in Neural Information Processing Systems 35 (2022) 4328–4343.

[59] S. Gong, M. Li, J. Feng, Z. Wu, L. Kong, DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models, arXiv preprint arXiv:2210.08933 (2022).

[60] OpenAI, ChatGPT: Optimizing Language Models for Dialogue, accessed: Jul. 13, 2023.
URL `https://openai.com/blog/chatgpt/`

[61] M. Corporation, Reinventing Search with A New AI-powered Microsoft Bing and Edge, your Copilot for the Web, accessed: Jul. 13, 2023.
URL `https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copil`

[62] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, et al., Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, arXiv preprint arXiv:2201.11990 (2022).

[63] Anthropic, Introducing the Next Generation of Claude, accessed: Mar. 4, 2024.
URL `https://www.anthropic.com/news/claude-3-family`

[64] PaintMe, PaintMe AI: Home, accessed: Jul. 19, 2023.
URL `https://www.paintme.ai/index.php`

[65] Vizcom, The Next Generation of Product Visualization, accessed: Jul. 19, 2023.
URL `https://www.vizcom.ai/`

[66] Steve.AI, Convert Photos to Videos Instantly Using Steve.AI, accessed: Jul. 19, 2023.
URL `https://www.steve.ai/photo-video-maker`

[67] A. Van Den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: International Conference on Machine Learning, PMLR, 2016, pp. 1747–1756.

[68] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., Conditional Image Generation with PixelCNN Decoders, Advances in neural information processing systems 29 (2016).

[69] H. Huang, R. He, Z. Sun, T. Tan, et al., IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis, Advances in neural information processing systems 31 (2018).

[70] A. Razavi, A. Van den Oord, O. Vinyals, Generating Diverse High-Fidelity Images with VQ-VAE-2, Advances in neural information processing systems 32 (2019).

[71] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[72] T. Karras, S. Laine, T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[73] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2745–2754.

[74] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, L. Shao, Zero-VAE-GAN: Generating Unseen Features for Generalized and Transductive Zero-Shot Learning, IEEE Transactions on Image Processing 29 (2020) 3665–3680.

[75] D. P. Kingma, P. Dhariwal, Glow: Generative Flow with Invertible 1x1 Convolutions, Advances in neural information processing systems 31 (2018).

[76] L. Dinh, J. Sohl-Dickstein, S. Bengio, Density Estimation Using Real NVP, arXiv preprint arXiv:1605.08803 (2016).

[77] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, Advances in neural information processing systems 33 (2020) 6840–6851.

[78] J. Song, C. Meng, S. Ermon, Denoising Diffusion Implicit Models, arXiv preprint arXiv:2010.02502 (2020).

[79] S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, MoCoGAN: Decomposing Motion and Content for Video Generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1526–1535.

[80] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. J. Fleet, Video Diffusion Models, arXiv:2204.03458 (2022).

[81] Murf.AI, Transform Your Voiceover from A Home Recording to A Professional AI Voice, accessed: Jul. 19, 2023.
URL `https://murf.ai/voice-changer`

[82] Resemble.AI, Real-time Speech-to-Speech Voice Conversion, accessed: Jul. 19, 2023.
URL `https://www.resemble.ai/speech-to-speech/`

[83] MetaVoice, accessed: Jul. 19, 2023. [link].
URL `https://themetavoice.xyz/`

[84] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A Generative Model for Raw Audio, arXiv preprint arXiv:1609.03499 (2016).

[85] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, Y. Bengio, SampleRNN: An Unconditional End-to-End Neural Audio Generation Model, arXiv preprint arXiv:1612.07837 (2016).

[86] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, A. Roberts, GANSynth: Adversarial Neural Audio Synthesis, arXiv preprint arXiv:1902.08710 (2019).

[87] C. Donahue, J. McAuley, M. Puckette, Adversarial Audio Synthesis, arXiv preprint arXiv:1802.04208 (2018).

[88] Y. Meng, W. Li, S. Lei, Z. Zou, Z. Shi, Large-Factor Super-Resolution of Remote Sensing Images With Spectra-Guided Generative Adversarial Networks, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–11. doi:10.1109/TGRS.2022.3222360.

[89] K. Akuzawa, K. Onishi, K. Takiguchi, K. Mametani, K. Mori, Conditional Deep Hierarchical Variational Autoencoder for Voice Conversion, in: 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2021, pp. 808–813.

[90] F. Schneider, ArchiSound: Audio Generation with Diffusion, arXiv preprint arXiv:2301.13267 (2023).

[91] OpenAI, DALL·E 2: Exploring the Limits of Data-Driven Image Generation with Transformers, accessed: Jul. 19, 2023.
URL `https://openai.com/dall-e-2`

[92] NightCafe Studio, NightCafe Studio - AI Art Generator, accessed: Jul. 19, 2023.
URL `https://creator.nightcafe.studio/`

[93] DreamStudio, DreamStudio - Image Generator, accessed: Jul. 19, 2023.
URL `https://dreamstudio.ai/generate`

[94] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot Text-to-Image Generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.

[95] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv preprint arXiv:2204.06125 (2022).

[96] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, T.-Y. Lin, Magic3D: High-Resolution Text-to-3D Content Creation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 300–309.

[97] B. Poole, A. Jain, J. T. Barron, B. Mildenhall, DreamFusion: Text-to-3D using 2D Diffusion, arXiv preprint arXiv:2209.14988 (2022).

[98] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: Text-Based Real Image Editing with Diffusion Models (2023). arXiv:2210.09276.

[99] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, K.-Y. K. Wong, Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models, Advances in Neural Information Processing Systems 36 (2024).

[100] Synthesia, Create Professional Videos without Mics, Cameras, or Actors, accessed: Jul. 19, 2023.
URL `https://www.synthesia.io/`

[101] Pictory, Script To Video Creation In Minutes, accessed: Jul. 19, 2023.
URL `https://pictory.ai/pictory-features/script-to-video`

[102] Meta AI, accessed: Jul. 19, 2023. [link].
URL `https://makeavideo.studio/`

[103] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, D. Erhan, Phenaki: Variable Length Video Generation From Open Domain Textual Description, arXiv preprint arXiv:2210.02399 (2022).

[104] W. Hong, M. Ding, W. Zheng, X. Liu, J. Tang, CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers, arXiv preprint arXiv:2205.15868 (2022).

[105] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, M. Z. Shou, Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7623–7633.

[106] Murf.AI, Go from Text to Speech with A Versatile AI Voice Generator, accessed: Jul. 19, 2023.
URL `https://murf.ai/`

[107] PlayHT, AI Powered Text to Voice Generator, accessed: Jul. 19, 2023.
URL `https://play.ht/blog/ai-text-to-speech-voice-cloning-technology-overvie`

[108] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards End-to-End Speech Synthesis, arXiv preprint arXiv:1703.10135 (2017).

[109] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, et al., AudioLM: A Language Modeling Approach to Audio Generation, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

[110] READ-COOP, Unlock Historical Documents with AI, accessed: Jul. 19, 2023.
URL `https://readcoop.eu/transkribus/`

[111] J. Chen, H. Guo, K. Yi, B. Li, M. Elhoseiny, VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18030–18040.

[112] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv preprint arXiv:2010.11929 (2020).

[113] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, M. Zhou, UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation, arXiv preprint arXiv:2002.06353 (2020).

[114] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, C. Feichtenhofer, VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, arXiv preprint arXiv:2109.14084 (2021).

[115] SpeakAI, Turn Your Language Data into Insights, Fast and with No Code, accessed: Jul. 19, 2023.
URL `https://speakai.co/`

[116] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., Deep Speech: Scaling Up End-to-End Speech Recognition, arXiv preprint arXiv:1412.5567 (2014).

[117] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[118] Apple: Siri, accessed: Jul. 19, 2023. [link].
URL `http://www.apple.com/ios/siri/`

[119] XiaoIce, accessed: Jul. 19, 2023. [link].
URL `https://www.xiaoice.com/`

[120] Google: Google Assistant., accessed: Jul. 19, 2023. [link].
URL `https://assistant.google.com/`

[121] Amazon Inc.: Alexa, accessed: Jul. 19, 2023. [link].
URL `https://developer.amazon.com/public/solutions/alexa/`

[122] L. Zhou, J. Gao, D. Li, H.-Y. Shum, The Design and Implementation of XiaoIce, an Empathetic Social Chatbot, Computational Linguistics 46 (1) (2020) 53–93.

[123] R. Batish, Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger, Packt Publishing Ltd, 2018.

[124] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, D. Niyato, Generative AI for Semantic Communication: Architecture, Challenges, and Outlook, arXiv preprint arXiv:2308.15483 (2023).

[125] H. Du, G. Liu, D. Niyato, J. Zhang, J. Kang, Z. Xiong, B. Ai, D. I. Kim, Generative AI-aided Joint Training-free Secure Semantic Communications via Multi-modal Prompts, arXiv preprint arXiv:2309.02616 (2023).

[126] C. Xu, M. B. Mashhadi, Y. Ma, R. Tafazolli, J. Wang, Generative Semantic Communications with Foundation models: Perception-error Analysis and Semantic-aware Power Allocation, IEEE Journal on Selected Areas in Communications (2025).

[127] L. Qiao, M. B. Mashhadi, Z. Gao, C. H. Foh, P. Xiao, M. Bennis, Latency-aware Generative Semantic Communications with Pre-trained Diffusion Models, IEEE wireless communications letters 13 (10) (2024) 2652–2656.

[128] L. Qiao, M. B. Mashhadi, Z. Gao, R. Tafazolli, M. Bennis, D. Niyato, Token Communications: A Large Model-driven Framework for Cross-modal Context-aware Semantic Communications, IEEE Wireless Communications 32 (5) (2025) 80–88.

[129] Y. Zhong, A Theory of Semantic Information, China Communications 14 (1) (2017) 1–17. doi:10.1109/CC.2017.7839754.

[130] A. Kolchinsky, D. H. Wolpert, Semantic Information, Autonomous Agency and Non-equilibrium Statistical Physics, Interface Focus 8 (6) (2018) 20180041.

[131] A. Rényi, On Measures of Entropy and Information, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Vol. 4, University of California Press, 1961, pp. 547–562.

[132] R. Carnap, Y. Bar-Hillel, An Outline of a Theory of Semantic Information, Journal of Symbolic Logic 19 (3) (1954) 230–232. doi:10.2307/2268645.

[133] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, J. A. Hendler, Towards A theory of Semantic Communication, in: 2011 IEEE Network Science Workshop, IEEE, 2011, pp. 110–117.

[134] A. Karapantelakis, P. Alizadeh, A. Alabassi, K. Dey, A. Nikou, Generative AI in Mobile Networks: A Survey, Annals of Telecommunications (2023) 1–19.

[135] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, L. Sun, A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT, arXiv preprint arXiv:2303.04226 (2023).

[136] J. Wu, W. Gan, Z. Chen, S. Wan, H. Lin, AI-Generated Content (AIGC): A Survey, arXiv preprint arXiv:2304.06632 (2023).

[137] M. Xu, H. Du, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, A. Jamalipour, D. I. Kim, V. Leung, et al., Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services, arXiv preprint arXiv:2303.16129 (2023).

[138] P. Jiang, C.-K. Wen, S. Jin, G. Y. Li, Deep Source-Channel Coding for Sentence Semantic Transmission With HARQ, IEEE Transactions on Communications 70 (8) (2022) 5225–5240. doi:10.1109/TCOMM.2022.3180997.

[139] M. Sana, E. C. Strinati, Learning Semantics: An Opportunity for Effective 6G Communications, in: 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), 2022, pp. 631–636. doi:10.1109/CCNC49033.2022.9700645.

[140] H. Xie, Z. Qin, A Lite Distributed Semantic Communication System for Internet of Things, IEEE Journal on Selected Areas in Communications 39 (1) (2021) 142–153. doi:10.1109/JSAC.2020.3036968.

[141] F. Zhou, Y. Li, X. Zhang, Q. Wu, X. Lei, R. Q. Hu, Cognitive Semantic Communication Systems Driven by Knowledge Graph, in: ICC 2022-IEEE International Conference on Communications, IEEE, 2022, pp. 4860–4865.

[142] S. Guo, Y. Wang, S. Li, N. Saeed, Semantic Importance-Aware Communications Using Pre-Trained Language Models, IEEE Communications Letters (2023).

[143] F. Zhao, Y. Sun, L. Feng, L. Zhang, D. Zhao, Enhancing Reasoning Ability in Semantic Communication through Generative AI-Assisted Knowledge Construction, IEEE Communications Letters (2024).

[144] H. Yoo, T. Jung, L. Dai, S. Kim, C.-B. Chae, Real-time Semantic Communications with a Vision Transformer, in: 2022 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2022, pp. 1–2.

[145] D. Huang, X. Tao, F. Gao, J. Lu, Deep Learning-Based Image Semantic Coding for Semantic Communications, in: 2021 IEEE Global Communications Conference (GLOBECOM), 2021, pp. 1–6. doi:10.1109/GLOBECOM46510.2021.9685667.

[146] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, A. Fernando, Wireless End-to-End Image Transmission System using Semantic Communications, IEEE Access (2023).

[147] D. Huang, F. Gao, X. Tao, Q. Du, J. Lu, Toward Semantic Communications: Deep Learning-Based Image Semantic Coding, IEEE Journal on Selected Areas in Communications 41 (1) (2023) 55–71. doi:10.1109/JSAC.2022.3221999.

[148] C.-H. Lee, J.-W. Lin, P.-H. Chen, Y.-C. Chang, Deep Learning-Constructed Joint Transmission-Recognition for Internet of Things, IEEE Access 7 (2019) 76547–76561. doi:10.1109/ACCESS.2019.2920929.

[149] X. Kang, B. Song, J. Guo, Z. Qin, F. R. Yu, Task-Oriented Image Transmission for Scene Classification in Unmanned Aerial Systems, IEEE Transactions on Communications 70 (8) (2022) 5181–5192. doi:10.1109/TCOMM.2022.3182325.

[150] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. Niyato, Y. Li, Personalized Saliency in Task-Oriented Semantic Communications: Image Transmission and Performance Analysis, IEEE Journal on Selected Areas in Communications 41 (1) (2023) 186–201. doi:10.1109/JSAC.2022.3221990.

[151] A. D. Raha, M. S. Munir, A. Adhikary, Y. Qiao, C. S. Hong, Generative AI-driven Semantic Communication Framework for NextG Wireless Network, arXiv preprint arXiv:2310.09021 (2023).

[152] H. Du, G. Liu, D. Niyato, J. Zhang, J. Kang, Z. Xiong, B. Ai, D. I. Kim, Generative AI-aided Joint Training-free Secure Semantic Communications via Multi-modal Prompts, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 12896–12900.

[153] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, M. Agrawala, Text-based Editing of Talking-Head Video, ACM Transactions on Graphics (TOG) 38 (4) (2019) 1–14.

[154] P. Tandon, S. Chandak, P. Pataranutaporn, Y. Liu, A. M. Mapuranga, P. Maes, T. Weissman, M. Sra, Txt2Vid: Ultra-Low Bitrate Compression of Talking-Head Videos via Text, IEEE Journal on Selected Areas in Communications 41 (1) (2023) 107–118. doi:10.1109/JSAC.2022.3221953.

[155] Z. Weng, Z. Qin, Semantic Communication Systems for Speech Transmission, IEEE Journal on Selected Areas in Communications 39 (8) (2021) 2434–2444. doi:10.1109/JSAC.2021.3087240.

[156] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, G. Y. Li, Deep Learning Enabled Semantic Communications with Speech Recognition and Synthesis, IEEE Transactions on Wireless Communications (2023) 1–1doi:10.1109/TWC.2023.3240969.

[157] H. Tong, Z. Yang, S. Wang, Y. Hu, O. Semiari, W. Saad, C. Yin, Federated Learning For Audio Semantic Communication, Frontiers in communications and networks 2 (2021) 734402.

[158] E. Grassucci, C. Marinoni, A. Rodriguez, D. Comminiello, Diffusion Models for Audio Semantic Communication, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 13136–13140.

[159] J. M. Wing, Trustworthy AI, Communications of the ACM 64 (10) (2021) 64–71.

[160] S. A. Seshia, D. Sadigh, S. S. Sastry, Toward Verified Artificial Intelligence, Communications of the ACM 65 (7) (2022) 46–55.

[161] D. Dalrymple, Safeguarded AI: Constructing Guaranteed Safety, UK Advanced Research and Invention Agency (2024).

[162] J. Chua, Y. Li, S. Yang, C. Wang, L. Yao, AI Safety in Generative AI Large Language Models: A Survey, arXiv preprint arXiv:2407.18369 (2024).

[163] R. Inam, A. Y. Hata, V. Prifti, S. A. Asadollah, A Comprehensive Study on Artificial Intelligence Algorithms to Implement Safety Using Communication Technologies, arXiv preprint arXiv:2205.08404 (2022).

[164] K. M. Cohen, S. Park, O. Simeone, S. S. Shitz, Calibrating AI Models for Wireless Communications via Conformal Prediction, IEEE Transactions on Machine Learning in Communications and Networking 1 (2023) 296–312.

[165] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, Advances in neural information processing systems 26 (2013).

[166] K. Marino, R. Salakhutdinov, A. Gupta, The More You Know: Using Knowledge Graphs for Image Classification, arXiv preprint arXiv:1612.04844 (2016).

[167] C. Xing, J. Lv, T. Luo, Z. Zhang, Representation and Fusion Based on Knowledge Graph in Multi-modal Semantic Communication, IEEE Wireless Communications Letters (2024).

[168] C. Liang, X. Deng, Y. Sun, R. Cheng, L. Xia, D. Niyato, M. A. Imran, VISTA: Video Transmission over A Semantic Communication Approach, in: 2023 IEEE International Conference on Communications Workshops (ICC Workshops), 2023, pp. 1777–1782. doi:10.1109/ICCWorkshops57953.2023.10283754.

[169] A. J. Kumar, C. Schmidt, J. Köhler, A Knowledge Graph Based Speech Interface for Question Answering Systems, Speech Communication 92 (2017) 1–12.

[170] S. Oramas, V. C. Ostuni, T. D. Noia, X. Serra, E. D. Sciascio, Sound and Music Recommendation with Knowledge Graphs, ACM Transactions on Intelligent Systems and Technology (TIST) 8 (2) (2016) 1–21.

[171] C. Liang, Y. Sun, C. K. Thomas, L. Mohjazi, W. Saad, Semantic Communication for the Internet of Sounds: Architecture, Design Principles, and Challenges, arXiv preprint arXiv:2407.12203 (2024).

[172] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: Enhanced Language Representation with Informative Entities, arXiv preprint arXiv:1905.07129 (2019).

[173] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding, arXiv preprint arXiv:1804.07461 (2018).

[174] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders, in: Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096–1103.

[175] D. A. Huffman, A Method for the Construction of Minimum-redundancy Codes, Proceedings of the IRE 40 (9) (1952) 1098–1101.

[176] R. Gallager, Low-density Parity-check Codes, IRE Transactions on information theory 8 (1) (1962) 21–28.

[177] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, Position-aware Attention and Supervised Data Improve Slot Filling, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), 2017, pp. 35–45. URL `https://nlp.stanford.edu/pubs/zhang2017tacred.pdf`

[178] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, Advances in neural information processing systems 26 (2013).

[179] G. Zhu, C. A. Iglesias, Computing Semantic Similarity of Concepts in Knowledge Graphs, IEEE Transactions on Knowledge and Data Engineering 29 (1) (2017) 72–85. doi:10.1109/TKDE.2016.2610428.

[180] Huggingface, BERT, accessed: Jun. 29, 2024.
URL `https://huggingface.co/transformers/v3.0.2/model_doc/bert.html`

[181] F. Baader, B. Sertkaya, A.-Y. Turhan, Computing the Least Common Subsumer w.r.t. a Background Terminology, Journal of Applied Logic 5 (3) (2007) 392–420.

[182] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and Application of a Metric on Semantic Nets, IEEE transactions on systems, man, and cybernetics 19 (1) (1989) 17–30.

[183] A. Hore, D. Ziou, Image Quality Metrics: PSNR vs. SSIM, in: 2010 20th international conference on pattern recognition, IEEE, 2010, pp. 2366–2369.

[184] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[185] J. Cao, X. Weng, R. Khirodkar, J. Pang, K. Kitani, Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking, arXiv preprint arXiv:2203.14360 (2022).

[186] M. Yang, H.-S. Kim, Deep Joint Source-Channel Coding for Wireless Image Transmission with Adaptive Rate Control, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 5193–5197.

[187] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, M. Rodrigues, Wireless Image Transmission Using Deep Source Channel Coding With Attention Modules, IEEE Transactions on Circuits and Systems for Video Technology 32 (4) (2021) 2315–2328.

[188] L. Lu, R. Wu, H. Lin, J. Lu, J. Jia, Video Frame Interpolation with Transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3532–3542.

[189] S. Meister, J. Hur, S. Roth, UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[190] R. Zabih, J. Woodfill, Non-parametric Local Transforms for Computing Visual Correspondence, in: European conference on computer vision, Springer, 1994, pp. 151–158.

[191] T.-W. Hui, X. Tang, C. C. Loy, LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8981–8989.

[192] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, P. Luo, Dancetrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20993–21002.

[193] K. Corona, K. Osterdahl, R. Collins, A. Hoogs, MEVA: A Large-Scale Multiview, Multimodal Video Dataset for Activity Detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1060–1068.

[194] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, Overview of the High Efficiency Video Coding (HEVC) Standard, IEEE Transactions on circuits and systems for video technology 22 (12) (2012) 1649–1668.

[195] Z. Cai, J. Hao, P. Tan, S. Sun, P. Chin, Efficient Encoding of IEEE 802.11n LDPC Codes, Electronics Letters 42 (25) (2006) 1.

[196] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, Y.-J. A. Zhang, The Roadmap to 6G: AI Empowered Wireless Networks, IEEE Communications Magazine 57 (8) (2019) 84–90. doi:10.1109/MCOM.2019.1900271.

[197] C. Cai, X. Yuan, Y.-J. A. Zhang, Multi-Device Task-Oriented Communication via Maximal Coding Rate Reduction, arXiv preprint arXiv:2309.02888 (2023).

[198] S. Kaul, R. Yates, M. Gruteser, Real-time Status: How Often Should One Update?, in: 2012 Proceedings IEEE INFOCOM, 2012, pp. 2731–2735. doi:10.1109/INFCOM.2012.6195689.

[199] R. A. Howard, Information Value Theory, IEEE Transactions on Systems Science and Cybernetics 2 (1) (1966) 22–26. doi:10.1109/TSSC.1966.300074.

[200] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, K. B. Letaief, Semantic Communication Meets Edge Intelligence, IEEE Wireless Communications 29 (5) (2022) 28–35.

[201] G. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, et al., Semantic Communications for Artificial Intelligence Generated Content (AIGC) Toward Effective Content Creation, arXiv preprint arXiv:2308.04942 (2023).

[202] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, G. Y. Li, Robust Semantic Communications with Masked VQ-VAE Enabled Codebook, IEEE Transactions on Wireless Communications (2023) 1–1doi:10.1109/TWC.2023.3265201.

[203] G. Shi, Y. Xiao, Y. Li, X. Xie, From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems, IEEE Communications Magazine 59 (8) (2021) 44–50. doi:10.1109/MCOM.001.2001239.

[204] L. Xia, Y. Sun, D. Niyato, D. Feng, L. Feng, M. A. Imran, xURLLC-Aware Service Provisioning in Vehicular Networks: A Semantic Communication Perspective, arXiv preprint arXiv:2302.11993 (2023).

[205] X. Chen, S. Jia, Y. Xiang, A Review: Knowledge Reasoning over Knowledge Graph, Expert Systems with Applications 141 (2020) 112948.

[206] T. Vu, T. D. Nguyen, D. Q. Nguyen, D. Phung, et al., A Capsule Network-based Embedding Model for Knowledge Graph Completion and Search Personalization, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2180–2189.

[207] Y. Lin, X. Wang, H. Ma, L. Wang, F. Hao, Z. Cai, An Efficient Approach to Sharing Edge Knowledge in 5G-Enabled Industrial Internet of Things, IEEE Transactions on Industrial Informatics 19 (1) (2023) 930–939. doi:10.1109/TII.2022.3170470.

[208] H. Chai, S. Leng, Y. Chen, K. Zhang, A Hierarchical Blockchain-Enabled Federated Learning Algorithm for Knowledge Sharing in Internet of Vehicles, IEEE Transactions on Intelligent Transportation Systems 22 (7) (2021) 3975–3986. doi:10.1109/TITS.2020.3002712.

[209] Z. Cai, X. Zheng, A Private and Efficient Mechanism for Data Uploading in Smart Cyber-Physical Systems, IEEE Transactions on Network Science and Engineering 7 (2) (2020) 766–775. doi:10.1109/TNSE.2018.2830307.

[210] R. Mullins, M. T. Barros, Cognitive Network Management for 5G, Tech. rep., 5GPPP Working Group on Network Management and QoS (2017). arXiv:https://5g-ppp.eu/wp-content/uploads/2017/03/NetworkManagement$_{W}hitePaper_{1}.pdf$.

[211] T. Maksymyuk, S. Dumych, M. Brych, D. Satria, M. Jo, An IoT Based Monitoring Framework for Software Defined 5G Mobile Networks, in: Proceedings of the 11th international conference on ubiquitous information management and communication, 2017, pp. 1–4.

[212] H. Du, J. Wang, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, AI-generated Incentive Mechanism and Full-duplex Semantic Communications for Information Sharing, arXiv preprint arXiv:2303.01896 (2023).

[213] C. Zhong, J. Havlicek, LDPC Codes for Robust Transmission of Images over Wireless Channels, in: Conference Record of Thirty-Fifth Asilomar Conference on Signals, Systems and Computers (Cat. No. 01CH37256), Vol. 1, IEEE, 2001, pp. 797–800.

[214] Q. Ye, W. Shi, K. Qu, H. He, W. Zhuang, X. Shen, Joint RAN Slicing and Computation Offloading for Autonomous Vehicular Networks: A Learning-Assisted Hierarchical Approach, IEEE Open Journal of Vehicular Technology 2 (2021) 272–288.

[215] K. Qu, W. Zhuang, Q. Ye, W. Wu, X. Shen, Model-Assisted Learning for Adaptive Cooperative Perception of Connected Autonomous Vehicles, IEEE Transactions on Wireless Communications (2024).

[216] W. Zhuang, Q. Ye, F. Lyu, N. Cheng, J. Ren, SDN/NFV-Empowered Future IoV With Enhanced Communication, Computing, and Caching, Proceedings of the IEEE 108 (2) (2019) 274–291.

[217] S. Zhang, I.-L. Yen, F. Bastani, H. Moeini, D. Moore, A Semantic Model for Information Sharing in Autonomous Vehicle Systems, in: 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017, pp. 32–39. doi:10.1109/ICSC.2017.93.

[218] A. Deb Raha, M. Shirajum Munir, A. Adhikary, Y. Qiao, S.-B. Park, C. Seon Hong, An Artificial Intelligent-Driven Semantic Communication Framework for Connected Autonomous Vehicular Network, in: 2023 International Conference on Information Networking (ICOIN), 2023, pp. 352–357. doi:10.1109/ICOIN56518.2023.10049005.

[219] L. Carvalho, Smart Cities from Scratch? A Socio-technical Perspective, Cambridge Journal of Regions, Economy and Society 8 (1) (2015) 43–60.

[220] S. Rinaldi, J. Peerenboom, T. Kelly, Identifying, Understanding, and Analyzing Critical Infrastructure Interdependencies, IEEE Control Systems Magazine 21 (6) (2001) 11–25. doi:10.1109/37.969131.

[221] M. Austin, P. Delgoshaei, M. Coelho, M. Heidarinejad, Architecting Smart City Digital Twins: Combined Semantic Model and Machine Learning Approach, Journal of Management in Engineering 36 (4) (2020) 04020026.

[222] P. Antonios, K. Konstantinos, G. Christos, A Systematic Review on Semantic Interoperability in the IoE-enabled Smart cities, Internet of Things (2023) 100754.

[223] C. Wang, Y. Li, F. Gao, D. Deng, J. Xu, Y. Liu, W. Wang, Adaptive Semantic-Bit Communication for Extended Reality Interactions, IEEE Journal of Selected Topics in Signal Processing (2023) 1–13doi:10.1109/JSTSP.2023.3310654.

[224] B. Zhang, Z. Qin, Y. Guo, G. Y. Li, Semantic Sensing and Communications for Ultimate Extended Reality, arXiv preprint arXiv:2212.08533 (2022).

[225] B. Zhang, Z. Qin, G. Y. Li, Semantic Communications with Variable-Length Coding for Extended Reality, arXiv preprint arXiv:2302.08645 (2023).

[226] Y. Lin, Z. Gao, H. Du, D. Niyato, J. Kang, A. Jamalipour, X. S. Shen, A Unified Framework for Integrating Semantic Communication and AI-Generated Content in Metaverse, arXiv preprint arXiv:2305.11911 (2023).

[227] J. Park, J. Choi, S.-L. Kim, M. Bennis, Enabling the Wireless Metaverse via Semantic Multiverse Communication, in: 2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), IEEE, 2023, pp. 85–90.

[228] J. Chen, J. Wang, C. Jiang, Y. Ren, L. Hanzo, Trust-Worthy Semantic Communications for the Metaverse Relying on Federated Learning, arXiv preprint arXiv:2305.09255 (2023).

[229] Y. Zhu, X. Wu, N. Gotawala, D. M. Higdon, H. Z. Yu, Thermal Prediction of Additive Friction Stir Deposition Through Bayesian Learning-Enabled Explainable Artificial Intelligence, Journal of Manufacturing Systems 72 (2024) 1–15.

[230] X. Wu, A. Rastogi, N. Gotawala, M. A. Pandol, Y. Zhu, H. Z. Yu, Shear-driven Solid-state Additive Manufacturing of Aerospace Aluminum on Impurity Contaminated Surfaces, Materials & Design (2025) 114312.

[231] S. Y. Yetim, T. M. Duman, O. Arikan, Hidden Semi-Markov Models for Semantic-Graph Language Modeling, Journal of the Franklin Institute 361 (16) (2024) 107032.

[232] L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, Q. Wu, Adaptive Resource Allocation for Semantic Communication Networks, IEEE Transactions on Communications (2024).

[233] Osvaldo Simeone, Beyond Best Effort: How to Ensure Reliability in AI-Based Wireless Systems, accessed: Mar. 26, 2025.
URL https://www.balkancom.info/2024/presentations/0_Simeone-Keynote_BalkanCom_2024.pdf

[234] Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.

[235] L. Qiao, M. B. Mashhadi, Z. Gao, R. Tafazolli, M. Bennis, D. Niyato, Token Communications: A Unified Framework for Cross-modal Context-aware Semantic Communications (2025). arXiv:2502.12096.
URL https://arxiv.org/abs/2502.12096