Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.046837

ARTICLE

Generative Multi-Modal Mutual Enhancement Video Semantic Communications

Yuanle Chen[1], Haobo Wang[1], Chunyu Liu[1], Linyi Wang[2], Jiaxin Liu[1] and Wei Wu[1,*]

[1]The College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

[2]The College of Science, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

*Corresponding Author: Wei Wu. Email: weiwu@njupt.edu.cn

Received: 16 October 2023 Accepted: 19 December 2023 Published: 11 March 2024

ABSTRACT

Recently, there have been significant advancements in the study of semantic communication in single-modal scenarios. However, the ability to process information in multi-modal environments remains limited. Inspired by the research and applications of natural language processing across different modalities, our goal is to accurately extract frame-level semantic information from videos and ultimately transmit high-quality videos. Specifically, we propose a deep learning-based Multi-Modal Mutual Enhancement Video Semantic Communication system, called M3E-VSC. Built upon a Vector Quantized Generative Adversarial Network (VQGAN), our system aims to leverage mutual enhancement among different modalities by using text as the main carrier of transmission. With it, the semantic information can be extracted from key-frame images and audio of the video and perform differential value to ensure that the extracted text conveys accurate semantic information with fewer bits, thus improving the capacity of the system. Furthermore, a multi-frame semantic detection module is designed to facilitate semantic transitions during video generation. Simulation results demonstrate that our proposed model maintains high robustness in complex noise environments, particularly in low signal-to-noise ratio conditions, significantly improving the accuracy and speed of semantic transmission in video communication by approximately 50 percent.

KEYWORDS

Generative adversarial networks; multi-modal mutual enhancement; video semantic transmission; deep learning

1 Introduction

Over the past 30 years, wireless communication systems have undergone rapid development. From the introduction of the first-generation (1G) mobile communication system to the gradual commercial deployment of the current fifth-generation (5G) mobile communication system [1], wireless communication technology has experienced generational upgrades approximately every ten years, continuously advancing towards higher efficiency and intelligence.

In recent years, with the emergence of the concept of semantic communication, communication is no longer confined to the mere transmission of complete bit data. Semantic communication is a revolutionary communication approach based on artificial intelligence technology, aiming to achieve

more efficient, accurate, and natural information exchange through the understanding and processing of language and semantics [2]. By leveraging techniques such as natural language processing, machine learning, and deep learning, semantic communication enables communication systems to understand and analyze the meaning of language and generate appropriate responses and interactions based on this understanding. This concept has been demonstrated in many related works [3–5].

In semantic communication, textual or voice information is transformed into interpretable and processable semantic representations. Through the use of semantic analysis and inference techniques, the system can understand and extract the key elements of the information, implicitly embedding semantic knowledge in the training parameters. This deep learning-based approach enhances the efficiency and flexibility of communication, as information can be transmitted and processed more rapidly, without overreliance on extensive bandwidth.

Video communication plays a crucial role in global communication as a significant means of transmitting information. According to recent statistical data, video content constitutes a substantial portion of internet traffic, and platforms like YouTube and Netflix, due to their delivery of a large volume of video content, have contributed to a significant increase in global bandwidth consumption [6]. Similar to the importance and research value emphasized in [7] regarding real-time communication and virtual meetings through video conferencing, the importance of updating video transmission communication methods is evident.

However, as discussed in [8], traditional methods of video communication transmission predominantly rely on the transmission of audiovisual signals and depend on source-channel coding. These conventional approaches have limitations and issues concerning video quality and bandwidth constraints.

Firstly, video quality presents a significant challenge for traditional video communication. As shown in [9], Transmitting high-resolution videos requires substantial bandwidth, while network bandwidth is typically limited. This can lead to buffering, freezing, or a decrease in video quality. Applications that require high-definition imagery and smooth video, such as remote healthcare or video conferences, might experience severe user experience issues due to video quality problems. Secondly, reference [9] indicated that most wireless video transmissions employ a modular approach, and this separation-based design can lead to the so-called "cliff effect." When channel conditions cannot meet the transmission expectations, the overall communication performance of the system is significantly reduced, and the channel capacity decreases. Additionally, extra error correction processing is required to maintain the quality of video transmission.

As research deepens, although in single-modal scenarios, semantic communication has demonstrated significant potential, offering precise information transmission, including textual-related studies [10–12], image-related studies [13–15], audio-related studies [16,17], and more, the effectiveness of processing information in multi-modal settings remains limited. In multi-modal scenarios, semantic communication encounters several challenges. One of the key difficulties is achieving consistency across different modalities, ensuring that information conveyed through various channels aligns and complements each other. This necessitates the efficient integration and synchronization of textual, visual, and auditory elements to create a cohesive and meaningful communication experience [18]. Furthermore, understanding and interpreting the rich and complex information present in multiple modalities poses another challenge. It involves extracting and processing relevant features, patterns, and semantics from different modal inputs and integrating them to form a comprehensive understanding of the communicated content. Additionally, the fusion of heterogeneous modalities, each with its distinct features and representations, presents a challenge in effectively combining and leveraging

information from different sources. This requires advanced multi-modal fusion techniques, such as feature-level fusion, decision-level fusion, or semantic-level fusion, to maximize the advantages of each modality and enhance overall communication performance.

The research on multi-modal semantic communication has gained increasing attention in academia. In this paper, an end-to-end model for extracting and transmitting multi-model information is considered at the semantic level. The purpose is to accurately extract visual and audio information from video clips and transmit semantic information based on the textual modality, aiming to achieve efficient information recovery with low transmission overhead. For the multi-modal video semantic communication we are studying, the following issues are taken into account:

*Question*1 Semantic Representation: How to represent and encode the extracted information in a semantic-rich format that facilitates efficient transmission and interpretation.

*Question*2 Modality Fusion: How to effectively integrate and fuse information from different modalities, such as video, audio and text, to enhance the overall semantic understanding and transmission.

*Question*3 Multi-Modal Alignment: How to align and synchronize the different modalities to ensure coherent and meaningful communication across modalities.

*Question*4 Robustness and Adaptability: How to handle challenges such as noisy or incomplete input data and variations in input quality, and adaptability to different communication scenarios.

In light of the advancements in natural language processing across various modalities, this paper proposes a deep learning-based Multi-Modal Mutual Enhancement Video Semantic Communication model (M3E-VSC). The objective of this model is to leverage mutual enhancement among different modalities to achieve accurate semantic transmission while minimizing spectrum resource consumption. The main contributions of this paper are as follows:

- A novel M3E-VSC model is proposed, which advances the field of multi-modal communication by addressing the difficulties in multi-modal information interaction, weak correlation between audio and images, and the potential loss of semantic information during multi-modal transitions, offering a promising solution for maintaining consistency and coherence across different modalities.
- A multi-modal mutual enhancement network is designed and trained that leverages generative adversarial networks and information differential mechanisms to achieve complementarity and error correction among different modalities, optimizing the extraction of semantic information. Ultimately, the preservation of the most essential semantic features in image frames aligns with crucial semantic information from multiple modalities, thereby achieving high-quality video restoration tasks.
- Simulation results demonstrate that our proposed M3E-VSC performs well in the extraction of semantic information from multi-modal data. It exhibits significantly improved restoration performance and robustness compared to traditional communication methods under low signal-to-noise ratio transmission conditions, resulting in approximately 30% to 40% enhancement in the quality of the restored images and audio. Additionally, the overall transmission overhead of the entire communication framework is reduced by 1 to 2 orders of magnitude. These results highlight the effectiveness and robustness of our approach in realistic communication scenarios.

The remainder of this paper is organized as follows: Section 2 provides a review of the related literature on semantic communication and multi-modal information processing. Section 3 introduces the proposed M3E-VSC model, including its model architecture and implementation details. The experimental results and analysis are presented in Section 4. Finally, Section 5 concludes the paper and discusses potential directions for future research.

2 Related Work

2.1 Semantic Communication

In 1948, Weaver and Shannon first introduced the concept of semantic communication in their seminal work [19], expanding the scope of research in the field of communication beyond the transmission of traditional bit characters. With the advancements in natural language processing, semantic communication gained increased attention in the 1980s and 1990s. Building upon the foundations laid by Carnap and Bar-Hillel in their seminal work on semantic information theory in 1952, a general model for semantic communication was established by Shi et al. [20], who also introduced the concept of semantic noise in semantic encoding and decoding.

In recent years, with the advancement of artificial intelligence, semantic communication has entered a new stage. Leveraging the powerful feature extraction and modeling capabilities of large-scale deep neural networks, there has been significant progress in the quantitative analysis and communication of semantic information. The current paradigm of semantic communication involves embedding the original data into a low-dimensional space that encapsulates semantic information, aiming to achieve compression of the source information [21]. In [22], the authors proposed a joint source-channel coding framework based on natural language processing techniques, utilizing bidirectional long short-term memory (BiLSTM) or Transformer networks to extract and compress the semantic information contained in the original data. In addition, in literature [14], the receiver utilized a shared knowledge base to achieve the inverse recovery of information from the sender for implementing semantic communication. By leveraging the inference rules, the transmission of semantic information can be achieved more robustly [10].

Semantic communication has been continuously developed in the field of single modality, primarily focusing on text, image, or audio modalities. In the context of text modality, researchers have been devoted to utilizing natural language processing and text encoding techniques to achieve the transmission and understanding of semantic information. By transforming text data into vector representations or utilizing pre-trained language models for encoding, semantic information can be extracted from text and transmitted. Weng et al. [23] proposed a semantic communication approach by designing an end-to-end speech recognition system called DeepSC-SR. This system aims to learn and extract semantic features related to text while exploring robust models to cope with complex environments. Yan et al. [24] introduced the concept of semantic spectral efficiency (S-SE) as an optimization criterion for resource allocation in text semantic communication. They optimized the allocation of resources in terms of the number of transmitted semantic symbols. In practical applications of text semantic communication, Xu et al. [25] proposed a collaborative semantic-aware architecture to propagate basic semantics from collaborating users to servers, aiming to reduce data traffic. To evaluate the advantages of the proposed architecture, they presented a case study of vehicle image retrieval tasks in an intelligent transportation system (ITS).

In the field of image modality, researchers have focused on transforming images into compact representations using image encoding and feature extraction techniques to facilitate the transmission of semantic information. Common approaches include using convolutional neural networks (CNNs)

to extract image features or mapping images to low-dimensional representations in a latent space. Huang et al. [26] defined semantic concepts of image data, including categories, spatial arrangements, and visual features, as representation units. They proposed a convolutional neural network (CNN) for extracting semantic concepts and a semantic encoder based on a generative adversarial network (GAN) that incorporates attention modules to fuse local and global features. Bourtsoulatze et al. [7] introduced a joint source-channel coding (JSCC) technique for wireless image transmission. They designed an autoencoder that parameterizes the encoder and decoder functions through two CNNs. Zhang et al. [27] proposed a novel neural network-based semantic communication system for image transmission, employing a dynamic training process for transfer learning.

For practical applications, Sun et al. [28] presented a new paradigm for image transmission in the context of aerial scene classification. Their approach focused on semantic block transmission for images and channel-condition-aware perception on the frontend unmanned aerial vehicle (UAV). They employed deep reinforcement learning (DRL) to explore the contributions of the optimal semantic blocks to the backend classifier under various channel conditions. These advancements in image semantic communication showcase various techniques and architectures that aim to efficiently transmit and utilize semantic information in the context of image data.

In the field of audio research, semantic communication has become a key technology for achieving more intelligent human-machine interaction and speech communication. In their studies, Tong et al. [29] focused on wireless network semantic communication based on audio. To extract semantic information from audio signals, they proposed a CNN-based auto-coder based on the wav2vec architecture, which enabled high-precision audio transmission with a small amount of data. Following this, Han et al. [30] proposed a novel DL-based end-to-end transceiver that can extract and encode semantic information from the input speech spectrogram at the transmitter and output the corresponding transcriptions of decoded semantic information at the receiver. For speech-to-speech transmission, they further introduced a connectionist temporal classification (CTC) alignment module to extract a small amount of additional speech-related but semantically irrelevant information to better reconstruct the speech signal on the receiver for the original text. Overall, research on semantic communication in the audio modality primarily focuses on speech emotion recognition, semantic understanding and generation, and speech synthesis. Accurately extracting semantic information and achieving more natural and realistic speech generation remain important areas for further investigation.

In addition to studying the semantic context of information in different modalities, improving the efficiency of semantic transmission in practical communication settings is also an important research direction. In [31], Wang et al. proposed an adaptive semantic-bit communication structure based on resource efficiency enhancement for extended reality (XR), in which part of the XR users employ semantic communication, while others employ the conventional way. By employing a paradigm selection scheme based on the Signal-to-Interference-plus-Noise Ratio (SINR) and a power allocation algorithm based on genetic algorithms, they utilized adaptive communication and power allocation to maximize the achievable system-level performance indicated by the effective semantic rate.

Furthermore, Wang et al. in [32] built a bidirectional caching task model to achieve enhanced computation through caching. They proposed a content popularity-based Deep Q-Network (CP-DQN) algorithm to make caching decisions. Subsequently, CP-DQN was extended to the cache-computation coordination optimization algorithm (CCCA) to achieve the tradeoff of computing, caching, and communication (3C) resources. This work is highly beneficial for the practical communication process of semantic transmission as it reduces computational costs while improving the

utilization of edge resources. Although our work does not focus on resource optimization in real transmission scenarios, the proposed model architecture by the authors provides valuable insights for optimizing our approach.

2.2 Multi-Modal Information Processing

The research in the field of multi-modal information processing aims to investigate and enhance methods and techniques for effectively handling diverse types of information, such as text, images, and speech, to improve the processing of information from multiple sensory modalities. Multi-modal information processing involves capturing diverse types of information from multiple sensors or data sources, such as images, speech, text, etc., and aims to extract and integrate meaningful information from these different modalities. In recent years, numerous scholars and research teams have extensively studied multi-modal information processing and proposed various methods and techniques to address its challenges.

Deep learning-based methods have shown great potential in multi-modal information processing. For instance, Xin et al. [33] presented a multi-modal fusion network that combines visual and textual information for image classification tasks, achieving improved accuracy compared to using individual modalities alone. In addition, they proposed an interpretable deep multi-modal fusion (DMFusion) framework based on deep canonical correlation analysis (CCA). The DMFusion framework employs CCA loss to leverage inter-modal correlations and optimizes multi-modal information in a low-dimensional latent fusion space through reconstruction loss and cross-entropy loss, effectively incorporating both within-modal structure and discriminative information. Similarly, Wu et al. [34] proposed a multi-modal attention network that captures the interactions between visual and textual modalities for image captioning, leading to more accurate and descriptive captions. Graph-based methods have also been widely explored in multi-modal information processing. In [35], the authors introduced a graph convolutional network for joint image-text representation learning, leveraging the graph structure to model the interactions between different modalities. This approach has demonstrated promising results in tasks such as image retrieval and cross-modal matching.

In addition, research on data fusion, feature extraction, modal alignment and cross-modal inference plays a crucial role in multi-modal information processing. These techniques aim to better utilize information from multiple modalities and achieve effective integration and interaction between different modalities.

In the context of multi-modal information processing, feature extraction involves extracting the most representative and meaningful feature representations from multi-modal data. Traditional methods include handcrafted feature extractors such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG). However, with the rise of deep learning, feature extraction methods based on deep neural networks have become mainstream. For example, Zuo et al. [36] proposed a deep multi-modal fusion network for medical image classification tasks. This network can extract discriminative feature representations simultaneously from different modalities of images, thereby improving classification performance.

Modal alignment tackles the challenge of mapping data from different modalities to a shared representation space, facilitating cross-modal information transfer and fusion. One challenge in this task is the heterogeneity between different modalities, such as structural differences between images and text. To address this issue, a common approach is to use GAN for modal alignment. For instance, Zhu et al. [37] proposed a CycleGAN-based method that achieves cross-modal translation from images to text. By learning the mapping relationship between two modalities, modal alignment and cross-

modal information transfer can be achieved. In addition to CycleGAN, various other GAN networks can be used for modal alignment tasks. For example, the Unsupervised Image-to-Image Translation network enables unsupervised image translation, transforming images from one domain to another. The MUNIT (Multimodal Unsupervised Image-to-image Translation) network, on the other hand, is designed to handle multi-modal image translation tasks, allowing for mapping and conversion between multiple modalities. In [38], Esser et al. introduced Vector Quantized Generative Adversarial Network (VQGAN), which combines the effectiveness of CNN's positional biases with the expressiveness of transformers, enabling them to model and synthesize high-resolution images. Their approach is easily applicable to conditional synthesis tasks, where both non-spatial information (such as object class) and spatial information (such as segmentation) can control the generated images. Subsequently, the method of semantic-guided image generation using VQGAN+CLIP was proposed, which is also one of the important research foundations referenced in our proposed M3E-VSC.

Cross-modal inference, another important aspect of multi-modal information processing, aims to leverage multi-modal information for comprehensive and accurate reasoning and decision-making. It requires integrating information from different modalities and performing cross-modal associations and inferences. In [39], Wang et al. proposed a knowledge graph reasoning method based on multi-modal fusion, which combines information from images, text, and knowledge graphs to solve complex reasoning problems. This method improves the accuracy and robustness of inference by utilizing the complementarity between different modalities. In [40], Qin et al. proposed two transformer-based models named DeepSC-IR and DeepSC-MT, which perform image retrieval and machine translation tasks, respectively. Their groundbreaking contribution lies in the design of a novel hierarchical transformer that incorporates connections between each encoder layer and decoder layer to facilitate the fusion of multi-modal data, which greatly benefited our work.

Fine-grained image retrieval, as one of the tasks in both single-modal and cross-modal semantic communication, has received extensive attention in recent years. In [41], Ma et al. proposed a deep progressive asymmetric quantization (DPAQ) method based on causal intervention to learn compact and robust descriptions for fine-grained image retrieval tasks. Later, in [42], they further proposed a complementary part contrastive learning method for weakly supervised object co-localization in fine-grained images, representing similar parts of fine-grained objects with similar features in the feature space. They integrated local and contextual cues through self-supervised attention, channel attention, and spatial attention, to improve the discriminability of image objects. Additionally, they introduced a cross-class object complementary part contrastive learning module, which pulls similar part features closer and pushes different part features apart to recognize extractive part regions with different semantic information, alleviating confusion biases caused by co-occurrence environments within specific categories.

For multi-modal scenarios, particularly in video transmission tasks, several noteworthy research efforts have been made. In [43], Zhou et al. proposed a clean and effective framework to generate controllable talking faces, focusing on the generation of speaking facial expressions. They manipulated unaligned original facial images using a single photograph as an identity reference. The key innovation was designing implicit low-dimensional pose codes to modularize audio-visual representations, enabling the accurate generation of synchronized talking faces with controllable poses that could be influenced by other videos.

In [44] by Ji et al., they introduced Emotion Video Portraits (EVP), a system for synthesizing high-quality video portraits with vivid emotional dynamics driven by audio. They introduced cross-reconstruction emotion disentanglement techniques, breaking down audio into two decoupled spaces:

an emotion space independent of duration and a content space related to duration. Through these disentangled features, dynamic 2D emotional facial features could be inferred. Furthermore, in [45], Wang et al. designed a novel class of efficient deep joint source-channel encoding methods for end-to-end video transmission over wireless channels. This approach utilized nonlinear transformations and conditional coding architectures to adaptively extract semantic features across video frames and transmit them over wireless channels through deep joint source-channel encoding in the semantic feature domain. In [9], Jiang et al. proposed an Incremental Redundancy Hybrid Automatic Repeat reQuest (IR-HARQ) framework with a novel semantic error detector for video conferencing communication tasks. This framework aimed to enhance video communication quality by incorporating semantic error detection capabilities.

These recent studies exemplify advancements in data fusion, feature extraction, modal alignment, and cross-modal inference within the domain of multi-modal information processing. While many works are relevant to video transmission tasks in the multi-modal field, there remains a lack of research on video semantic transmission in natural scenes. This shortage primarily stems from the complexity of multi-modal information in natural scenes, presenting challenges in extracting meaningful semantics. By delving deeper into these techniques and innovating upon them, we aim to construct a semantic communication framework for video transmission tasks in natural scenes. This framework seeks to achieve accurate and comprehensive semantic information analysis and understanding of complex videos.

3 System Model and Problem Formulation

3.1 Semantic Communication Framework

As shown in Fig. 1, at the sender side, the video segments are first processed by extracting key frames. Using neural network recognition, which is represented by the "recognition" module in Fig. 1, the obtained images and audio modalities are analyzed, and a text modality is generated by incorporating them into a corpus. This generates feature vectors for the text, audio, and image modalities. Next, the system utilizes a multi-modal mutual enhancement network, referred to as the "M3E" module in Fig. 1, which utilizes VQGAN and set difference processing to enhance the multi-modal semantic information. This module generates the textual representation of the complete video semantic information and the differential information between the audio and image modalities. Then, the three-modal feature vectors of each key frame's video undergo an attention mechanism network, which separately outputs multi-frame semantic feature vectors for the three modalities. These feature vectors, denoted as audio modality H_t, text modality H_e, and image modality H_v, which exhibit high semantic continuity and modality correlation over time. Subsequently, the semantic encoding layer encodes these feature vectors using a BERT pre-training network. The encoded information is then transmitted through a physical channel, overcoming transmission obstacles such as distortion and noise. The channel encoding module and semantic encoding module subsequently decode the transmitted information, and the multi-modal information is restored using VQGAN networks. Finally, the multi-frame modal information is individually extracted and input into the audio frame processing module (AFP) and the visual frame processing module (VFP). These modules perform bit synchronization and inter-frame semantic error processing, resulting in the final reconstructed original audio and image frames and achieving video reconstruction.

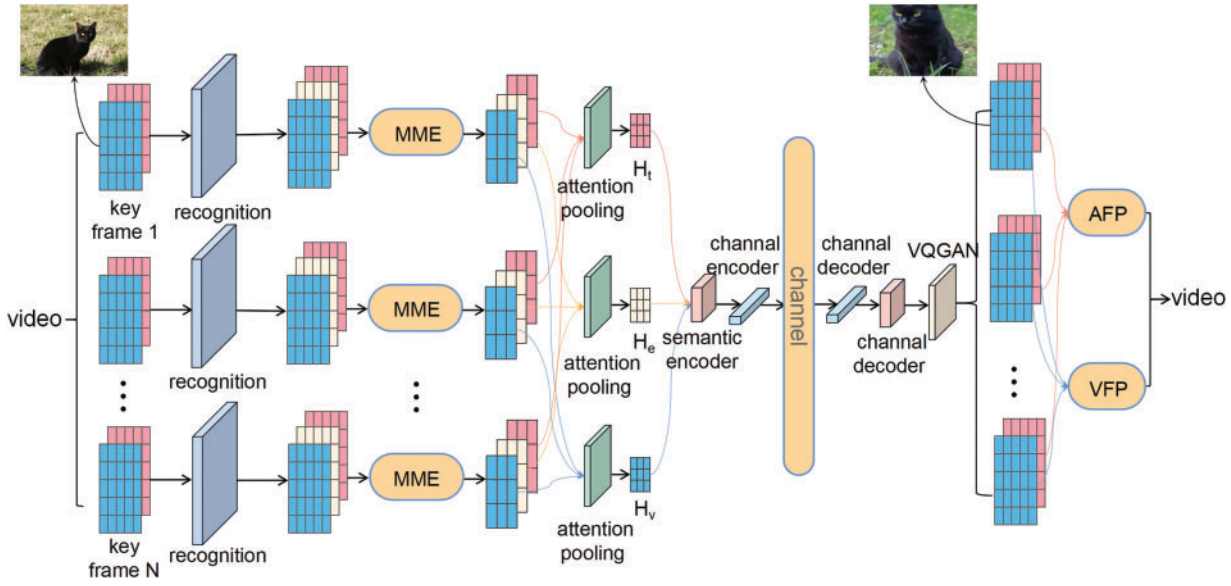

Figure 1: The proposed multi-modal mutual enhancement video semantic communication system

Specifically, to facilitate the mutual enhancement of multi-modal information, it is essential to ensure the temporal consistency of modal information within a certain time domain. In this model, the range of selected video keyframes is determined by assessing the cross-correlation coefficients, which quantify the similarity between two images, represented as f_i and f_j, and can be calculated as

$$\mathrm{C}(f_i,f_j) = \frac{\sum_{m,n}(f_i(m,n)-\mu_i)(f_j(m,n)-\mu_j)}{\sqrt{\sum_{m,n}(f_i(m,n)-\mu_i)^2 \sum_{m,n}(f_j(m,n)-\mu_j)^2}}, \tag{1}$$

where (m,n) represents the pixel coordinates, and μ_i and μ_j are respectively the mean values of the images f_i and f_j. The numerator of the equation calculates the sum of the pixel-wise product of the difference between the pixel value of each image and its mean value. The denominator normalizes the result by dividing the square root of the sum of squared differences between the pixel values and their mean values for each image in it.

To filter keyframes based on a similarity threshold, the code snippet can be used as

$$\mathrm{n}(f_k) = \{f_i | f_i \in F, f_j \in F, \mathrm{C}(f_i,f_j) < T\}, \tag{2}$$

where F represents the set of all frames in the video. The function $\mathrm{C}(f_i,f_j)$ measures the similarity between frames f_i and f_j as mentioned above. If the similarity between f_i and all other frames in F is below the specified threshold, f_i is treated as a selected keyframe and added to the set selected_frames.

To achieve more precise modal conversion with text as the transmission backbone, the semantic information is extracted and processed from the image and audio data of each selected keyframe to be sent. Additionally, the key frame information is subjected to further processing to enable target recognition for image and audio modalities. Due to the dependence of target detection algorithms in the image and audio modalities on different databases, the detection results between the two modalities will inevitably exhibit slight discrepancies. Besides, the confidence levels for the same targets may also vary. Therefore, to comprehensively extract semantic information from video frames, hyperparameters are set to strike a balance in the textual semantics of the target objects and generate the final

detection results. By incorporating text predicates using the BERT corpus, complete descriptive text representations are constructed as $\mathbf{e_i} = [o_1, o_2, \ldots, o_n]$, where $\mathbf{o_i}$ is the text semantic targets extracted from each frame modality. The key frame texts are combined to form a textual sequence represented as $\mathbf{E} = [\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}]$.

Next, multi-modal information is extracted from the key frames of the video. The objective of this system is to simulate end-to-end transmission using text as the backbone. Therefore, how to efficiently leverage the interaction and error correction among different modalities, specifically image and audio, to improve the accuracy of information extraction is a problem that needs to be addressed. To accomplish this, a multi-modal mutual enhancement network is proposed to eliminate redundancy in the multi-modal information and enhance the complementarity among the modalities. The semantic information obtained from the multi-modal mutual enhancement network can be denoted as

$$T' = \mathbf{H}_T(T, E, V), \tag{3}$$

where T' represents the transformed output of the audio modality after undergoing the multi-modal mutual enhancement network. Similarly, E' and V' denote the transformed outputs for the text and image modalities, respectively.

However, extracting semantics from individual key frames of a video does not guarantee the quality of the final video reconstruction. Therefore, following the multi-modal mutual enhancement network, a multi-frame verification mechanism is employed to ensure the correlation between different frames. Specifically, a Recurrent Neural Network (RNN) is utilized as the foundational framework, integrated with attention networks and pooling operations. This mechanism verifies and normalizes the multi-modal information from multiple frames to achieve consistency in semantic information across video frames.

In this framework, the cross-entropy function is adopted as the loss function to measure the difference in multi-frame multi-modal information. Each frame has multiple modal features represented as X_i. For each modal feature, there are corresponding labels and reference information Y_i. The cross-entropy loss function can be defined as

$$\mathscr{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{3} Y_{ij} \log\left(P_{ij}\right), \tag{4}$$

where Y_{ij} represents the label or reference information for the j th class of the i th modality, and P_{ij} is the predicted probability of the j th class for the i th modality.

The loss function calculates the difference between the predicted probabilities and the corresponding labels for each modality and averages the differences across all modalities. Minimizing the loss function encourages our model to learn more accurate representations and predictions for multi-frame multi-modal information.

After verifying and correcting the multi-frame semantic information, the multi-modal semantic information of the obtained keyframes is processed into tensors, as shown in Fig. 1, and finally, a concatenated tensor is obtained, denoted as y. In the M3E-VSC framework, the transmission of semantic information is simulated using the established semantic communication framework. On the transmitting side, semantic encoders and channel encoders are designed to perform the transmission of semantic information in the physical channel, ensuring the integrity of semantics. In this model, the processed information x is inputted into the pre-trained BERT model, and the encoded semantic

symbols can be represented as

$$x' = \mathscr{C}_{\beta}(\mathscr{B}_{\alpha}(x)), \tag{5}$$

where $\mathscr{B}_{\alpha}(\cdot)$ represents the semantic encoder based on the pre-trained BERT model with parameter set 'α', and $\mathscr{C}_{\beta}(\cdot)$ represents the channel encoder with parameter set 'β'. For the collection of multiple frame texts and residual data x to be transmitted, the process of simulating transmission can be represented as

$$y = hx' + w, \tag{6}$$

where h represents the fading coefficient, which is a random variable, and its probability density function (PDF) can be represented by the Rayleigh distribution, and w represents the Additive White Gaussian Noise (AWGN). The Probability Density Function (PDF) of the fading coefficient can be expressed as the Rayleigh distribution and is given by

$$f(h) = \frac{h}{\sigma^2} e^{-\frac{h^2}{2\sigma^2}}, \tag{7}$$

where σ is the scale parameter of the fading channel, determining the fading severity. The fading coefficient h takes non-negative real values. This study simulates and controls the signal-to-noise ratio (SNR) by taking into account the dual effects of fading gain and fading loss in the fading channel. This allows us to examine the stability and performance of M3E-VSC in complex environments.

On the receiver side, a semantic decoder and a channel decoder are designed. The decoding process, which is the inverse of the encoding process, can be represented as

$$S = \mathscr{C}_{\gamma}^{-1}(\mathscr{B}_{\delta}^{-1}(y)), \tag{8}$$

where S represents the decoded information, $\mathscr{C}_{\gamma}(\cdot)$ is the semantic decoder, $\mathscr{B}_{\delta}(\cdot)$ is the channel decoder, and γ and δ denote their respective parameters.

It is worth noting that our model replaces traditional bit symbols with semantic symbols as the main form of transmission. Therefore, semantic similarity is introduced to measure the quality of transmission. The formula of this measure can be expressed as

$$\xi = \frac{B(s)B(\hat{s})^T}{||B(s)||||B(\hat{s})||}, \tag{9}$$

where ξ represents the semantic similarity, which is a continuous value ranging from 0 to 1. $B(s)$ and $B(\hat{s})$ denote the mapped representations of the original semantic symbol s and the decoded semantic symbol $\hat{s}$ at the receiving end. respectively, $||B(s)||$ and $||B(\hat{s})||$ represent the norms of $B(s)$ and $B(\hat{s})$. For each frame of transmitted information, after decoding by the decoder, the minimum semantic similarity threshold $\xi_{\min}$ is set to 0.6. Frames with a similarity below $\xi_{\min}$ are flagged by the system and retransmitted to ensure the quality of semantic transmission.

When our model successfully decodes the received semantic symbols at the receiving end, generating high-quality videos accurately becomes our primary concern. By combining the multi-frame semantic detection mechanism and applying smooth processing, the restoration of video using frames of images and audio is successfully achieved Throughout the entire process, the spatial and temporal characteristics of videos, along with the continuity and consistency of semantic information, play a vital role in supervising the generation of videos.

Accurately generating high-quality videos based on the decoded semantic symbols presents a significant challenge. To overcome this challenge, a multi-frame semantic detection mechanism is

employed, which incorporates object detection and tracking algorithms to identify semantic objects in the video and extract their position and motion information. Specifically, assuming that images and audio corresponding to key-frame semantic text have been generated using a VQGAN network, denoted as I_t for a total of T frames, where t represents the frame index. By utilizing object detection and tracking algorithms, the target positions and bounding box information for each frame are obtained and represented as B_t.

To achieve smoother transitions between frames and enhance the temporal continuity of the video, a smooth processing technique is introduced. By leveraging the obtained smoothed semantic object positions and motion information, the information can be utilized for image restoration. A memory-based approach is employed in the multi-frame image semantic detection to ensure consistency of semantic representations within a specific range, resulting in a final generated video closely resembling a real video. Additionally, optical flow estimation is utilized to estimate motion vectors between frames, which are then used to generate intermediate frames through interpolation techniques. Assuming the goal is to restore the image I_t for the t-th frame, the interpolated frame I_t^{interp} can be calculated using the motion vector V_t obtained from optical flow estimation. This process can be expressed as follows:

$$I_t^{interp} = \textbf{Interpolate}(I_{t-1}, I_{t+1}, V_t), \tag{10}$$

where **Interpolate**$(\cdot)$ represents the interpolation function, which interpolates based on the motion vector and the images of adjacent frames to generate intermediate frames.

In the case of audio processing, a similar approach is employed as that of image frames. Filters are utilized to enhance audio quality, and interpolation techniques are employed to ensure temporal continuity. By applying filters, the audio signal can be improved by reducing noise or adjusting the frequency response, thereby enhancing its overall quality. Furthermore, interpolation techniques are used to maintain the temporal coherence of the audio frames. As a result, the complete video segment, encompassing both the visual and audio components, is successfully restored.

3.2 Multi-Modal Mutual Enhancement Network

As shown in Fig. 2, the aim is to achieve precise information extraction from the determined keyframes by utilizing the interaction and error correction among different modalities. To achieve this, a multi-modal mutual enhancement network is proposed, which eliminates redundancy in multi-modal information and enhances the complementarity among modalities. Specifically, the network consists of three modules: an encoder, a Vector Quantized Generative Adversarial Network (VQGAN), and an information difference processing module. The network takes audio, text descriptions, and image modalities as inputs. Each modality is processed by VGG, BERT, and Transformer networks, respectively, to obtain their discrete representations. These representations then pass through the encoder layer to generate the three-modal information. The network introduces an information difference algorithm to obtain the output, which includes the main text information T_i' as well as the audio difference information E_i' and image difference information V_i' of this key frame i.

During the network input processing, encoding is applied to both the image and audio modalities to reduce the dimensionality of their features while preserving their semantic information. This dimensionality reduction simplifies the subsequent information processing. Within our multi-modal semantic transmission framework, the interaction between text, image, and audio modalities is intricate. Through the dimensionality reduction operation of the encoder, the semantic information of each modality can be extracted using pre-trained models. Specifically, the VGG network is employed for image processing, and the Transformer network is used for audio processing. Utilizing operations

such as convolution, pooling, and sequential layers, self-encoding operations are achieved on both the image and audio modalities.

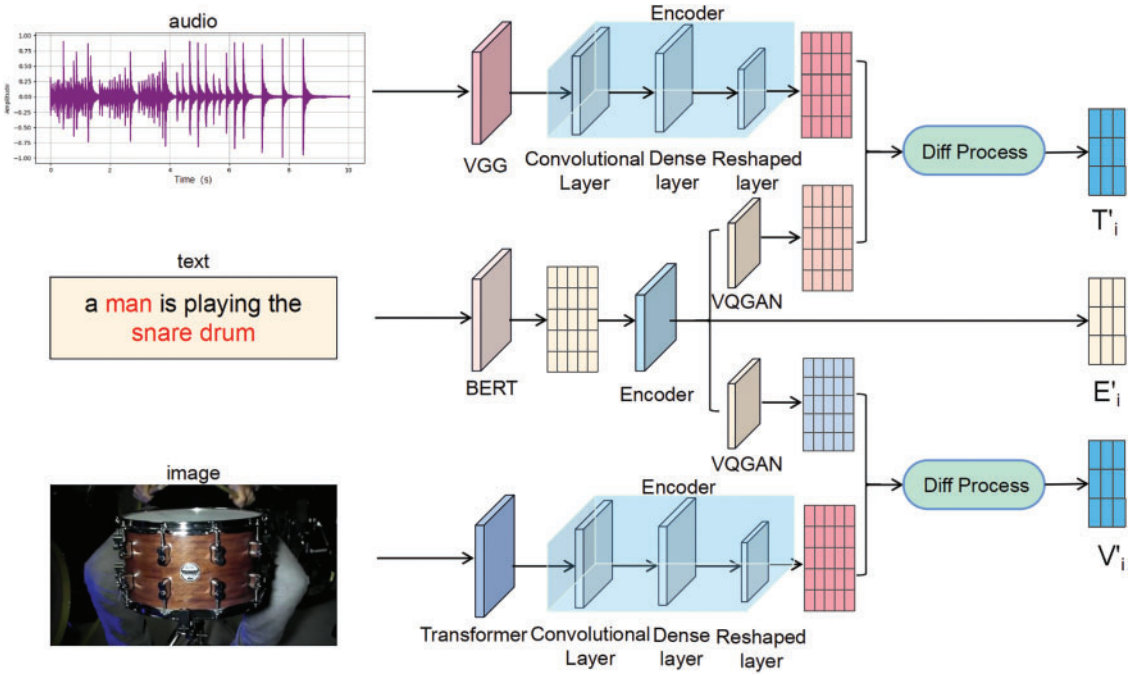

Figure 2: The proposed multi-modal information processing network

In the mutual enhancement network, the primary approach for modality conversion is the application of VQGAN, enabling the transfer of semantic information during the conversion process. In the proposed overall framework, the model obtains descriptive textual information (E) by recognizing the audio and image modalities of video keyframes. This information is a comprehensive text generated by the pre-trained BERT model under the unified condition of multi-modal semantic correction. To map the descriptive text to specific feature subspaces, the text is embedded into a VQGAN network, leveraging pre-trained models to accomplish the mapping from text to the audio and image modalities.

Specifically, two separate VQGAN networks are independently trained: one for mapping text to audio and another for mapping text to image. Firstly, a VQGAN network is trained for the text-to-audio modality conversion task. The generator takes an input text description and generates the corresponding audio waveform, while the discriminator aims to distinguish between the generated audio waveform and the real audio waveform. Secondly, for the text-to-image modality conversion task, another independent VQGAN network is trained. The generator in this network receives a text description as input and generates the corresponding image, while the discriminator differentiates between the generated images and real images. Through adversarial training of the generator and discriminator, the generator gradually learns to generate realistic audio and images that closely resemble the real ones. The optimization problems for the two VQGAN networks can be represented as follows:

$$\min_{G} \max_{D} \mathscr{L}(G, D), \tag{11}$$

where D refers to the discriminator and G represents the generator. The loss function $\mathscr{L}(D, G)$ can be expressed as

$$\mathscr{L}(G, D) = \mathbb{E}\mathbf{x} \sim p_{\mathbf{data}}(\mathbf{x})[\log D(\mathbf{x})] + \mathbb{E}\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})[\log(1 - D(G(\mathbf{z})))], \tag{12}$$

where $\mathbf{x}$ represents the real samples drawn from the true data distribution $p_{\mathbf{data}}(\mathbf{x})$, $\mathbf{z}$ represents the noise samples drawn from a prior distribution $p_{\mathbf{z}}(\mathbf{z})$, $G(\mathbf{z})$ denotes the generated samples from the generator, and $D(\mathbf{x})$ represents the output of the discriminator for real samples.

In our proposed approach, modality conversion relies on the information description provided by the text, which may result in information loss or redundancy during processing. Therefore, in the third part of our multi-modal mutually-enhancing network, an information difference processing module is included, which is shown in Fig. 2 as the "diff process" module, to automatically learn and process the difference between the generated vectors by the VQGAN network and the original modalities. Following the principle of information processing, a dynamic network is also designed to perform the differential processing of cross-modal features and real modality features, aiming to obtain information that cannot be reflected in modality conversion or remove redundant information.

As illustrated in Fig. 2, the true modality information of audio and image is denoted as T and V, respectively, while E represents the descriptive text. The difference set information obtained after the information difference processing is denoted as T' and V'. The processing procedure can be summarized as

$$T' = T - \hat{T}_s \cdot \Theta, \tag{13}$$

$$V' = V - \hat{V}_s \cdot \Gamma, \tag{14}$$

$$\Theta = \textbf{Attention}\ (E, \hat{T}_s), \tag{15}$$

$$\Gamma = \textbf{Attention}\ (V, \hat{V}_s), \tag{16}$$

where $\hat{T}_s$ and $\hat{V}_s$ represent the intermediate result of text-to-image and text-audio modality conversion in the VQGAN network, Θ is the similar weights of T' and T, while Γ represents the same relation between V' and V, which is learned automatically by our designed attention mechanism network. They capture the relationship and similarity between the processed information and the original information. By adjusting the values of Θ and Γ, the network can emphasize or downplay certain aspects of the information, allowing for more effective information difference processing. For the attention mechanism network, the attention function is defined to calculate the weights, which can be expressed as

$$\text{Attention}\ (\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{17}$$

where $\mathbf{Q}$ represents the semantic information of the images generated by VQGAN, $\mathbf{K}$ represents the semantic information of the original audio and images, and $\mathbf{V}$ represents the value vector. The dimension of the key vector is denoted by d_k. The equation performs a dot product operation between the query vector and the key vector and scales the attention using the scaling factor $\frac{1}{\sqrt{d_k}}$. It then applies the softmax function to compute the attention weights. Finally, the attention weights are multiplied by the value vector to obtain the attention-weighted result.

Through the application of semantic difference processing and attention mechanism, our multi-modal mutual enhancement network can extract and modulate semantic differences in information

across text, image, and audio modalities, enabling more effective modality conversion and semantic processing. Ultimately, the network outputs a sentence vector combining foundational textual semantics with the supplemented semantic information obtained through differential processing. This output is used to complete the intended task.

Algorithm 1: Multi-Modal Mutual Enhancement Network Training Algorithm.

Initialization

1. Initialize the image encoder network VGG with pre-trained weight V;
2. Initialize the audio encoder network WaveNet with pre-trained weight W;
3. Input: Text modality E, Image modality V, audio modality T.

Function: Training for text-image and text-audio VQGAN network

1. Initialize the noise vector λ, episode T;
2. Compute the encoding of real images using the picture encoder network VGG;
3. Compute the encoding of real images using the audio encoder network WaveNet;
4. for episode $t = 1 \rightarrow$ T do
5. Train the VQGAN network for text → image, using weight V of VGG;
6. Train the VQGAN network for text → audio, using weight W of WaveNet;
7. Output text → image network $R(\cdot)$, text → audio network $Y(\cdot)$;
8. end for.

Function: Training for multi-modal mutual enhancement network

1. Initialize text → image network $R(\cdot)$, text → audio network $Y(\cdot)$, episode T, batch size M;
2. randomly initialize parameters λ;
3. Input: original information E, V, T of video samples in the training set D;
4. for episode $t = 1 \rightarrow$ T do
5. Sample M examples from training set D;
6. Computer the loss functions $L_M(E, V, T, R(E), Y(E))$;
7. Generating gradient ∇ of the network and update parameters λ;
8. end for.

Output: The trained multi-modal mutual enhancement network.

3.3 Semantic Encoder and Decoder Design

The semantic encoding and decoding module is of paramount importance in the framework while designing a semantic communication system using text as the transmission basis. The descriptive text obtained after processing with the multi-modal mutual enhancement network undergoes tokenization as the first step. Tokenization involves splitting the text into individual sub-word units, such as Word-piece or Sentence-piece, which serve as the basic units for the BERT model. After tokenization, special tokens are added to the beginning and between sentences of the tokenized sequence. The special token "[CLS]" (standing for classification) is added at the beginning of the sequence, while the special token "[SEP]" is inserted between each pair of sentences. Assuming the descriptive text is represented as $X = x_1, x_2, ..., x_n$, the tokenized sequence can be denoted as

$$X_{\text{tokenized}} = \{[\text{CLS}], x_1, x_2, ..., x_n, [\text{SEP}]\}. \tag{18}$$

This tokenization process breaks down the text into smaller meaningful units, enabling the BERT model to process and understand the semantics at a more granular level. Then, position embeddings are added to each word vector to retain the positional information of words within the sentence. After tokenization and position embedding, the processed text is input into the BERT pre-trained model.

In the encoding process, the self-attention mechanism is utilized in each encoder layer to capture the relationships and dependencies among different words in the input sequence. This mechanism allows the model to focus on relevant words and their interactions, enabling it to build contextual representations. To achieve this, the self-attention mechanism calculates attention weights for each token in the input sequence. These attention weights determine the importance of each word concerning the others in the sequence. The mechanism is "self-attention" because it computes the weights based on the similarity of each word to all other words in the sequence.

The self-attention mechanism in BERT is specifically referred to as "multi-head attention" which means that BERT utilizes multiple parallel self-attention heads, each capturing different aspects of contextual information and relationships. The multiple heads allow the model to attend to various patterns and dependencies simultaneously, enhancing its ability to understand complex contexts. For a given input sequence $X = x_1, x_2, ..., x_n$, the output of the self-attention mechanism can be represented as $H = h_1, h_2, ..., h_n$, where h_i represents the representation of the i th token after incorporating information from all other tokens in the sequence. Each h_i carries contextual information from the entire input sequence, thanks to the ability of the self-attention mechanism to effectively gather information from all words in the sequence based on their relevance to each other.

After the self-attention mechanism, each word vector is further transformed through a feed-forward neural network to introduce non-linearity. Specifically, for the input sequence X, the output of the feed-forward neural network can be represented as

$$\mathrm{FFN}(X) = \mathbf{ReLU}(XW_1 + b_1)W_2 + b_2, \tag{19}$$

where ReLU is the rectified linear unit activation function, W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors. This feed-forward transformation enables the model to capture more complex relationships and dependencies between words in the sequence.

Afterward, the pooling operation is performed on the text data, using the output of the "[CLS]" token as the representation of the entire text. This representation is considered as a summarized and compressed semantic vector representation of the input sequence. Mathematically, it can be expressed as

$$\lambda = \mathbf{Pool}\ (\text{BERT Output}[[CLS]]), \tag{20}$$

where BERT Output[[CLS]] refers to the output of the BERT model corresponding to the "[CLS]" token, and $\mathbf{Pool}(\cdot)$ denotes the pooling operation, which aggregates the information from the entire sequence into a single vector, capturing the overall semantic meaning of the text to be transmitted.

The semantic decoding process involves designing a reverse Transformer decoder to convert the semantic vector into a text sequence. By generating the next word iteratively, the complete text sequence is gradually constructed. In each step of word generation, the decoder model utilizes its output and the historical context to predict the next word. This can be represented by the function as

$$P(y_i|y_{i+1}, ..., y_m, z) = \mathbf{f}(y_{i+1}, ..., y_m, z), \tag{21}$$

where y_i represents the i th generated word, $y_i + 1, ..., y_m$ denotes the previously generated prefix sequence, and z represents the semantic vector.

4 Experimental Results

4.1 Simulation Settings

Due to the utilization of multiple neural networks in our communication framework, each neural network is specialized for specific tasks and operates on distinct datasets. For the audio and image recognition modalities at the receiver end, we train them on the FSD50K and MSCOCO datasets, respectively. Audio segments are uniformly sampled at a rate of 16 kHz. Textual descriptions undergo processing using a pre-trained BERT model, incorporating predicate connection labels to create complete sentences. For instance, the label "plane, fly, sky" is transformed into the sentence "a plane is flying in the sky." In the evaluation of audio modality tasks, separate simulation experiments are conducted to compare the generative model against traditional compression methods such as AAC, FLAC, PCM, and MP3, especially in complex channel environments. For the proposed multi-modal mutual enhancement network, we perform classification filtering on the VGG-sound audio-visual dataset. It consists of 34 different categories of objects, and the test set and experimental set are split in a 2:8 ratio. Each video sample is defined with a resolution of 720 × 480, a length of 10 seconds, and a target frame rate of 30 fps. The learning rate is set to 0.01.

4.2 Simulation Results

In the simulation experiments, various evaluation metrics are set to assess the performance of the model from multiple aspects. These metrics primarily include the transmitted bit rate, text recovery effectiveness (semantic similarity metric), image recovery structural similarity index (SSIM) metric, image object detection accuracy, and audio recovery timbre similarity.

Fig. 3 showcases the restored videos by M3E-VSC at the receiver end for different video tasks, with detection categories being zebra, snare drum, cat, and train. Each transmission task is achieved through the joint processing of audio and image information using the multi-modal mutually beneficial enhancement network.

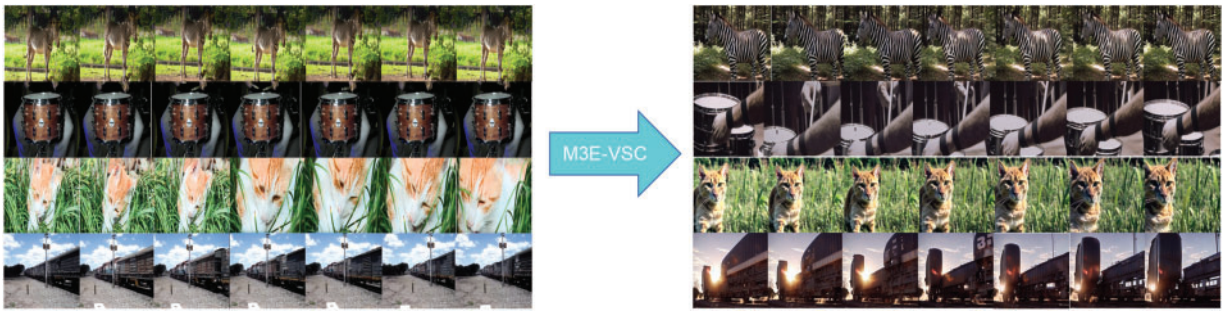

Figure 3: Presentation of multi-modal mutual enhancement video semantic communication system

In the following sections, a comparison is made between our model and some traditional video and image transmission methods. In particular, for the primary comparison schemes, an H.265 video compression encoder [46] is used for source encoding, employs low-density parity-check codes (LDPC) coding for channel encoding, and utilizes quadrature amplitude modulation (QAM) modulation. Additionally, we also take into account different LDPC encoding modes at varying bit rates, providing a comprehensive comparison of metrics.

In Fig. 4, the graph displays the bit sizes required for each Gop (Group of Pictures) in various video transmission schemes under AWGN channel conditions. Signal-to-Noise Ratio (SNR) is employed to assess changes in channel transmission conditions. This comparison specifies that

each scheme operates at a maximum efficient frame rate of 30 fps with a fixed-quality video, resulting in a noticeable decline in metrics as SNR increases. Notably, the model presented here, which utilizes multi-modal mutually beneficial enhancement for semantic processing rather than pixel-level compression, substantially reduces the necessary bit size during transmission compared to traditional approaches. This underscores that our communication model can allocate more substantial transmission bandwidth for video communication, thereby significantly enhancing communication efficiency.

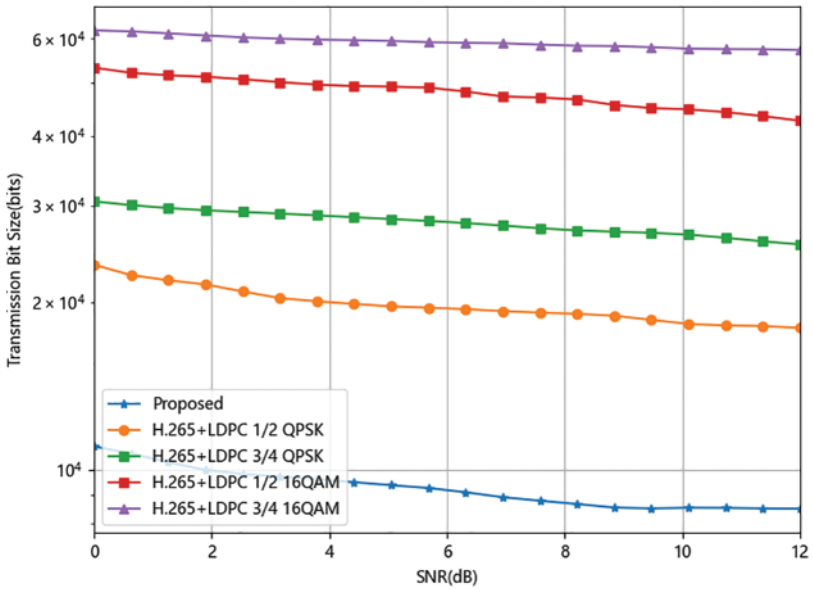

Figure 4: Comparison of bit rates for single-frame image compression and transmission between M3E-VSC and other compression methods in AWGN channel

Fig. 5 examines the transmission performance of our proposed semantic model, comparing it with several traditional text compression methods as baseline models including Huffman, 5-bit, and Brotli. The experimental results show that compared to these traditional compression models, the semantic transmission model exhibits approximately 30% to 40% performance improvement in restoring both text and its additional information, especially under low signal-to-noise ratio conditions. This highlights the robust anti-interference capability and stability of the semantic encoding and decoding approach.

Fig. 6 investigates the restoration performance of our proposed model on single-frame video images, focusing on the requirements imposed on our generation model in complex transmission environments. The experiment includes several baseline schemes, including H.265 + LDPC 3/4 QPSK, H.265 + LDPC 3/4 16 QAM, and JPEG + 16 QAM. The experimental results demonstrate that our generation model exhibits significant improvement in the restoration of image frames compared to other schemes, particularly in low signal-to-noise ratio transmission environments. This improvement can be attributed to the reception of relatively complete textual semantic information by our model. As the signal-to-noise ratio increases, the advantage of our proposed model gradually diminishes. However, the SSIM metric remains relatively stable, indicating the consistent performance of our approach.

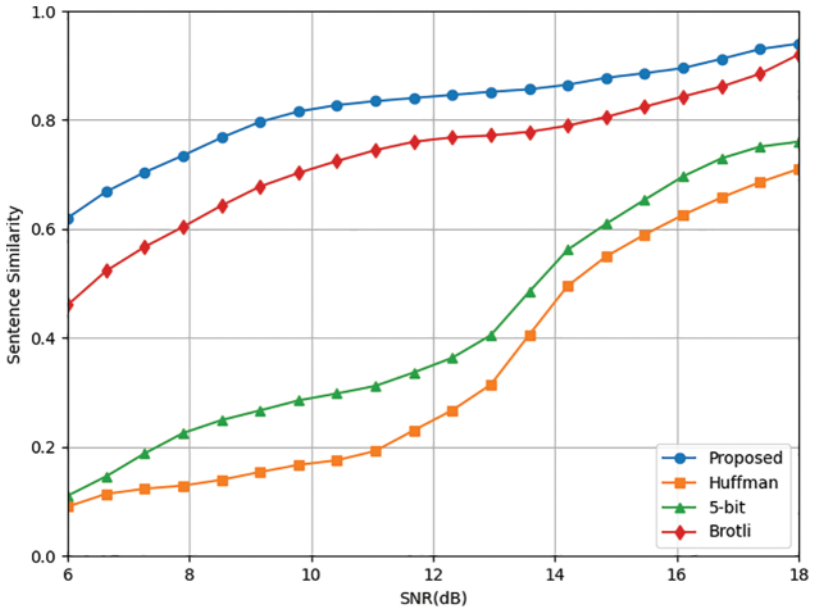

Figure 5: Comparison of similarity in recovering transmitted text at the receiver between M3E-VSC and other conventional text compression methods in AWGN channel

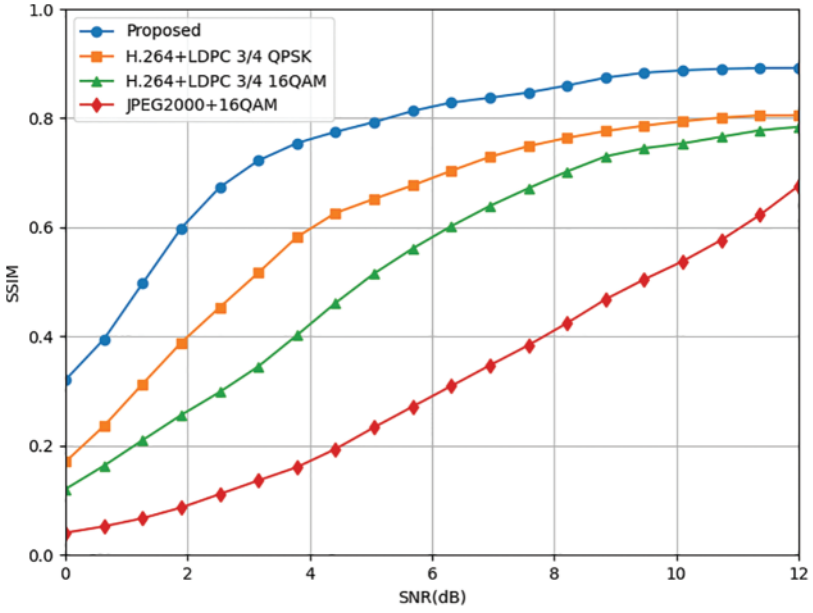

Figure 6: Comparison of SSIM index between M3E-VSC and other image compression and transmission methods for recovered images at the receiver in AWGN channel

Fig. 7 evaluates the preservation of semantic information in video frame generation based on object detection accuracy. The baseline schemes, H.265 + LDPC 3/4 QPSK, H.265 + LDPC 3/4 16 QAM, represent a comparison with traditional image compression and transmission methods. DeepSC: text and DeepSC: image represent cases where only textual semantic information and only

image semantic information are transmitted, respectively, using the classical semantic text transmission model, DeepSC. The simulation results demonstrate that our communication model outperforms other schemes regarding image object detection within the specified signal-to-noise ratio range. This advantage is primarily attributed to the error-correcting capabilities of the multi-modal mutual enhancement network in our proposed model, which preserves semantic information effectively.

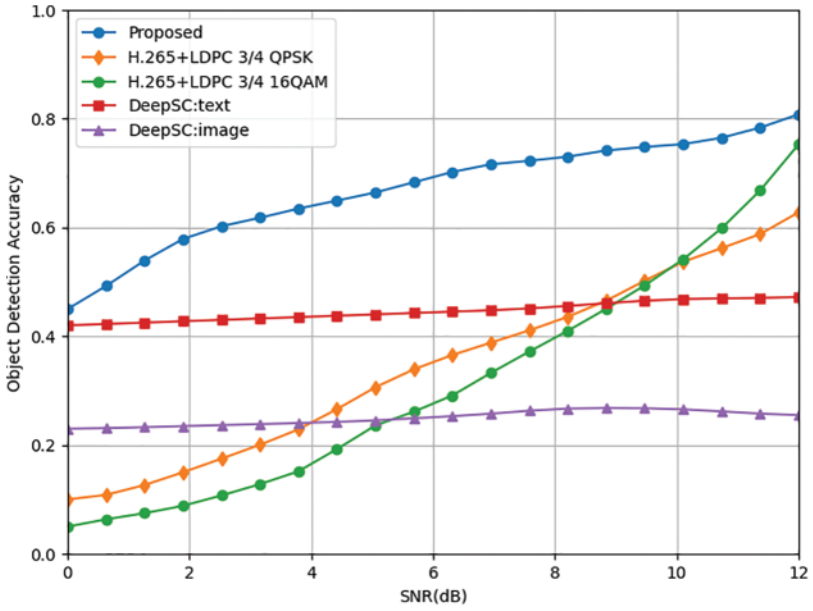

Figure 7: Comparison of object detection accuracy scores between M3E-VSC and traditional image compression methods for recovered images at the receiver in AWGN channel. Here, DeepSC: text and DeepSC: image represent conventional semantic transmission methods that solely transmit text information and image information, respectively

Table 1 presents the results of the ablation experiment. We evaluate the models based on the object recognition scores of video transmission image frames. The YOLOv6 model is used for the detection task, and multiple ablation models are compared with the traditional approach. The evaluated models include:

- **Ours:** This refers to our proposed multi-modal mutually enhanced semantic communication model. It incorporates multiple modalities to improve video semantic communication performance.
- **Ours (without M3E):** This corresponds to our proposed semantic communication model without the M3E component. It serves as a comparison to assess the impact of the M3E scheme.
- **Ours (only text):** This corresponds to our proposed semantic communication model without the semantic-based disjoint network component. In this model, only the text feature vectors are encoded for semantic purposes at the sending end, without involving multi-modal semantic processing tasks or the M3E component. It serves as a comparison to assess the impact of the semantic-based disjoint network on our overall approach.
- **Traditional communications:** This represents traditional semantic communication methods, specifically referring to the results of H.265 video compression and transmission techniques.

Table 1: Results of ablation experiments

Model (SNR = 8)	Car	Dog	Zebra	Airplane	Train	Umbrella	Class avg.
Ours	0.748	0.712	0.723	0.560	0.682	0.528	0.659
Ours (without M3E)	0.572	0.631	0.476	0.520	0.651	0.516	0.561
Ours (only text)	0.421	0.372	0.423	0.460	0.312	0.378	0.394
Traditions	0.313	0.267	0.210	0.132	0.215	0.180	0.219

The table presents evaluation metrics for different classes such as Car, Dog, Zebra, Airplane, Train, Umbrella, Chair, and the overall class average. For each class, we conduct simulations using 30 different video clips with a signal-to-noise ratio of 8 dB. The goal is to study the accuracy of object detection in keyframe images for the video transmission task and assess the importance of our proposed semantic-based disjoint network and multi-modal mutual enhancement network in the overall system.

The results indicate that our complete model outperforms the two ablation models and significantly surpasses the traditional H.264 model in terms of image semantic transmission, specifically under low signal-to-noise ratio conditions. The comparison among the three scenarios highlights the substantial effectiveness of our semantic-based disjoint network and multi-modal mutual enhancement network in preserving image reconstruction quality and enhancing semantic information. It is worth noting that, for certain classes such as Car, Dog, and Zebra, there is a noticeable difference between the "Ours (without M3E)" model and the complete model. This divergence is due to the pronounced audio modality features of these classes, which are greatly influenced by the semantic optimization of the multi-modal mutual enhancement network. Furthermore, the relatively poorer image reconstruction performance of the "Ours (only text)" model, compared to the first two comparison scenarios, can be attributed to the absence of the image semantic representation module at the sending end, resulting in the loss of important image feature information.

Fig. 8 investigates the restoration performance of our approach for audio. The model primarily focuses on transmitting audio semantics and does not specifically emphasize detailed audio restoration. The main goal is to achieve similar semantic representations in the audio. To assess this, the timbre similarity metric is used as a comparison scheme to determine whether the audio maintains similar semantic information. Several baseline schemes are considered, including traditional audio compression methods like AAC, FLAC, PCM, and MP3. The simulation results indicate that the approach exhibits significant performance advantages over the baseline schemes in low signal-to-noise ratio transmission environments, resulting in better audio restoration effects.

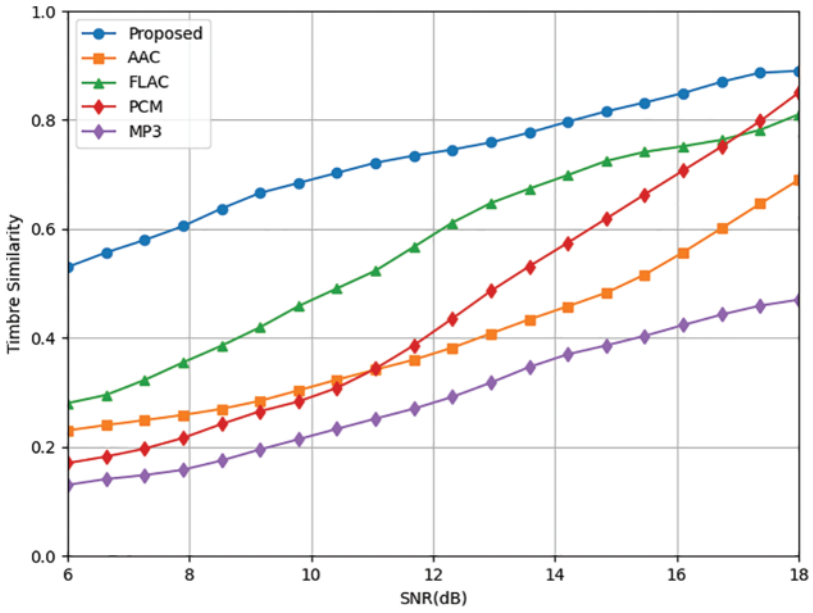

Figure 8: Comparison of timbre accuracy scores between M3E-VSC and traditional audio compression methods for recovered audio at the receiver in AWGN channel

5 Conclusion

In this paper, we proposed a multi-modal mutual enhancement video semantic communication system called M3E-VSC, which leveraged computer vision and natural language processing techniques to achieve more effective and efficient transmission and understanding of semantic information in videos. Additionally, our framework incorporated advanced machine learning algorithms, such as deep neural networks, to facilitate the fusion and integration of multi-modal information. By jointly optimizing the encoding, decoding, and video restoration processes, we enhanced the overall efficiency and effectiveness of video semantic communication. Simulation results demonstrated that our multi-modal mutual enhancement approach exhibited a significant advantage in terms of transmission volume while ensuring performance in semantic content transmission and comprehension. By leveraging the complementary nature of different modalities, our framework achieved higher semantic accuracy and robustness in conveying video content.

Acknowledgement: We would like to thank the reviewers who remained anonymous for their constructive criticism and recommendations. We also thank the journal, CMES, for their support for the publication of this Special Issue. The authors are grateful for the support by the National Key Research and Development Project and the National Natural Science Foundation of China.

Funding Statement: This work was supported by the National Key Research and Development Project under Grant 2020YFB1807602, Key Program of Marine Economy Development Special Foundation of Department of Natural Resources of Guangdong Province (GDNRC[2023]24) and the National Natural Science Foundation of China under Grant 62271267.

Author Contributions: The authors confirm their contribution to the paper as follows: write the main manuscript text and perform the experiments: Yuanle Chen; conduct research and design the basic

system framework: Haobo Wang, Chunyu Liu; collect data and assist in the training process of deep learning: Linyi Wang; assist in analysis and interpretation of results: Jiaxin Liu; give guidance and suggestions for method of this article, as well as reviewed and edited the manuscript: Wei Wu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All the reviewed research literature and used data in this research paper consists of publicly available scholarly articles, conference proceedings, books, and reports. The references and citations are contained in the reference list of this manuscript and can be accessed through online databases, academic libraries, or by contacting the respective publishers.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Kojima, F., Matsumura, T. (2021). NICT's R&D activities on the future terrestrial wireless communication systems toward B5G/6G by harmonizing requirements with environments. *2021 IEEE VTS 17th Asia Pacific Wireless Communications Symposium (APWCS)*, Osaka, Japan.
2. Luo, X., Chen, H. H., Guo, Q. (2022). Semantic communications: Overview, open issues, and future research directions. *IEEE Wireless Communications, 29(1),* 210–219.
3. Wheeler, D., Natarajan, B. (2023). Engineering semantic communication: A survey. *IEEE Access, 11,* 13965–13995.
4. Qin, Z., Tao, X., Lu, J., Tong, W., Li, G. Y. (2022). Semantic communications: Principles and challenges. arXiv:2201.01389.
5. Xie, H., Qin, Z., Li, G. Y. (2022). Task-oriented multi-user semantic communications for VQA. *IEEE Wireless Communications Letters, 11(3),* 553–557.
6. Cheng, X., Liu, J., Dale, C. (2013). Understanding the characteristics of internet short video sharing: A youtube-based measurement study. *IEEE Transactions on Multimedia, 15(5),* 1184–1194.
7. Mehdi, M., Ala, A. F., Sameh, S., Mohsen, G. (2018). Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys Tutorials, 20(4),* 2923–2960.
8. Li, Y., Zhou, Z. (2020). Subjective video quality assessment and the analysis of coding strategies in video communication scene. *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Chengdu, China.
9. Jiang, P., Wen, C. K., Jin, S., Li, G. Y. (2023). Wireless semantic communications for video conferencing. *IEEE Journal on Selected Areas in Communications, 41(1),* 230–244.
10. Peng, X., Qin, Z., Huang, D., Tao, X., Lu, J. et al. (2022). A robust deep learning enabled semantic communication system for text. *GLOBECOM 2022–2022 IEEE Global Communications Conference*, Rio de Janeiro, Brazil.
11. Wang, Y., Chen, M., Luo, T., Saad, W., Niyato, D. et al. (2022). Performance optimization for semantic communications: An attention-based reinforcement learning approach. *IEEE Journal on Selected Areas in Communications, 40(9),* 2598–2613.
12. Jiang, P., Wen, C. K., Jin, S., Li, G. Y. (2022). Deep source-channel coding for sentence semantic transmission with HARQ. *IEEE Transactions on Communications, 70(8),* 5225–5240.
13. Huang, X., Chen, X., Chen, L., Yin, H., Wang, W. (2021). A novel convolutional neural network architecture of deep joint source-channel coding for wireless image transmission. *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, Changsha, China.
14. Kurka, D. B., Gündüz, D. (2020). DeepJSCC-*f* Deep joint source-channel coding of images with feedback. *IEEE Journal on Selected Areas in Information Theory, 1(1),* 178–193.

15. Dai, J., Wang, S., Tan, K., Si, Z., Qin, X. et al. (2022). Nonlinear transform source-channel coding for semantic communications. *IEEE Journal on Selected Areas in Communications, 40(8),* 2300–2316.
16. Pedrelli, L., Hinaut, X. (2022). Hierarchical-task reservoir for online semantic analysis from continuous speech. *IEEE Transactions on Neural Networks and Learning Systems, 33(6),* 2654–2663.
17. Wang, H., Zha, Z. J., Li, L., Chen, X., Luo, J. (2023). Semantic and relation modulation for audio-visual event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6),* 7711–7725.
18. Luo, X., Gao, R., Chen, H. H., Chen, S., Guo, Q. et al. (2022). Multi-modal and multi-user semantic communications for channel-level information fusion. *IEEE Wireless Communications*, pp. 1–18.
19. You, X., Wang, C., Jie, H., Gao, X., Zhang, Z. (2021). Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts. *Science China Information Sciences, 64(1),* 1–74.
20. Shi, G., Gao, D., Song, X., Chai, J., Yang, M. et al. (2021). A new communication paradigm: From bit accuracy to semantic fidelity. arXiv:2101.12649.
21. Kountouris, M., Pappas, N. (2021). Semantics-empowered communication for networked intelligent systems. *IEEE Communications Magazine, 59(6),* 96–102.
22. Zhou, F., Li, Y., Zhang, X., Wu, Q., Lei, X. et al. (2022). Cognitive semantic communication systems driven by knowledge graph. *ICC 2022–IEEE International Conference on Communications*, Seoul, Korea.
23. Weng, Z., Qin, Z., Tao, X., Pan, C., Liu, G. et al. (2023). Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications, 22(9),* 6227–6240.
24. Yan, L., Qin, Z., Zhang, R., Li, Y., Li, G. Y. (2022). Resource allocation for text semantic communications. *IEEE Wireless Communications Letters, 11(7),* 1394–1398.
25. Kang, X., Song, B., Guo, J., Qin, Z., Yu, F. R. (2022). Task-oriented image transmission for scene classification in unmanned aerial systems. *IEEE Transactions on Communications, 70(8),* 5181–5192.
26. Huang, D., Gao, F., Tao, X., Du, Q., Lu, J. (2022). Towards semantic communications: Deep learning-based image semantic coding. *IEEE Journal on Selected Areas in Communications, 41(1),* 55–71.
27. Zhang, H., Shao, S., Tao, M., Bi, X., Letaief, K. B. (2023). Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data. *IEEE Journal on Selected Areas in Communications, 41(1),* 170–185.
28. Sun, Q., Guo, C., Yang, Y., Chen, J., Tang, R. et al. (2022). Deep joint source-channel coding for wireless image transmission with semantic importance. *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, London, UK.
29. Tong, H., Yang, Z., Wang, S., Hu, Y., Saad, W. et al. (2021). Federated learning based audio semantic communication over wireless networks. *2021 IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain.
30. Han, T., Yang, Q., Shi, Z., He, S., Zhang, Z. (2023). Semantic-preserved communication system for highly efficient speech transmission. *IEEE Journal on Selected Areas in Communications, 41(1),* 245–259.
31. Wang, C., Li, Y., Gao, F., Deng, D., Xu, J. et al. (2023). Adaptive semantic-bit communication for extended reality interactions. *IEEE Journal of Selected Topics in Signal Processing, 17(5),* 1080–1092.
32. Wang, C., Yu, X., Xu, L., Wang, Z., Wang, W. (2023). Multimodal semantic communication accelerated bidirectional caching for 6G MEC. *Future Generation Computer Systems, 140,* 225–237.
33. Xin, B., Huang, J., Zhou, Y., Lu, J., Wang, X. (2021). Interpretation on deep multimodal fusion for diagnostic classification. *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China.
34. Wu, H., Shao, Y., Mikolajczyk, K., Gündüz, D. (2022). Channel-adaptive wireless image transmission with OFDM. *IEEE Wireless Communications Letters, 11(11),* 2400–2404.
35. Zhang, Q., Xu, Z., Liu, H., Tang, Y. (2021). KGAT-SR: Knowledge-enhanced graph attention network for session-based recommendation. *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, Washington DC, USA.

36. Zuo, Q., Zhang, J., Yang, Y. (2021). DMC-fusion: Deep multi-cascade fusion with classifier-based feature synthesis for medical multi-modal images. *IEEE Journal of Biomedical and Health Informatics, 25(9),* 3438–3449.
37. Zhu, J. Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy.
38. Esser, P., Rombach, R., Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA.
39. Wang, J., Bao, B. K., Xu, C. (2022). DualVGR: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia, 24,* 3369–3380.
40. Xie, H., Qin, Z., Tao, X., Letaief, K. B. (2022). Task-oriented multi-user semantic communications. *IEEE Journal on Selected Areas in Communications, 40(9),* 2584–2597.
41. Ma, L., Hong, H., Meng, F., Wu, Q., Wu, J. (2023). Deep progressive asymmetric quantization based on causal intervention for fine-grained image retrieval. *IEEE Transactions on Multimedia*, pp. 1–13.
42. Ma, L., Zhao, F., Hong, H., Wang, L., Zhu, Y. (2023). Complementary parts contrastive learning for fine-grained weakly supervised object co-localization. *IEEE Transactions on Circuits and Systems for Video Technology, 33(11),* 6635–6648.
43. Zhou, H., Sun, Y., Wu, W., Loy, C. C., Wang, X. et al. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA.
44. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C. C. et al. (2021). Audio-driven emotional video portraits. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA.
45. Wang, S., Dai, J., Liang, Z., Niu, K., Si, Z. et al. (2023). Wireless deep video semantic transmission. *IEEE Journal on Selected Areas in Communications, 41(1),* 214–229.
46. Ohm, J. R., Sullivan, G. J. (2013). High efficiency video coding: The next frontier in video compression [standards in a nutshell]. *IEEE Signal Processing Magazine, 30(1),* 152–158.