Engineering 56 (2026) 163–172

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
6G: From Theory to Practice—Article

Generative Video Communications: Concepts, Key Technologies, and Future Research Trends

Wenjun Zhang [a], Guo Lu [a,*], Zhiyong Chen [a], Geoffrey Ye Li [b]

[a] School of Information Science and Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK

ARTICLE INFO

Article history:
Received 9 March 2025
Revised 24 May 2025
Accepted 9 June 2025
Available online 19 June 2025

Keywords:
Video communications
Video compression
Video transmission
Video evaluation

ABSTRACT

With the rapid growth of video traffic and the evolution of video formats, traditional video communication systems are encountering many challenges, such as limited data compression capacity, high energy consumption, and a narrow range of services. These challenges stem from the constraints of current systems, which rely heavily on discriminative methods for visual content reconstruction and achieve communication gains only in the information and physical domains. To address these issues, this paper introduces generative video communication, a novel paradigm that leverages generative artificial intelligence technologies to enhance video content expression. The core objective is to improve the expressive capabilities of video communication by enabling new gains in the cognitive domain (i.e., content dimension) while complementing existing frameworks. This paper presents key technical pathways for the proposed paradigm, including elastic encoding, collaborative transmission, and trustworthy evaluation, and explores its potential applications in task-oriented and immersive communication. Through this generative approach, we aim to overcome the limitations of traditional video communication systems, offering more efficient, adaptable, and immersive video services.

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Since the advent of third-generation (3G) wireless communications, video communication has evolved into a cornerstone service. The International Telecommunication Union (ITU) highlights its importance and identifies immersive video as a key service scenario for future sixth-generation (6G) networks. Traditional video communication usually focuses on video encoding, transmission, display, and quality evaluation. The primary goal is to optimize key performance metrics, including spectrum utilization, transmission rate, capacity, and latency, through efficient data compression and robust transmission techniques. These technologies can be broadly classified into two categories. The first category aims to increase expression efficiency in the information domain, such as by employing advanced video compression algorithms to reduce data redundancy. The second category focuses on improving resource utilization in the physical domain, such as by leveraging multiple-input multiple-output (MIMO) technologies to fully exploit available communication resources including time, space, and frequency. However, throughout this process, key aspects related to the processing, understanding, and application of video content within the cognitive domain have been largely overlooked, leading to a significant underestimation of its communication potential in this domain.

With continued expansion and the diversification of user demands, the limitations of traditional video communication are becoming increasingly apparent. Taking video encoding technology as an example, the computational complexity has surged from H.264 to newer standards such as versatile video coding (H.266/VVC), despite ongoing advancements in compression efficiency. However, the performance gains from individual key technologies have become marginal, often less than 1%. Additionally, the power consumption of wireless network infrastructure has risen sharply, with fourth-generation (4G) base stations consuming significantly more power than previous generations—a trend that is expected to escalate with the advent of 6G technology. Furthermore, current video services are primarily limited to conventional two-dimensional (2D) streaming and on-demand viewing, which fail to meet the growing demand for personalized user experiences.

* Corresponding author.
E-mail address: luguo2014@sjtu.edu.cn (G. Lu).

https://doi.org/10.1016/j.eng.2025.06.018
2095-8099/© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

These challenges are inherent to the limitations of the existing communication framework, and incremental technological improvements alone are insufficient to address them. Therefore, paradigm-shifting innovations are essential.

The fundamental issue underlying these limitations is that current video communication systems place disproportionate emphasis on optimizing information and physical domain performance, often at the expense of the cognitive domain. More specifically, current systems conceptualize the cognitive domain of video content in an overly simplified and inflexible manner, often reducing videos to mere collections of discrete pixels. Consequently, the communication objective is narrowly confined to the accurate transmission of these pixel values, thereby disregarding the crucial semantic information embedded within the pixel data. While such an approach might be adequate for certain traditional communication modalities, such as fiber optics, where data fidelity is paramount, it is fundamentally inadequate for video communication, whose ultimate purpose is to effectively engage human cognitive processes.

Empirical findings from neuroscience provide strong support for the preceding argument regarding the overlooked potential of the cognitive domain in current communication paradigms. As demonstrated in Ref. [1], the human brain's information-processing capacity is estimated at a modest 10 bits·s^{-1}, whereas the sensory system's data input rate is on the order of 10^9 bits·s^{-1}, illustrating a significant quantitative gap. This significant difference highlights the remarkable capability of human cognition to effectively manage and distill the overwhelming volume of sensory information received. Nevertheless, contemporary video communication paradigms largely emphasize the faithful transmission of sensory data, frequently neglecting the critical cognitive stages of analysis, processing, and perceptual interpretation. Viewed from a comprehensive standpoint of video communication, this cognitive domain, acting as the terminal processing stage, holds substantial promise for advancement in areas such as bandwidth optimization and perceptual fidelity.

The rapidly advancing field of generative artificial intelligence (AI) is unlocking transformative potential for video communication, presenting a paradigm shift in how video content is created, transmitted, and experienced. Generative technologies, exemplified by variational autoencoders [2], generative adversarial networks (GANs) [3], and diffusion models [4], have revolutionized content creation and data modeling by effectively learning and replicating the inherent statistical distribution within complex datasets. These models excel not only in generating high-fidelity multimedia content—including text, audio, images, and videos—but also in driving innovations in visual data compression and related domains. In contrast to traditional video communication, which is predicated on pixel-level transmission, generative technologies fundamentally augment the representational capacity of video, facilitating a transition from a passive transmission paradigm to an active generation framework. The advanced intelligent capabilities of generative AI provide a compelling foundation for developing novel video communication paradigms, including innovative approaches to leverage and optimize the cognitive domain.

Building upon the analysis above, this paper introduces the concept of "generative video communication," with the aim of thoroughly exploring its research background, core concepts, technical pathways, and future development prospects. We first systematically outline the new paradigm of generative video communication, highlighting its core essence, technical advantages, and implementation framework. Next, we explore potential technical pathways across three dimensions: encoding, transmission, and evaluation. Finally, we summarize and predict future development directions, offering valuable insights for both theoretical research and practical applications in this field.

2. Generative video communications

The core characteristic of traditional video communication is its reliability, which has long been the primary goal of communication systems. However, a major challenge in incorporating generative AI into video communication is the inherent uncertainty of generative techniques. This suggests that video communication systems relying solely on generative AI algorithms may struggle to ensure reliability. As illustrated in Fig. 1, the AI-generated image at the decoding end may differ significantly from the original, while transmitting a small set of keywords can describe an entire image, greatly reducing storage and transmission domain. Therefore, it is crucial to carefully consider how to seamlessly integrate generative AI technology with traditional video communication systems to construct the novel paradigm of generative video communication.

Here, we propose a preliminary definition: Generative video communication involves a seamless integration of generative AI with traditional discriminative video communication, rather than completely replacing the latter, with the aim of addressing multi-dimensional constraints, such as generation quality and computational power, while fully harnessing the benefits of generative technologies. Based on the discussion above, generative video communication has the following three core characteristics:

(1) **Discriminative technology as a cornerstone.** Discriminative technology is foundational to generative video communication, ensuring compatibility, quality control, and reliability. It also offloads computation from generative AI, which improves system efficiency. Consequently, discriminative technology acts as both a constraint and a safeguard for stable and effective generative video communication.

(2) **Generative technology for enhanced expressiveness.** Generative technology significantly expands video communication's expressive capacity, potentially reducing transmission costs and evolving communication paradigms. By learning patterns in video content, it enhances discriminative methods, enabling more efficient and dynamic systems.

(3) **Evolutionary upgrade, not overhaul.** Generative video communication takes into account computational power, generation quality, and bandwidth constraints as part of an evolutionary upgrade. It maintains traditional reliability while incorporating generative innovation.

Building upon these characteristics, generative video communication can be further defined as a video communication paradigm integrated with generative AI that deeply understands and characterizes the intrinsic patterns of video. Its goal is to maximize the expressive capability of video within a given bitrate, thereby achieving more efficient and reliable video communication. The core feature of this paradigm lies in the fusion of generative and discriminative information. Discriminative information—primarily drawn from traditional video communication—ensures reliability and interpretability, while generative information enhances its expressive potential by uncovering the underlying patterns of video.

$$\max \text{Ability}, \ R \leq R_\tau \tag{1}$$

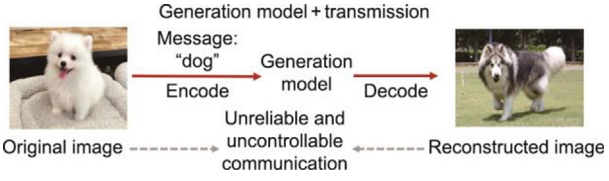

Fig. 1. Basic integration of generative models and communication.

where R denotes the bitrate consumption, constrained by bitrate limits R_τ at timestamp τ.

From the perspective of information theory, traditional discriminative video communication is a typical form of syntactic communication, with its core objective being pixel-level consistency. Generative video communication, built upon the foundation of traditional discriminative video communication, goes beyond the realm of syntactic communication, involving a composite of syntax, semantics, and pragmatics. Building on this, we can further define the capabilities of generative video communication, which encompass the following three levels:

(1) **Reconstruction capability at the syntactic level:** the ability to accurately reconstruct the original video at the pixel level, forming the foundational capability of generative video communication;

(2) **Utility capability at the semantic level:** the ability to reconstruct video content that exhibits similar effects as the original video in downstream tasks, reflecting the practical value of generative video communication;

(3) **Comprehension capability at the pragmatic level:** the ability to reconstruct video that produces the same perceptual effects as the original video at the human cognitive level, representing the highest level of generative video communication.

These three levels build progressively from concrete to abstract. The capability of generative video communication depends on the ratio of generative to discriminative information. More generative information means less data transmission, moving communication toward pragmatics but potentially reducing reconstruction fidelity. Generative video communication is categorized into three levels based on this ratio:

- **Discriminative information dominance (full-reference generation).** At this level, the typical application involves semantic transformation of the input video, such as converting a pixel-based video into a semantically enhanced version. This level aligns with elastic transmission, in which base video content is broadcasted, and personalized semantic enhancements are delivered via unicast cellular networks. The corresponding utility evaluation focuses on consistency, ensuring accurate semantic alignment with the reference content.
- **Balanced discriminative and generative information (semi-reference generation).** At this level, the typical application involves a dimensionality expansion of the input content, such as converting a 2D video into a three-dimensional (3D) video. This level is supported by scalable transmission, which dynamically adjusts the balance between discriminative and generative information according to the network conditions. The utility evaluation emphasizes reasonableness, assessing whether the generated content is coherent and adheres to physical plausibility.
- **Generative information dominance (no-reference generation).** At this level, the typical application involves modality transformation of the input signal, such as converting language or brainwave signals into video signals. This level corresponds to generative reception technologies, in which the receiver reconstructs high-quality video content using minimal control signals with the help of generative AI models. The associated utility evaluation highlights usability, evaluating the effectiveness of the generated content in specific downstream tasks or user scenarios.

Given the diversity of video content and varying network conditions, the appropriate choice among the three elastic encoding modes depends not only on the transmission bitrate but also on the semantic and structural complexity of the source video. To provide a practical guideline for mode selection, we define indicative thresholds based on bits per pixel (bpp):

- Full-reference generation is recommended when the available bitrate exceeds 0.05 bpp, allowing sufficient capacity to preserve both pixel-level and semantic features.
- Semi-reference generation is suitable for intermediate bitrate conditions between 0.005 and 0.05 bpp, where partial generative modeling supplements reference signals.
- No-reference generation is adopted under extremely low bitrate conditions (below 0.005 bpp), prioritizing semantic-level communication through compact control signals such as text.

These thresholds serve as a coarse guideline for dynamic mode switching; further research is needed to refine adaptive policies that balance compression efficiency, visual quality, and system stability.

3. Key technologies of generative video communication

The fundamental framework of generative video communication consists of three stages: encoding, transmission, and evaluation, as illustrated in Fig. 2. The key challenges are ① how to efficiently represent both discriminative information and generative information (i.e., how to optimize encoding); ② how to collaboratively transmit compressed bitstreams (i.e., how to achieve complementary cooperation); and ③ how to evaluate generated content (i.e., how to conduct personalized analysis). To address these challenges, we propose three potential technical pathways: elastic encoding, intelligent transmission, and utility evaluation.

3.1. Elastic encoding

Elastic encoding aims to address the challenge of efficiently encoding both reference and generative information. Based on the degree of integration between the sampled reference and the generative information, encoding strategies can be categorized into three levels, as mentioned above: full-reference generation, semi-reference generation, and no-reference generation.

3.1.1. Full-reference generation

Full-reference generation takes a reference video as input and produces a semantically enhanced video. This level is primarily designed for scenarios that involve video semantic enhancement for downstream tasks. Utilizing self-supervised learning methods, this framework uncovers the intrinsic characteristics of the video, generating feature information as the generative information stream. This process enables the reconstruction of a video with enhanced expressive power.

As illustrated in Fig. 3, the full-reference generative video encoding framework [5] combines the efficient content encoding capabilities of traditional video codecs with the strengths of neural networks in semantic encoding. More specifically, the original video is first compressed using a traditional video codec (e.g., H.265 [6]), resulting in a lossy video. A neural-network-based generative stream is then introduced to efficiently transmit semantic information within the video. The encoding of this generative stream leverages the video stream, aiming to reduce the overall bitrate while increasing the accuracy of semantic representation. The framework also incorporates adaptive and dynamic modeling strategies to optimize the extraction of semantic information. At the receiving end, the video and generative stream are integrated through an attention-based cross-bitstream feature fusion scheme, generating the final decoded video. This process effectively supports various machine intelligence tasks under low-bitrate conditions.

In addition to the aforementioned full-reference generative video encoding framework, several other studies have explored the use of generative models to enhance perceptual quality and semantic preservation in video compression. For instance, Yang et al. [7] proposed a perceptual learned video compression

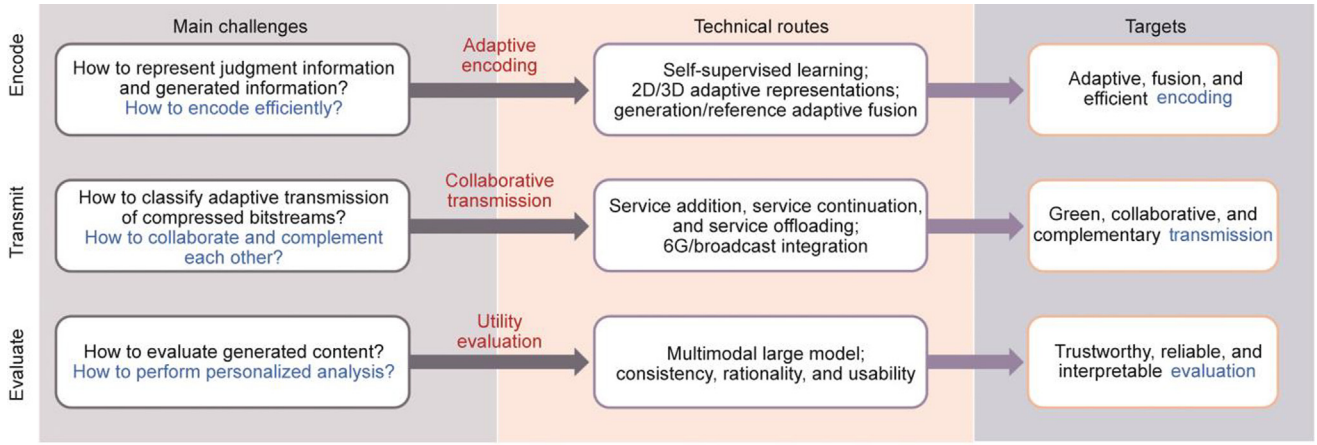

Fig. 2. Framework of generative video communication.

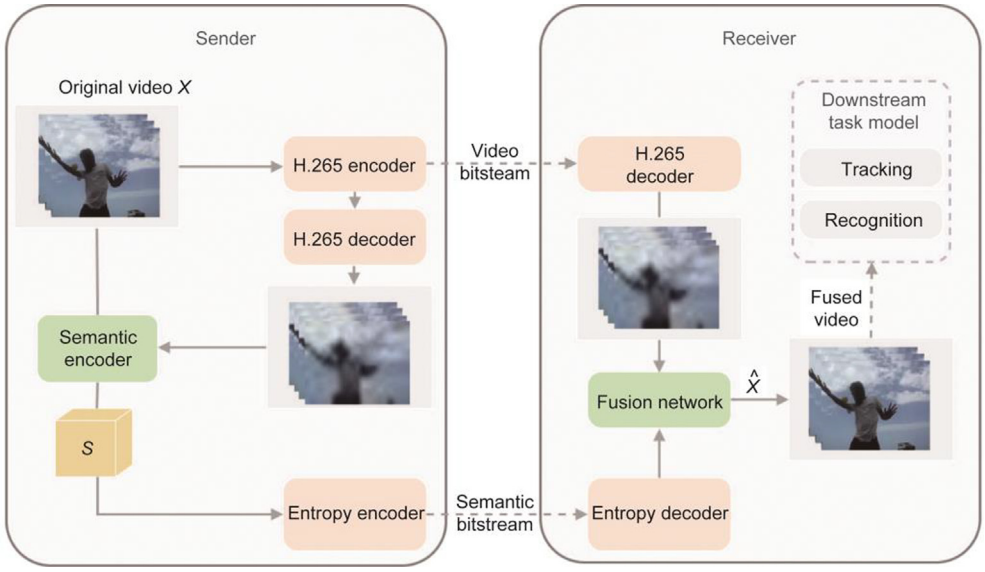

Fig. 3. An unsupervised-learning-based video encoding framework. X: the orginal video; $\hat{X}$: the fused video; S: the semantic latent representations.

method using a recurrent conditional GAN that improves visual quality by explicitly modeling temporal dependencies and optimizing for perceptual metrics. Similarly, Li et al. [8] introduced a high-visual-fidelity learned video compression system that leverages generative models to increase semantic retention, yielding more visually pleasing reconstructions at low bitrates. These works demonstrate the broader applicability of generative approaches in improving perceptual and semantic fidelity in full-reference scenarios.

3.1.2. Semi-reference generation

Semi-reference generation aims to create more immersive 3D scenes by expanding the dimensionality of the input video. This method is commonly used in multi-view video rendering, in which a single-view video serves as the reference information, while 3D implicit representations are employed to supplement the generative information, enabling the transition from 2D to 3D expansion. This approach is particularly suitable for scenarios such as virtual reality, which require high degrees of viewing freedom and flexible user experiences.

As illustrated in Fig. 4, we propose an elastic 3D joint encoding scheme [9] that combines explicit and implicit representations. In this framework, a 2D video codec based on explicit representation is first used to encode the single-view video, ensuring the reliable transmission of reference information. Then, a codec based on implicit neural representations (INRs) is employed to encode the information from the remaining views. The INR codec generates corresponding implicit reconstruction frames by using the temporal and view indices of the multi-view video as input coordinates. This encoding method efficiently captures the potential correlations between multiple views, reducing redundant information. To further enhance reconstruction quality, the framework introduces high-quality reconstruction frames from the explicit codec to compensate for inter-view information. A weighted fusion strategy is adopted to integrate the explicit representations with the implicit reconstructions generated by INRs, ultimately producing high-quality reconstruction frames. This design enables the collaborative optimization of explicit and implicit information, effectively leveraging the strengths of both approaches. Similarly, Chen et al. [10] proposed HNeRV, a hybrid neural representation that combines explicit and implicit encoding to effectively capture both spatial details and temporal dynamics in videos. This approach demonstrates strong performance in reconstructing high-quality frames while maintaining compact representations,

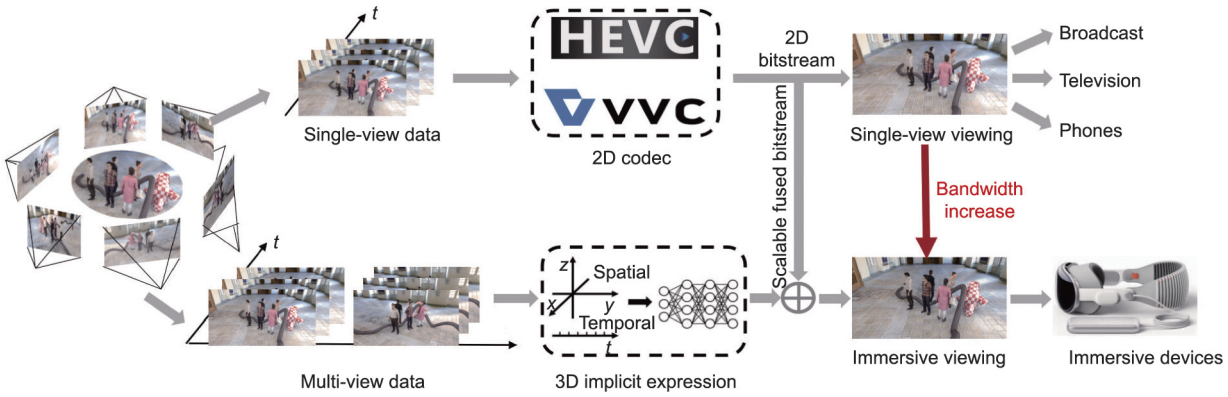

Fig. 4. An elastic 3D joint encoding scheme with explicit and implicit representations. *x* and *y*: the spatial dimensions; *z*: the depth dimension; *t*: the temporal dimension; HEVC: high efficiency video coding; VVC: versatile video coding.

further validating the effectiveness of semi-reference generation strategies based on mixed representations.

3.1.3. No-reference generation

No-reference generation focuses on modality transformation, with its core emphasis on the compact cross-modal representation of visual content. Utilizing the fundamental encoding paradigm of visual signal → textual signal → visual signal and incorporating weak visual references as generative control signals, this method enables high-quality generative video reconstruction. In particular, no-reference generation demonstrates exceptional performance in low-bitrate scenarios, being capable of generating video content with high visual quality even under extremely low-bandwidth conditions. This makes it a valuable solution for scenarios that are traditionally inaccessible to conventional communication methods.

Fig. 5 illustrates our generative encoding framework, which is based on a multimodal large model [11]. First, a multimodal image-to-text model extracts semantic information from the input image, which is then used as the input for a text-to-image diffusion model. Next, a mapping encoder identifies the image regions corresponding to the semantic information, enhancing the spatial consistency of the generated content. To further improve the precise alignment between the generated image and the original image, we introduce highly compressed information from the original image as a control signal, providing additional constraints for the text-to-image model. Finally, the decoder integrates the multimodal information to generate a high-quality image.

In extremely bandwidth-constrained environments, no-reference generation emphasizes the preservation and reconstruction of high-level semantics rather than achieving pixel-level fidelity with the original video. This tradeoff is particularly important in scenarios where only minimal control signals—such as textual prompts—can be transmitted. However, ambiguity in natural language input can lead to deviations from the intended content. To improve coherence and controllability, users may be allowed to manually input or refine intent-driven textual descriptions, thereby supplementing the generative model with clearer semantic guidance. While this approach may not guarantee pixel-wise similarity to the original source, it ensures that the core communicative intent is preserved under severe transmission constraints.

3.2. Collaborative transmission

Collaborative transmission explores methods to transmit compressed bitstreams with generative AI models collaboratively, primarily involving elastic transmission, scalable transmission, and generative reception technologies.

3.2.1. Elastic transmission

To increase transmission efficiency and reliability, our collaborative transmission strategy leverages the strengths of both cellular base stations and broadcast towers. This strategy covers several typical service modes, including service addition, service continuation, and service offloading.

(1) **Service addition.** This mode refers to transmitting shared common content (e.g., public-view video streams) through the broadcast system while using cellular base stations for unicast transmission to provide personalized enhancement streams (e.g., user-specific viewpoints or video content generated based on user preferences) to different terminals. Eventually, the broadcast common stream and the cellular personalized stream are combined to satisfy the individualized communication requirements of different users, as illustrated in Fig. 6. This mode fully leverages the complementary characteristics of broadcast and cellular systems, addressing the diverse requirements of immersive media communication.

(2) **Service continuation.** This mode is mainly designed to address service interruptions caused by unreliable broadcast signals. It utilizes cellular base stations to provide supplementary

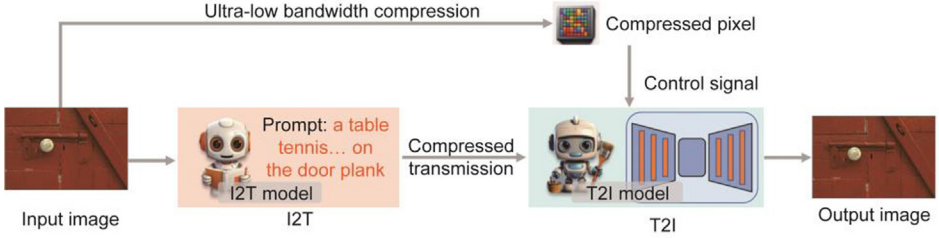

Fig. 5. A multimodal large-model-driven generative encoding scheme. I2T: image-to-text; T2I: text-to-image.

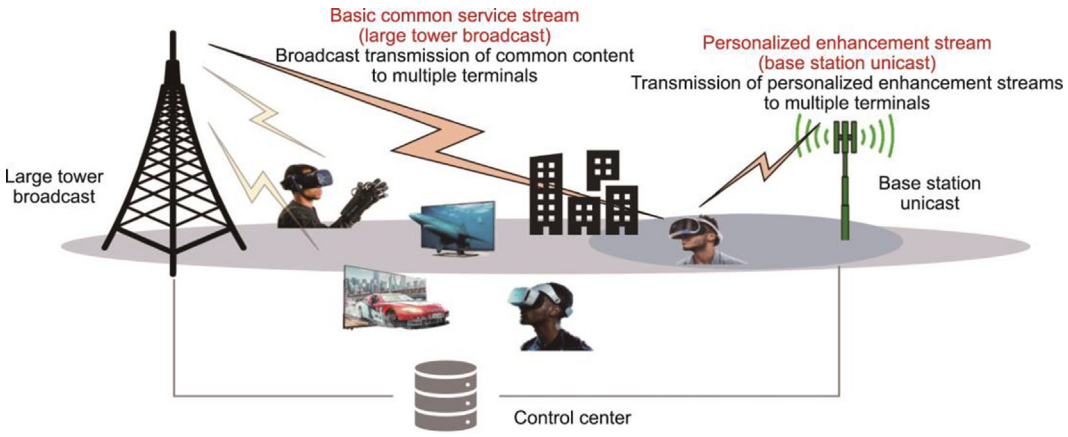

Fig. 6. Service addition mode with broadcast common stream and cellular personalized stream.

coverage, ensuring the continuity and reliability of communication services. Specifically, under normal circumstances, users receive services through broadcast towers. When the core network detects that a large number of users expect to enter areas with poor broadcast signal quality, posing a risk of degraded service quality, the broadcast core network forwards these users' services to the cellular core network. Subsequently, cellular base stations assist the broadcast system to provide high-quality service continuation and packet recovery service, ensuring the maintenance of service continuity and quality for users even in areas with weak signals.

(3) **Service offloading.** As shown in Fig. 7, the service offloading mode aims to alleviate the pressure on the cellular network by offloading part of the content through broadcast towers during peak traffic periods. This strategy is particularly suitable for large-scale live-streaming or high-concurrency scenarios. Under normal circumstances, users receive services through the cellular network. However, when the cellular core network detects a sharp increase in the number of users for common (unicast) services, it may occupy a significant number of time–frequency resources, leading to risks of network congestion and increased latency. To address this issue, the cellular core network offloads part of the traffic to the broadcast core network. Subsequently, the broadcast towers and cellular base stations collaborate to ensure high-quality live-streaming services even during sudden surges in user numbers. This mechanism effectively guarantees users a stable and smooth communication experience during peak network load periods.

3.2.2. Scalable transmission

Generative video communication technology leverages generative AI to achieve scalable transmission by dynamically adjusting the transmitted content based on factors such as network conditions, user requirements, and computational capabilities. This approach can alleviate network congestion and improve service quality. Specifically, in scenarios with limited network bandwidth, generative video communication intelligently optimizes the transmitted content. The sender adopts a contraction strategy, transmitting only the core information of the video, such as key frames and semantics. The receiver generates high-quality complete video content based on the received signals by means of advanced generative AI techniques. This transmission strategy ensures communication quality while effectively reducing bandwidth consumption. Conversely, when the network bandwidth is sufficient, the sender implements an expansion strategy, generating the full video content and transmitting it. This ensures high-fidelity and error-free video transmission, allowing the receiver

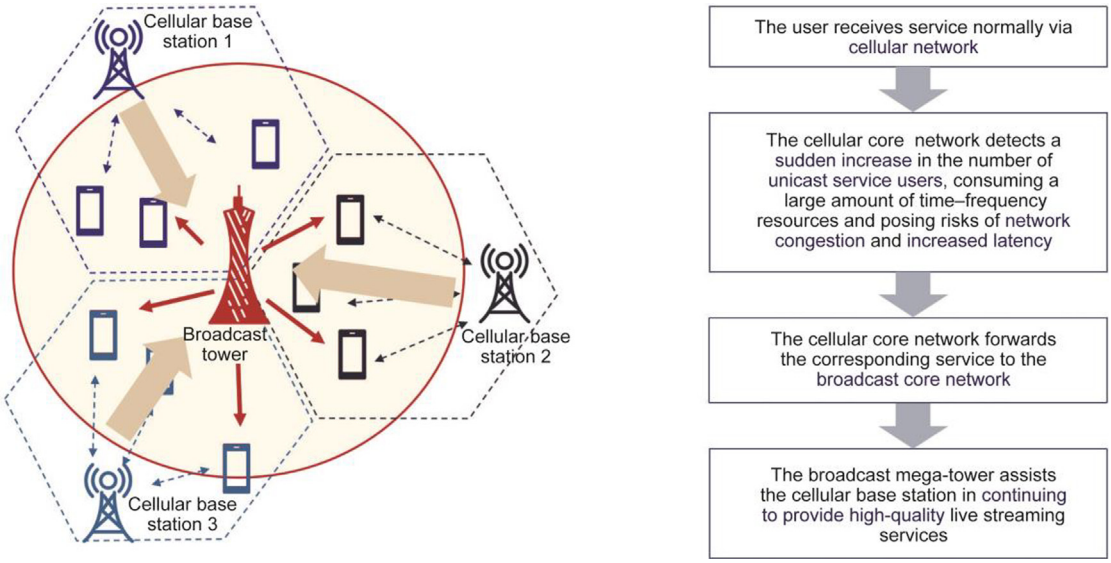

Fig. 7. Service offloading mode for broadcast traffic shunting during cellular traffic surge in demand, overloading resources and risking congestion and latency.

to enjoy the original video experience without relying on generative AI.

Fig. 8 illustrates two classic scenarios of scalable transmission based on generative AI in a multi-node network.

Scenario 1: In Fig. 8, node A needs to send information to nodes B, C, and D. Considering the limited bandwidth, node A uses generative AI to compress the data sent to nodes B and D in order to optimize bandwidth allocation. This allows more bandwidth to be allocated to the link between nodes A and C. As a result, node A can transmit the complete information directly to node C, improving the overall transmission efficiency and link resource utilization. This approach meets the diverse information-transmission needs between different nodes and enhances the system's data-interaction capabilities and stability in complex scenarios.

Scenario 2: Node A aims to transmit information to node H through a multi-hop approach. During this process, the transmission strategy is dynamically adjusted based on network congestion and the computational capabilities of each node to select the appropriate scaling method. More specifically, node A first compresses the information using AI and transmits it to node D. Node D then evaluates the real-time conditions of its link and decides to forward the received information unchanged to node F. Subsequently, node F uses generative AI to reconstruct and amplify the information before transmitting it to the final destination, node H. In this A–D–F–H transmission chain, the information undergoes three key stages: compressed transmission, regular transmission, and amplified generative transmission. This process effectively addresses the complexities of dynamic network environments, ensuring stable and efficient information delivery. It increases the adaptability and reliability of the overall transmission chain, ensuring that information reaches the target node accurately and completely, thereby meeting communication needs in diverse scenarios and strengthening the network's resilience and interaction capabilities under complex conditions.

Through this adaptive, scalable transmission mechanism, generative video communication not only improves the efficiency and quality of video services but also increases the resilience and adaptability of the network. It intelligently responds to varying network conditions, providing users with a consistently high-quality video communication experience.

3.2.3. Generative reception technology

In generative video communication, generative AI technology deeply empowers every reception stage of the receiver. Leveraging advanced generative techniques, it can significantly enhance received signals, thus improving the decoding quality to generate high-quality, personalized video media and creating a next-generation, high-performance generative receiver.

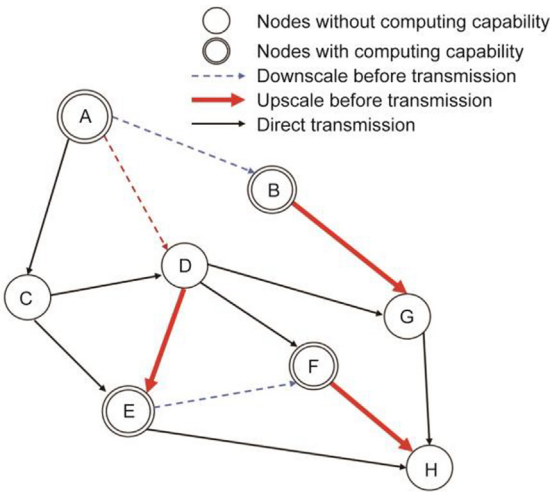

Fig. 8. Scalable transmission based on generative AI. A–H: different nodes.

Fig. 9 presents an application of the diffusion model in the receiver-based channel denoising with diffusion model (CDDM) method [12], which can serve as a new physical layer module following channel equalization to learn the distribution of the channel input signals and subsequently utilize this knowledge to remove channel noise. CDDM achieves an accurate and rapid signal-denoising process by using the received signal as the initial variable for the reverse sampling process and selecting an appropriate initial sampling step based on the channel noise level. Also, CDDM can adapt to the different fading channels by introducing the equalization matrix into the sampling process. When the distribution of the joint source and channel coding (JSCC) output changes, online training can be adopted. The experimental results demonstrate that systems employing CDDM outperform those without CDDM in terms of both signal distortion and content distortion. For example, CDDM improved the peak signal-to-noise ratio (PSNR) performance from 25.99 to 26.52 dB on the DIV2K dataset under a Rayleigh fading channel with a signal-to-noise ratio (SNR) of 10 dB and a compression ratio of 0.0208.

3.3. Utility evaluation

The utility evaluation of generative video communication differs from traditional video communication quality assessments (e.g., PSNR or structural similarity index measure (SSIM)), as it focuses more on the practicality and performance in specific tasks. As mentioned earlier, utility evaluation is primarily measured across the following three dimensions:

(1) **Consistency:** Whether the generated video accurately conveys the required semantic information;

(2) **Reasonableness:** Whether the generated video adheres to physical laws and ethical norms;

(3) **Usability:** Whether the generated video meets personalized needs in specific tasks.

To this end, we propose a utility evaluation framework based on a multimodal large model, as illustrated in Fig. 10. This framework first extracts visual features from reference and generated images using a visual encoder and measures the similarity between these features by calculating the cosine distance. Simultaneously, we design a series of question examples as textual inputs, which are converted into token vectors through tokenization and embedding. To achieve information synchronization across different modalities, the visual features are mapped to the same space as the textual features using a mapper. Subsequently, the aligned visual and textual features are fed into a large language model (LLM) to enable a comprehensive evaluation of the generated video quality. During this process, we employ low-rank adaptation (LoRA) technology to efficiently fine-tune the parameters of the visual encoder and visual mapper while maintaining the parameters of other parts in order to fully leverage the powerful prior knowledge of the pre-trained model. This utility evaluation framework provides a more accurate and comprehensive assessment method for generative video communication, offering quantitative support for task requirements in practical applications.

3.3.1. LLM-based image quality assessment

In Ref. [13], we explored how to teach multimodal large models to perform visual scoring consistent with human preference, as shown in Fig. 10. Inspired by the observation that human evaluators typically use discrete textual definitions for rating levels during subjective scoring, we proposed a method to mimic this subjective scoring process. More specifically, during the training

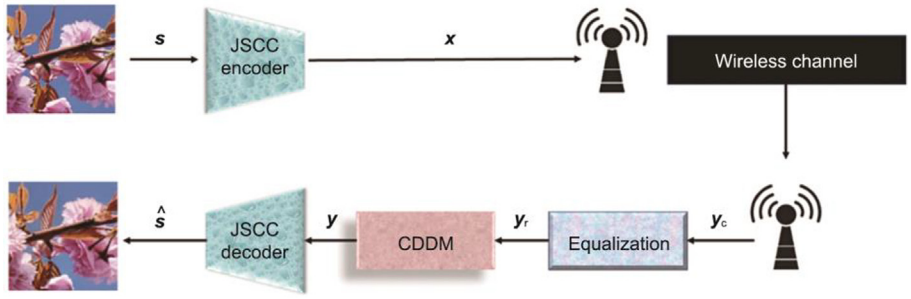

Fig. 9. A channel-denoising diffusion model. $\boldsymbol{s}$: source image; $\hat{\boldsymbol{s}}$: reconstructed image; $\boldsymbol{x}$: latent representation; $\boldsymbol{y}$: decoded latent representation; $\boldsymbol{y}_r$: received signal; $\boldsymbol{y}_c$: real signal; JSCC: joint source and channel coding; CDDM: channel denoising with diffusion model.

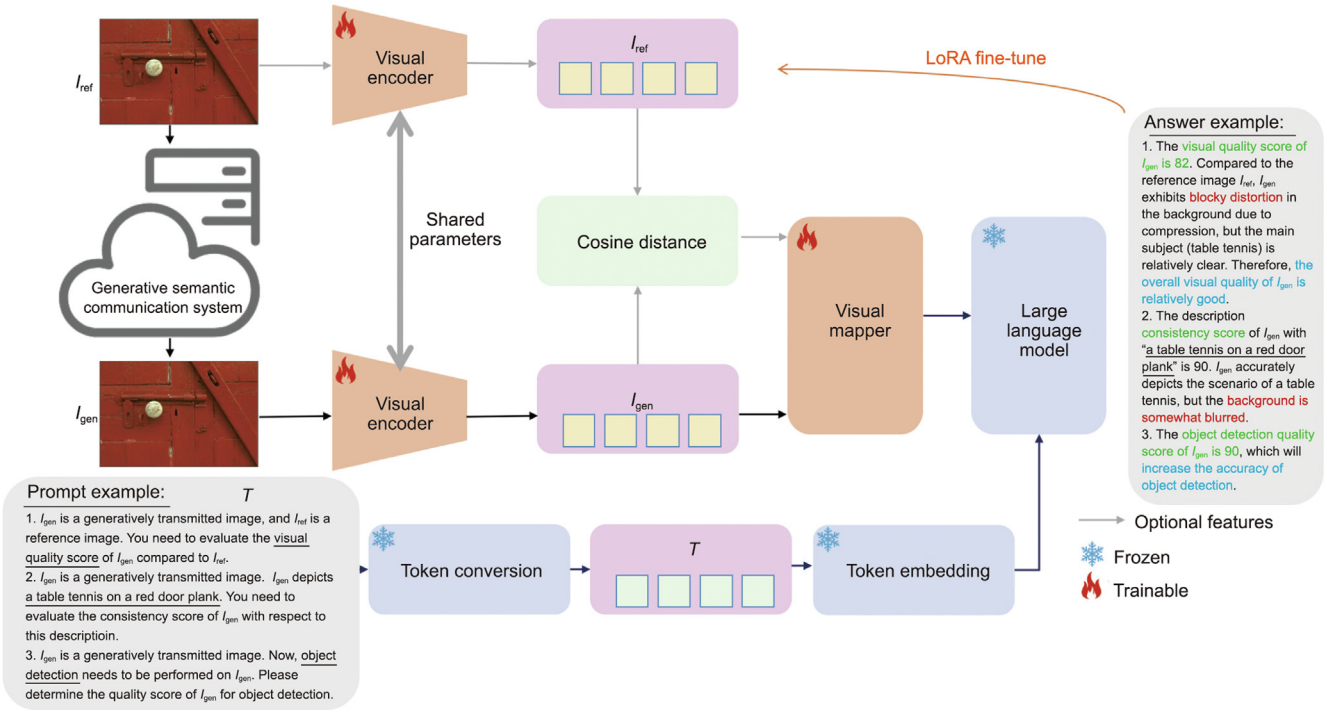

Fig. 10. A utility evaluation framework based on a large language model. T: tokens; LoRA: low-rank adaptation; I_{ref}: reference image; I_{gen}: generatively transmitted image.

phase, the mean opinion scores (MOSs) from the existing datasets are first converted into five discrete textual rating levels (e.g., excellent, good, fair, poor, and very poor). These levels are formatted into instruction–response pairs, and the model is then fine-tuned through visual instruction to learn how to generate corresponding rating levels based on input images or videos. During the inference phase, the model first predicts the probability of each textual rating level and then converts these levels into scores using a defined inverse mapping. Finally, a weighted average method is applied, multiplying the predicted probability of each rating level by its corresponding score and summing them to obtain the final model-predicted score. This approach enables multimodal large models to simulate the scoring process of human evaluators, achieving ratings consistent with human evaluation standards in visual scoring tasks.

3.3.2. Enhancing the low-level visual perception capabilities of models based on multi-model dataset augmentation

In Ref. [14], we introduced the first multimodal dataset focused on low-level vision, including the Q-Pathway dataset, with 58 000 human textual feedback entries covering 18 973 images from diverse sources. Based on this feedback, we used the generative pre-trained transformer (GPT) model to transform it into 200 000 instruction–response pairs, forming the derived Q-Instruct dataset. This dataset can be used for the visual instruction tuning of multimodal models, improving their performance in low-level vision tasks such as visual question answering (VQA) and extended dialogue tasks. The experimental results demonstrated that fine-tuning multiple foundational models with our dataset significantly improves their low-level visual capabilities, particularly in terms of generalization to unseen datasets. More specifically, for LLaVA-v1.5 (7B), fine-tuning with Q-Instruct improved the overall perception accuracy on the LLVisionQA test from 60.07% to 69.30% and the low-level description score from 3.21 to 3.86.

Although the proposed LLM-based multimodal evaluation framework provides a powerful means of assessing semantic consistency and perceptual quality, it raises concerns regarding computational cost, latency, and feasibility for deployment on resource-constrained edge devices such as smartphones. In practical scenarios, a potential solution is to adopt a cloud–edge collaborative architecture, in which lightweight feature extraction can be performed locally, while the more computationally intensive inference steps are offloaded to cloud servers. Alternatively, recent advances in model compression and knowledge distillation allow the use of compact multimodal models to approximate the performance of large-scale language models, making on-device

deployment more practical. Moreover, the evaluation process can be adaptively triggered based on specific task requirements or application contexts to balance latency and accuracy. These considerations highlight a tradeoff between evaluation precision and system responsiveness, and further exploration of lightweight deployment strategies will be an important direction for future work.

3.4. Summary

Generative video communication offers both significant advantages and notable challenges. On the positive side, it enhances video communication's ability of expression by enabling richer, more semantically meaningful video representations. It also improves bandwidth efficiency through high-fidelity content reconstruction from compact signals, allowing low-bitrate transmission. The integration of cognitive priors and multimodal prompts further supports personalization and adaptive task-oriented video generation. Additionally, collaborative transmission strategies dynamically adapt heterogeneous device capabilities and fluctuating network conditions. However, generative video communication presents several challenges. First, model reliability is a key concern, as the uncertainty of generative outputs raises issues of consistency and controllability. Additionally, real-time generation and decoding impose substantial computational costs, especially for resource-constrained receivers. Traditional quality metrics, such as PSNR and SSIM, are insufficient for evaluating perceptual or task-level quality, necessitating new frameworks. Furthermore, the use of generative AI introduces risks such as misinformation, privacy leakage, and ethical concerns, highlighting the need for refined encoding schemes, intelligent transmission strategies, and robust utility evaluation systems.

4. Potential directions

This section outlines several potential future research directions for generative video communication.

4.1. Theoretical framework

Guided by Shannon's first and second theorems, the performance limits of traditional video communication have been clearly defined. However, under the framework of generative video communication, its mathematical representation and theoretical foundation remain incomplete. The performance limitation of generative video transmission is a multidimensional and complex issue, involving factors such as the theoretical limits of AI generative models, network computational capabilities, communication bandwidth, and human perceptual limits. Therefore, constructing a foundational theory of generative video communication from an information-theoretic perspective will be a core direction for future research.

4.2. 6G networks

Generative video communication imposes higher demands on network bandwidth and computational capabilities, necessitating research in 6G networks on multidimensional resource scheduling, low-latency transmission, and privacy protection to promote its application. Specifically, research is needed on dynamically adjusting the allocation of communication, computing, and storage resources to enable adaptive optimization based on video content complexity, generative model computational requirements, and user needs, thereby improving video communication quality and reducing bandwidth waste. In terms of low-latency transmission, optimizing the inference speed of generative models and designing low-latency generative video transmission mechanisms in 6G networks will be critical to meet the demands of real-time video calls, extended reality (XR), and other latency-sensitive applications. Regarding privacy protection, generative video communication involves large amounts of personal data, and research is needed on how to achieve privacy-preserving video generation and transmission based on generative models in 6G networks to prevent data leaks or forgeries.

4.3. Optimized design of generative models

Traditional video communication typically involves multiple steps, such as video capture, encoding, transmission, decoding, and display. In contrast, generative video communication requires end-to-end generative models to accomplish these tasks, avoiding the bottlenecks of traditional encoding and decoding. Therefore, an important research direction is the optimized design of efficient, lightweight, end-to-end generative models tailored for video services under the computational resource constraints of various terminals. These models should be capable of the real-time processing of spatiotemporally dynamic video generation, enhancing the overall efficiency and performance of the system.

5. Summary and outlook

Generative video communication, driven by generative AI technology, redefines the objectives and implementation strategies of video communication. By integrating reference sampling with generative information, this paradigm significantly increases the expressive capabilities of video content under limited bitrates while substantially reducing transmission resource consumption. It not only achieves excellent performance in low-bitrate video transmission but also demonstrates broad application potential in cutting-edge fields, such as immersive media communication and cross-modal content generation, offering solutions to break the limitations of traditional communication systems. However, the field still faces several technical challenges, including improving the efficiency of information expression and fusion, controlling compression and transmission energy consumption, and supporting personalized needs. With the continuous advancement of generative AI technology in the future, generative video communication is expected to be widely applied in emerging scenarios such as personalized video communication, virtual reality, and augmented reality, delivering richer, more efficient, and personalized video communication experiences to users.

CRediT authorship contribution statement

Wenjun Zhang: Writing – original draft, Investigation, Funding acquisition, Conceptualization. **Guo Lu:** Writing – review & editing, Methodology, Formal analysis. **Zhiyong Chen:** Writing – review & editing, Methodology, Investigation. **Geoffrey Ye Li:** Writing – review & editing, Project administration, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62431015, 62471290, and 62222111).

References

[1] Zheng J, Meister M. The unbearable slowness of being: why do we live at 10 bits/s? Neuron 2025;113(2):192–204.
[2] Kingma DP, Welling M. Auto-encoding variational bayes. 2013. arXiv:1312.6114.
[3] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems; 2014 Dec 8–13; Montreal, QC, Canada. New York City: IEEE; 2014. p. 2672–80.
[4] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. Red Hook: Curran Associates Inc.; 2020. p. 6840–51.
[5] Tian Y, Lu G, Yan Y, Zhai G, Chen L, Gao Z. A coding framework and benchmark towards low-bitrate video understanding. IEEE Trans Pattern Anal Mach Intell 2024;46(8):5852–72.
[6] Sullivan GJ, Ohm JR, Han WJ, Wiegand T. Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol 2012;22 (12):1649–68.
[7] Yang R, Timofte R, Van Gool L. Perceptual learned video compression with recurrent conditional GAN. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22); 2022 Jul 23–29; Vienna, Austria. Sacramento: International Joint Conferences on Artificial Intelligence; 2022. p. 1537–44.
[8] Li M, Shi Y, Wang J, Huang Y. High visual-fidelity learned video compression. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. New York City: Association for Computing Machinery; 2023. p. 8057–66.
[9] Zhu C, Lu G, He B, Xie R, Song L. Implicit–explicit integrated representations for multi-view video compression. IEEE Trans Image Process 2025;34:1106–18.
[10] Chen H, Gwilliam M, Lim SN, Shrivastava A. HNeRV: a hybrid neural representation for videos. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. New York City: IEEE; 2023. p. 10270–9.
[11] Li C, Lu G, Feng D, Wu H, Zhang Z, Liu X, et al. MISC: ultra-low bitrate image semantic compression driven by large multimodal model. IEEE Trans Image Process 2025;34:335–49.
[12] Wu T, Chen Z, He D, Qian L, Xu Y, Tao M, et al. CDDM: channel denoising diffusion models for wireless semantic communications. IEEE Trans Wirel Commun 2024;23(9):11168–83.
[13] Wu H, Zhang Z, Zhang W, Chen C, Liao L, Li C, et al. Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria. New York City: JMLR; 2024. p. 54015–29.
[14] Wu H, Zhang Z, Zhang E, Chen C, Liao L, Wang A, et al. Q-Instruct: improving low-level visual abilities for multi-modality foundation models. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 25490–500.