Bridge: A Cross-Modal Learning Framework for Unified Semantic Representation in Noisy Communication

Liang Chen[1], **Yanze Huang**[2], **Limei Lin**[1*], **Xiaoding Wang**[1], **Wei Lou**[3], **Jie Wu**[4] and **Sun-Yuan Hsieh**[5]

[1]College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China
[2]School of Computing and Data Science, Fujian University of Technology, Fuzhou 350118, China
[3]The Hong Kong Polytechnic University, Hong Kong, China
[4]China Telecom Cloud Computing Research Institute, Beijing 100088, China
[5]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan
liangchen011208@gmail.com, yzhuang@fjut.edu.cn, linlimei@fjnu.edu.cn, wangdin1982@fjnu.edu.cn, csweilou@polyu.edu.hk, jiewu@temple.edu, hsiehsy@mail.ncku.edu.tw

Abstract

Multimodal semantic communication systems face a critical challenge in extracting and aligning semantic features across heterogeneous modalities within a unified representation space, particularly under noisy transmission conditions. To address this, we propose Bridge, a cross-modal learning framework that integrates video, audio, and text into a unified semantic space through feature disentanglement and contrastive alignment. Bridge separates modality-specific and modality-invariant representations, enhancing both intra-modal precision and inter-modal generalization. During decoding, it leverages heterogeneous foundation models to preserve semantic fidelity under channel noise. Extensive experiments on multiple multimodal benchmarks under varying SNR levels demonstrate that Bridge maintains high semantic consistency and reconstruction quality, improving image reconstruction by approximately 16.3% in low-SNR regimes. Our work provides a practical and theoretically grounded framework for next-generation multimodal semantic communication systems.

1 Introduction

Traditional communication systems predominantly rely on the accurate transmission of raw bit streams. However, with the rapid advancement of artificial intelligence (AI) and intelligent devices, increasing attention has been directed toward the semantic value of information [Yang *et al.*, 2023; Xia *et al.*, 2024]. Unlike raw data, semantic information aligns more closely with human cognition and task-specific requirements, enabling more efficient and purposeful communication under stringent bandwidth constraints. Semantic communication has thus emerged as a transformative paradigm, prioritizing the conveyance of intended meaning over the faithful delivery of original data [Wang *et al.*, 2023b]. This shift not only aligns with the intelligence-driven demands anticipated in future 6G networks but also paves the way toward communication systems endowed with cognitive and context-aware capabilities.

Current research efforts in semantic communication have primarily focused on unimodal data (e.g., text, audio, and video), emphasizing semantic extraction, compression, and transmission. For textual modalities, natural language processing techniques, including word embeddings and Transformer architectures, have enabled effective semantic representation and delivery [Miao *et al.*, 2024]. In the audio domain, the integration of automatic speech recognition and semantic inference has improved the preservation of speaker intent [Chen *et al.*, 2024]. For video-based communication, convolutional neural networks and temporal modeling approaches have been employed to capture and transmit core semantic content while minimizing redundancy [Hwang *et al.*, 2024]. Despite the progress, unimodal semantic systems still face several critical challenges, including ambiguous semantic definitions, limited channel robustness [Cheng *et al.*, 2024], and high system complexity, all of which hinder scalability and adaptability in dynamic real-world scenarios.

To overcome these limitations and enhance the depth and flexibility of semantic representation, research focus has increasingly shifted toward multimodal semantic communication [Lu *et al.*, 2024]. By integrating heterogeneous data sources, multimodal systems offer a more comprehensive understanding of complex semantics [Liu *et al.*, 2024]. Theoretical advances in this domain have introduced concepts such as multimodal semantic entropy and cross-modal mutual information, while early frameworks for joint semantic encoding have been proposed. Technologically, advances in multimodal Transformers, graph neural networks, and attention mechanisms have significantly enhanced the modeling of semantic relationships across diverse modalities [Cheng *et al.*, 2025]. These developments show promising potential for applications in autonomous driving, intelligent healthcare, and immersive virtual environments [Tang *et al.*, 2024]. Nev-

*Corresponding author.

ertheless, current multimodal semantic systems often face challenges such as inefficient cross-modal alignment, limited real-time processing capabilities [Nguyen *et al.*, 2024], and the absence of unified evaluation standards, all of which present obstacles to large-scale deployment.

To address these challenges, previous studies have shown that emergent protocols optimized solely for utility often fail to produce semantically consistent mappings [Ben Zion *et al.*, 2024], highlighting the need for a unified and reconstruction-oriented approach to semantic learning. In response, this paper introduces a novel multimodal semantic framework that projects heterogeneous modalities into a unified latent feature space and employs alignment mechanisms to ensure coherent semantic representation. This design significantly improves both the robustness of communication and semantic fidelity in complex and dynamic environments. Compared to existing methods, the proposed framework achieves notable improvements in semantic alignment accuracy, system adaptability, and resilience to channel disturbances [Zhao *et al.*, 2024].

The contributions of this paper are summarized as follows.

(1) We propose a cross-modal learning approach that effectively bridges the gap between video, audio, and text by learning a unified semantic representation. Our method disentangles modality-specific and modality-invariant features, enabling precise processing within each modality while ensuring robust generalization across modalities.

(2) We introduce a novel cross-modal alignment mechanism that enhances semantic fidelity by aligning features from different modalities in a shared latent space. This method is designed to maintain consistent semantic mapping across modalities, even in the presence of noise and fading, addressing the challenges of communication channels.

(3) Extensive experiments on various multimodal benchmarks demonstrate that BRIDGE outperforms existing methods, particularly in low-SNR conditions. Our results show that BRIDGE preserves semantic relationships across modalities while maintaining robust performance, improving image reconstruction (PSNR) by approximately 16.3% in low-SNR regimes, validating its effectiveness in applications.

2 Related Work

Unimodal Semantic Communication. Existing research on semantic communication has primarily focused on unimodal data (e.g., text, speech/audio, and vision), aiming to transmit task-relevant semantics rather than raw bitstreams to improve communication efficiency under bandwidth constraints and channel impairments [Shao *et al.*, 2022]. With the progress of deep learning, end-to-end semantic transceivers for text have become increasingly mature [Xie *et al.*, 2021]. For speech/audio scenarios, task-driven semantic transmission has been explored by integrating recognition and synthesis objectives, enabling robust intent-preserving delivery over noisy channels [Weng *et al.*, 2023]. For temporal signals such as video, semantic modeling often leverages spatiotemporal representation learning to capture both spatial cues and temporal dependencies for more faithful semantic extraction [Akbari *et al.*, 2021]. In the visual domain, recent advances further incorporate stronger generative priors (e.g., diffusion-based decoding) to enhance perceptual quality and robustness, especially in low-SNR regimes [Peng *et al.*, 2025]. Despite these advances, unimodal semantic systems still face persistent challenges, including the ambiguity between "semantics" and task objectives, limited robustness under channel mismatch, and non-trivial system complexity for practical deployment [Xie *et al.*, 2023].

Multimodal Semantic Communication. To overcome the limitations of unimodal systems and improve semantic completeness, research attention has gradually shifted to multimodal semantic communication, where heterogeneous sources are jointly exploited for more comprehensive understanding and more reliable downstream performance [Zhang *et al.*, 2024; Xie *et al.*, 2022]. Unlike unimodal settings, multimodal semantic communication must not only extract modality-specific features but also model inter-modal relationships and complementary cues to achieve context-aware and synergy-enhanced semantics . Methodologically, cross-modal alignment and fusion are often supported by contrastive and unified representation learning paradigms, exemplified by vision–language pretraining and mixture-of-experts style modeling that map heterogeneous inputs into shared semantic spaces [Radford *et al.*, 2021; Bao *et al.*, 2022]. However, practical multimodal semantic communication remains challenging due to the pronounced *modality gap*: heterogeneous modalities differ in distribution, granularity, and noise sensitivity, which can degrade alignment quality, and channel noise/fading further amplifies inconsistency at the receiver and causes unstable task performance [Xie *et al.*, 2022].

Unified Semantic Representation. A central goal of cross-modal learning is to build a unified semantic representation by projecting heterogeneous modalities into a shared latent space and enforcing alignment to preserve semantic consistency [Li *et al.*, 2021]. Recently, scalable pretraining frameworks have strengthened this direction by combining shared semantic spaces with modality-aware architectures (e.g., modality-specific branches or expert components), leading to more expressive and transferable cross-modal representations [Wang *et al.*, 2023a; Li *et al.*, 2023]. Nevertheless, most alignment strategies are developed under relatively stable data assumptions. When transferred to wireless semantic communication, channel perturbations can reshape the transmitted feature distribution and introduce uncertainty, which may induce semantic drift and cross-modal mismatch under noise and fading [Xie *et al.*, 2023; Zhang *et al.*, 2026]. Moreover, under joint constraints of bandwidth limitation, channel variation, and deployment requirements (complexity/latency), achieving accurate alignment, robust generalization, and practical efficiency simultaneously remains an open challenge [Shao *et al.*, 2022]. This is particularly difficult in real-world applications, where environmental factors constantly change and affect the performance of cross-modal models.

Motivation. Motivated by these observations, key questions arises for next-generation multimodal communication: how can we learn a unified semantic representation that is (i) consistent across heterogeneous modalities, (ii) robust to noise, fading, and channel mismatch, and (iii) efficient

enough for practical deployment under bandwidth and latency constraints. Addressing questions is crucial for enabling reliable multimodal semantic delivery and downstream performance in dynamic real-world environments.

3 Methodology

In this section, we provide a detailed description of the proposed multimodal semantic communication framework, Bridge. The core objective of Bridge is to enable collaborative expression and robust transmission of multimodal data within a unified semantic space through feature disentanglement and cross-modal alignment mechanisms. To facilitate a clearer understanding of the overall workflow and modular structure of the framework, Figure 1 illustrates the architecture of Bridge. Specifically, inputs from text, audio, and video modalities are first processed by modality-specific feature extractors, which generate both modality-specific and modality-invariant representations. The modality-invariant features are subsequently quantized and transmitted to the receiver side. A two-stage decoder is employed for semantic reconstruction: in the first stage, lightweight modality-specific decoders are used to reconstruct preliminary modality signals; in the second stage, large-scale foundation models are leveraged to verify and enhance semantic consistency.

Feature Extraction Module. In multimodal semantic communication, different modalities simultaneously carry modality-specific characteristics and invariant semantic content. To effectively extract and separate these aspects, we design a dual-branch encoder architecture. Each input modality is processed through two parallel pathways: one branch is responsible for extracting modality-specific features, while the other captures modality-invariant representations. These two branches are trained jointly but kept structurally independent to ensure specialization and avoid semantic interference.

This branch is designed to preserve the unique and irreplaceable characteristics intrinsic to each modality. For a given input x^m from modality $m \in \{a, b, c\}$, we first extract initial features using the modality-specific encoder $\phi^m(\cdot)$. These features are then projected and compressed using a projection head $\mathrm{MLP}^m_{\text{spec}}$. To improve robustness, Gaussian noise is added as a regularization term. The resulting modality-specific feature is defined as

$$h^m = \mathrm{MLP}^m_{\text{spec}}(\phi^m(x^m)) + \epsilon^m, \quad \epsilon^m \sim \mathcal{N}(0, \sigma^2 I). \quad (1)$$

This representation is not quantized and is used in downstream tasks where fidelity and integrity are important.

To facilitate semantic alignment across modalities, we introduce a parallel branch for extracting modality-invariant features. This branch is built upon a shared semantic encoder that projects all modalities into a common latent space.

Specifically, for each modality $m \in \{a, b, c\}$, the input x^m is first encoded by the same low-level modality-specific encoder $\phi^m(\cdot)$. The output is then passed to a shared encoder $g_{\text{shared}}(\cdot)$ to produce the semantic representation as follows.

$$z^m = g_{\text{sharcd}}(\phi^m(x^m)). \quad (2)$$

These modality-invariant features are subsequently used in the specific-invariant dependency minimization and the modality-aware cross-modal semantics for optimizing feature separation and cross-modal semantic consistency.

Specific-Invariant Dependency Minimization (SDMI). To ensure effective separation between modality-specific and modality-invariant features, we enhance the SDMI module by introducing auxiliary projection heads (distinct from $\mathrm{MLP}^m_{\text{spec}}$ used in feature extraction), which are solely used for robust mutual information estimation.

To facilitate mutual information estimation while preserving numerical stability, we introduce two auxiliary projection heads ψ_m and ψ_{inv}, which map z^m and z^{inv} into low-dimensional semantic spaces only for mutual information estimation. This design is based on the invariance property of mutual information, i.e. $I(z^m; z^{\text{inv}}) = I(\psi_m(z^m); \psi_{\text{inv}}(z^{\text{inv}}))$ under deterministic mappings.

We enhance the contrastive estimation framework by expanding the Mutual Information Neural Estimator (MINE), which approximates mutual information via the Donsker–Varadhan representation as follows.

$$\begin{aligned}\mathcal{L}_{\text{SDMI}} = \mathbb{E}_{p(z^m, z^{\text{inv}})}[T_\theta(z^m, z^{\text{inv}})] \\ - \log \mathbb{E}_{p(z^m)p(z^{\text{inv}})}[\exp(T_\theta(z^m, z^{\text{inv}}))],\end{aligned} \quad (3)$$

where T_θ is a neural critic operating on the projected features. It provides a tighter and smoother gradient signal and eliminates the need for explicit negative sampling.

These enhancements allow SDMI to more effectively suppress residual redundancy between the shared and specific latent factors, leading to improved disentanglement and more robust cross-modal generalization. across modalities.

Modality-Aware Cross-modal Semantics (MACS). To align modality-invariant features in a shared semantic space, we introduce MACS, a pairwise-gated contrastive module. MACS forms anchor–candidate pairs across modalities, maps them with two asymmetric projection heads ϕ_a and ϕ_b, and measures compatibility through bilinear similarity. A learnable *pairwise gate* then scales each pair's logit inside a negative InfoNCE objective to emphasize reliable matches and suppress noisy or weak alignments as follows.

$$\mathcal{L}_{\text{MACS}} = -\frac{1}{N}\sum\nolimits_{t=1}^{N} \log \frac{\exp\big(g_{t,t}\, s_{t,t}\big)}{\sum_{j=1}^{N} \exp\big(g_{t,j}\, s_{t,j}\big)}, \quad (4)$$

where the pairwise similarity is

$$s_{t,j} \;=\; \big(\phi_a(z_t^a)\big)^\top W\, \phi_b(z_j^b), \quad W \in \mathbb{R}^{d\times d}. \quad (5)$$

The gate for each pair (t, j) is produced from the concatenated projections and squashed by a sigmoid as follows.

$$g_{t,j} \;=\; \sigma(\mathrm{MLP}_{\text{gate}}([\phi_a(z_t^a) \,\|\, \phi_b(z_j^b)])) \in (0, 1]\,. \quad (6)$$

This design is equivalent to using an adaptive temperature with $g_{t,j} = 1/\tau_{t,j}$: low cross-modal consistency increases $\tau_{t,j}$, whereas high consistency decreases $\tau_{t,j}$, thereby enforcing explicit cross-modal structural consistency.

Vector Quantization. After the alignment phase, the modality-invariant features are discretized via vector quantization (VQ), converting continuous latent semantics into discrete code indices. This step improves compactness and maps

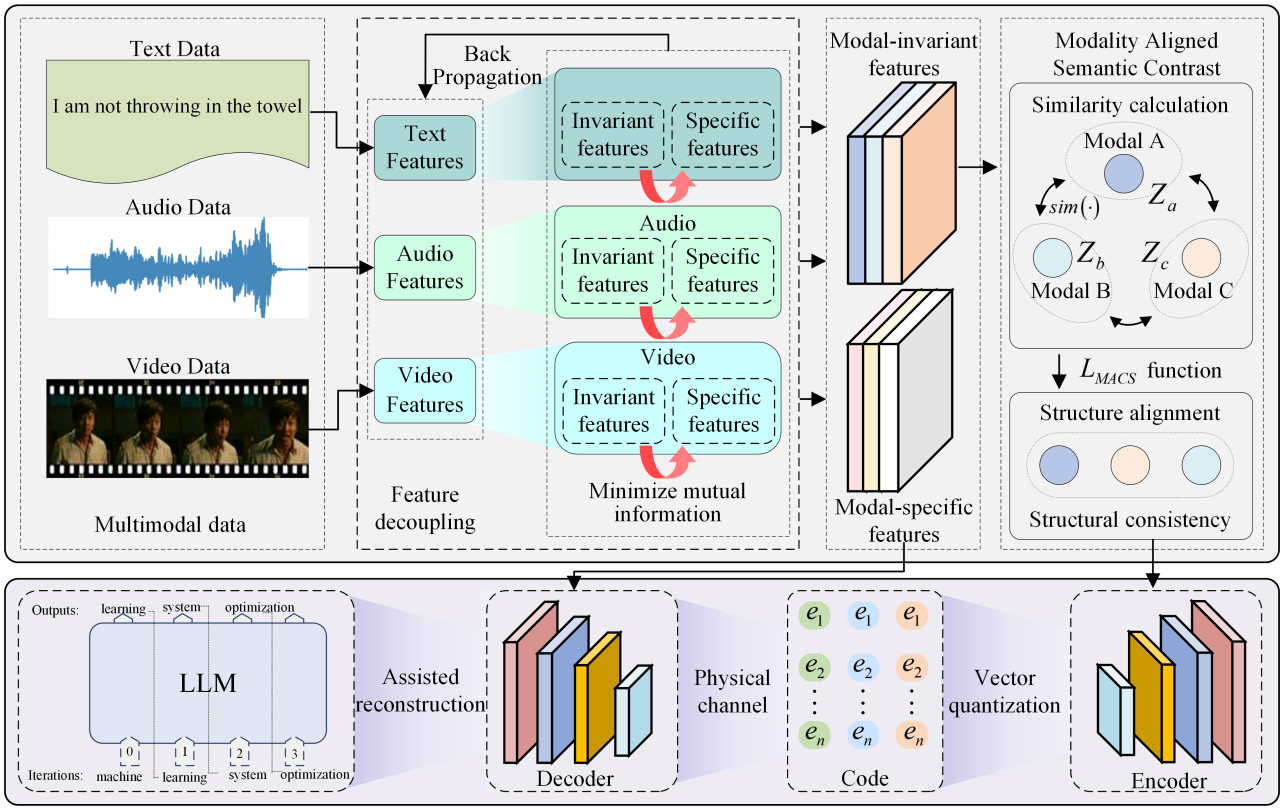

Figure 1: Overview of the multimodal semantic communication system: modality-specific and modality-invariant features are extracted from text, audio, and video, quantized by the encoder and transmitted through the channel, and the decoder assisted by a large language model recovers and reconstructs the multimodal representations.

cross-modal semantics to a unified discrete space for downstream reconstruction and reasoning.

Let $z^m \in \mathbb{R}^{T\times D}$ be the invariant representation of modality m. We quantize each time step by nearest-neighbor lookup in a *shared* codebook $\{e_j\}_{j=1}^{K} \subset \mathbb{R}^D$ as follows.

$$\hat{z}_t^m = \mathrm{Quant}(z_t^m) = e_k,\ k = \arg\min_j \|z_t^m - e_j\|_2. \quad (7)$$

Training Mechanism. We train the discrete bottleneck with the straight-through estimator (STE). In the backward pass the quantizer is treated as identity by using a stop-gradient operator $\mathrm{sg}[\cdot]$ as follows.

$$\tilde{z}_t^m \ = \ z_t^m + \mathrm{sg}\big[e_k - z_t^m\big]. \quad (8)$$

The VQ loss combines a *codebook* term and a *commitment* term as follows.

$$\mathcal{L}_{\mathrm{VQ}} = \sum_m \sum_t \Big\|\mathrm{sg}[z_t^m] - e_k\Big\|_2^2 + \beta \sum_m \sum_t \Big\|z_t^m - \mathrm{sg}[e_k]\Big\|_2^2, \quad (9)$$

where $\beta > 0$ (we use $\beta = 0.25$ unless stated otherwise). The first term updates code vectors toward assigned latents, and the second encourages latents to stay close to selected codes.

Cross-modal Usage Monitoring and Reset. To mitigate *codebook collapse* (few codes overused, many inactive), we adopt a windowed cross-modal usage rule. Let $\mathcal{M}$ be the set of modalities and $u_i^{(m)}(t') \in \{0,1\}$ indicate whether code e_i is used by modality m at step t'. Within the most recent N_{reset} steps, we count the number of *distinct* modalities that have used e_i as follows.

$$C_i \ = \ \Big|\big\{\, m \in \mathcal{M} \ \big|\ \sum_{t'=t-N_{\text{reset}}+1}^{t} u_i^{(m)}(t') > 0 \big\}\Big|. \quad (10)$$

If $C_i < 2$ (i.e., the code is not shared by at least two modalities within the window), we re-initialize it as follows.

$$C_i < 2 \ \Rightarrow \ e_i \sim \mathcal{N}(0, I). \quad (11)$$

This strategy periodically resets inactive or modality-exclusive codes, improving codebook utilization and promoting robust, shared semantic encoding across modalities.

Semantic Decoding with Foundation Models. On the receiver side, a multimodal semantic decoding and reconstruction module is designed to recover the underlying semantic representations of audio, video, and text from the compressed and quantized vectors transmitted over the channel. Since these vectors are modality-invariant and aligned in a shared latent space, decoding must focus not only on surface-level signal restoration, but also on preserving the high-level semantic structures originally embedded.

To achieve this, the system employs a two-stage decoding process. First, modality-specific decoders are used to reconstruct intermediate modality data—such as audio waveforms,

video frames, and text tokens—from quantized latent representations. These decoders are lightweight and trained jointly with the encoder to maintain low latency and acceptable fidelity under constrained transmission conditions.

We define the reconstruction objective as a weighted sum of modality-specific reconstruction losses as follows.

$$\mathcal{L}_{\text{recon}} = \lambda_a \cdot \|\hat{x}^a - x^a\|^2 + \lambda_v \cdot \|\hat{x}^v - x^v\|^2 + \lambda_t \cdot \text{CE}(\hat{x}^t, x^t), \tag{12}$$

where $\hat{x}^a, \hat{x}^v, \hat{x}^t$ are reconstructed audio/video/text signals, x^a, x^v, x^t denote their original counterparts. We use mean squared error (MSE) for continuous modalities (audio and video) and cross-entropy (CE) for discrete text prediction.

In the second stage, a foundation model alignment mechanism is introduced to perform semantic verification and enhancement. Each reconstructed modality-specific signal is passed through a pre-trained foundation model that maps it into a high-dimensional semantic embedding space.

$$f^{(a)}_{\text{sem}}(\hat{x}^a) \in \mathbb{R}^d, \quad f^{(v)}_{\text{sem}}(\hat{x}^v) \in \mathbb{R}^d, \quad f^{(t)}_{\text{sem}}(\hat{x}^t) \in \mathbb{R}^d, \tag{13}$$

where $f^{(a)}_{\text{sem}}$, $f^{(v)}_{\text{sem}}$, and $f^{(t)}_{\text{sem}}$ denote CLAP [Elizalde *et al.*, 2023], CLIP [Radford *et al.*, 2021], and BERT [Devlin *et al.*, 2019], respectively, which project the (reconstructed) signals into a shared d-dimensional semantic embedding space. This mapping enables evaluation of semantic consistency and quality using learned representations.

Specifically, for the audio modality, CLAP is employed to project the waveform into an audio-language embedding space, ensuring semantic alignment with the intended meaning. For the video modality, CLIP analyzes reconstructed frames using a dual-encoder architecture that bridges visual and textual semantics. For text, BERT encodes token sequences into contextualized embeddings, enabling assessment of both lexical accuracy and semantic coherence.

This foundation model-enhanced decoding mechanism enables semantic validation, cross-modal embedding consistency, and robustness under transmission degradation.

Overall Objective. To enable effective end-to-end optimization of BRIDGE, we aggregate the losses from each component into a single objective as follows.

$$\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{SDMI}} + \alpha_2 \mathcal{L}_{\text{MACS}} + \alpha_3 \mathcal{L}_{\text{recon}} + \alpha_4 \mathcal{L}_{\text{VQ}}. \tag{14}$$

Here, $\mathcal{L}_{\text{SDMI}}$ minimizes the mutual information between modality-*specific* and modality-*invariant* branches via a DV/MINE critic and a min–max schedule; $\mathcal{L}_{\text{MACS}}$ is a *negative* InfoNCE objective with *pairwise gating* $g_{t,j}$ (interpretable as an adaptive temperature) to enforce cross-modal semantic alignment; $\mathcal{L}_{\text{recon}}$ is the sum of modality-wise reconstruction losses (audio, video, and text); and $\mathcal{L}_{\text{VQ}}$ is the vector-quantization loss combining codebook and commitment terms under a straight-through estimator. The coefficients α_{1-4} are chosen on the validation set to balance gradient magnitudes and overall performance.

4 Experiments

In this section, we describe the experimental setup, including the datasets and tasks used. To comprehensively evaluate the robustness, semantic expressiveness, and downstream adaptability of the proposed semantic communication framework, we conduct experiments using the VGGSound [Chen *et al.*, 2020], COCO2017 [Lin *et al.*, 2014], CMU-MOSEI [Zadeh *et al.*, 2018], and MM-IMDb [Ovalle *et al.*, 2017] datasets.

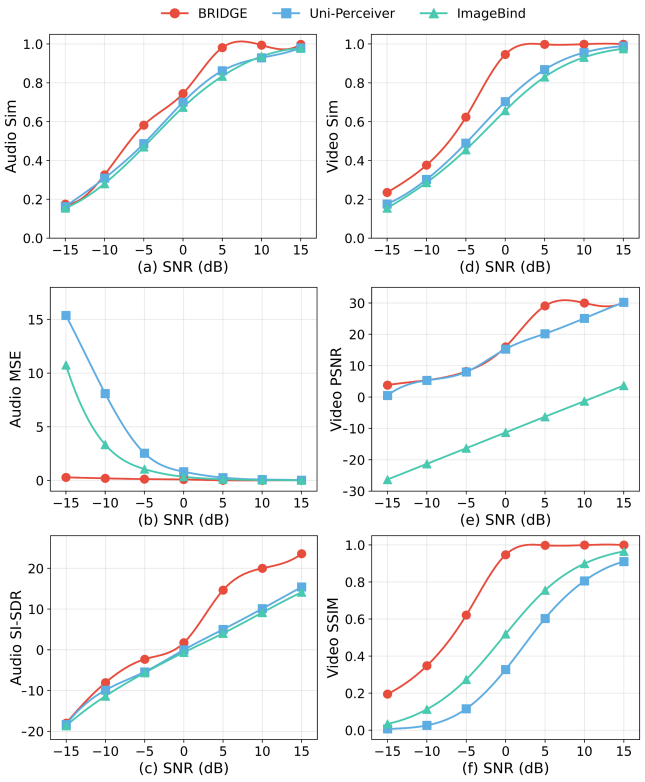

Figure 2: (a)-(c): Audio reconstruction quality under different SNRs. (d)-(f): Video reconstruction quality under different SNRs.

Audio and Video Reconstruction on VGGSound. Figures 2 summarizes reconstruction quality under AWGN across SNRs from -15 to 15 dB, comparing BRIDGE with ImageBind [Girdhar *et al.*, 2023] and the general-purpose Uni-Perceiver v2 [Zhu *et al.*, 2022]. Overall, BRIDGE exhibits consistently stronger robustness across the entire SNR spectrum: it achieves higher semantic similarity for both audio and video, with the most visible gains in low-to-medium SNR regimes, and quickly approaches near-perfect similarity once the channel becomes moderate. Meanwhile, BRIDGE attains markedly better signal/perceptual quality, showing much lower audio reconstruction error and higher audio SI-SDR as SNR increases, as well as substantially higher video PSNR and SSIM, indicating improved pixel-level fidelity and structural preservation of frames. These suggest that disentangled invariant representation together with modality-aware cross-modal alignment enables BRIDGE to simultaneously preserve high-level semantics and low-level reconstruction details under noise, outperforming both unified embedding baselines and a generic multimodal backbone.

Image Reconstruction Quantitative Comparison. On COCO2017, Table 1(a) reports PSNR and LPIPS under

Table 1: Overall performance comparison under different channel conditions and SNRs.

(a) Quantitative comparison of image reconstruction quality under AWGN and Rayleigh channels at different SNRs.

SNR	Channel	Metric	VAEJSCC	ADJSCC	DiffJSCC	DeepJSCC	JSCCformer	SGD-JSCC	Bridge
0	AWGN	PSNR↑	15.59	22.57	19.85	22.44	22.11	21.40	**26.97**
		LPIPS↓	0.300	0.241	0.260	0.2290	0.201	0.122	**0.057**
	Rayleigh	PSNR↑	10.52	18.47	18.52	18.25	19.31	16.31	**24.01**
		LPIPS↓	0.614	0.413	0.419	0.365	0.295	0.235	**0.118**
5	AWGN	PSNR↑	21.86	24.66	21.32	23.95	24.15	23.65	**30.16**
		LPIPS↓	0.099	0.173	0.190	0.167	0.150	0.080	**0.026**
	Rayleigh	PSNR↑	10.57	20.11	20.55	19.97	21.17	19.49	**27.08**
		LPIPS↓	0.556	0.343	0.349	0.305	0.235	0.144	**0.059**
10	AWGN	PSNR↑	25.20	26.66	22.15	25.11	25.62	25.41	**32.62**
		LPIPS↓	0.054	0.128	0.150	0.102	0.122	0.063	**0.012**
	Rayleigh	PSNR↑	10.61	21.17	21.25	21.96	21.78	22.67	**29.80**
		LPIPS↓	0.514	0.302	0.308	0.263	0.203	0.095	**0.029**
15	AWGN	PSNR↑	27.12	27.87	22.11	26.66	26.41	26.10	**33.77**
		LPIPS↓	0.037	0.113	0.142	0.105	0.122	0.063	**0.008**
	Rayleigh	PSNR↑	10.65	22.05	22.05	22.67	23.46	24.52	**31.82**
		LPIPS↓	0.495	0.273	0.281	0.235	0.184	0.075	**0.015**

(b) Classification accuracy (%) under different SNRs on CMU-MOSEI dataset.

Model	−6	−4	−2	0	2	4	6	8	10	12
U-DeepSC(12)	64.78	67.53	70.22	72.93	75.16	76.82	77.58	77.12	78.46	78.52
U-DeepSC(-2)	70.99	74.04	75.47	76.22	76.90	77.44	77.56	77.76	77.87	77.91
T-DeepSC(12)	65.49	68.71	71.58	74.33	76.62	78.98	79.04	79.74	79.96	80.04
T-DeepSC(-2)	71.66	74.82	76.90	78.19	78.84	79.24	79.34	79.43	79.56	79.56
Bridge	**72.53**	**75.78**	**78.51**	**79.33**	**79.35**	**80.11**	**80.78**	**81.21**	**81.72**	**82.15**

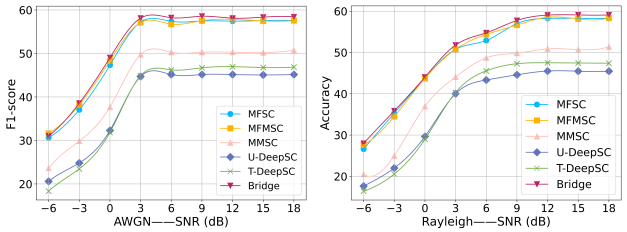

Figure 3: F1-score and accuracy performance of classification tasks.

AWGN and Rayleigh channels at SNR $\in \{0, 5, 10, 15\}$ dB. Compared with representative JSCC baselines, including the semantics-guided diffusion approach SGD-JSCC [Zhang *et al.*, 2026], BRIDGE consistently achieves the best reconstruction quality across all settings, yielding higher PSNR and lower LPIPS. The gain is most pronounced in low-SNR regimes, where BRIDGE better preserves structural details and perceptual fidelity under severe noise and fading. As SNR increases, BRIDGE continues to improve and maintains stable advantages, demonstrating robust performance under both noise and fading.

Multimodal Sentiment Classification Comparison. To further evaluate BRIDGE on multimodal task under different SNRs. As reported in Table 1 (b), BRIDGE achieves the best classification accuracy across the entire SNR range, showing clear advantages especially in low-SNR regimes and maintaining stable gains as channel conditions improve, which demonstrates stronger robustness for sentiment-related semantic preservation under noise.

Multimodal Genre Classification Comparison. Figure 3 compares BRIDGE with MFSC, MFMSC, MMSC [Zhu *et al.*, 2025], U-DeepSC, and T-DeepSC under different SNRs in AWGN and Rayleigh channels. Overall, BRIDGE consistently achieves the best performance across the SNR range in both settings, showing clear advantages especially as channel conditions improve. These results indicate that BRIDGE provides stronger robustness and more reliable semantic preservation under both additive noise and fading.

Image Reconstruction Trend Analysis. Figure 4 presents the PSNR and SSIM trends on the COCO2017 dataset under both AWGN and Rayleigh channels. Compared with the UIDSC approach [Ye *et al.*, 2025], the heatmap results further highlight the across-the-board robustness of Bridge: it maintains stronger overall performance over the entire SNR spectrum and exhibits better generalization under varying and mismatched channel conditions.

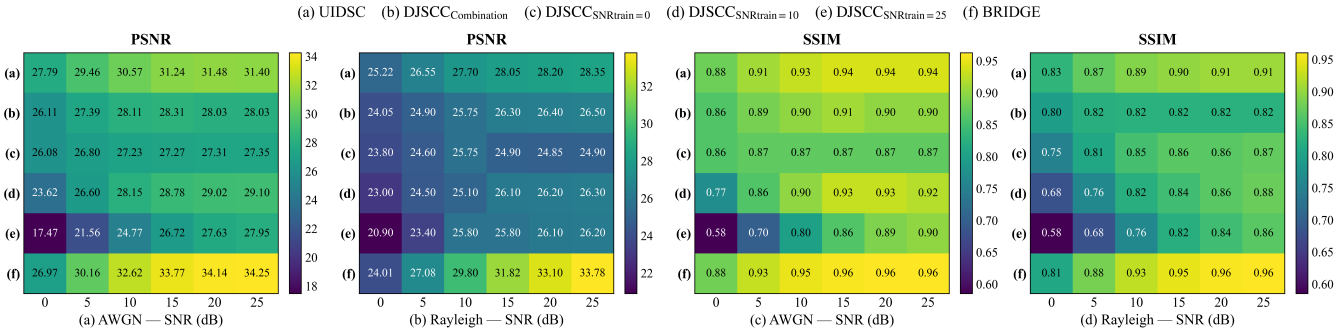

Figure 4: Reconstruction quality under different SNRs in AWGN and Rayleigh channels.

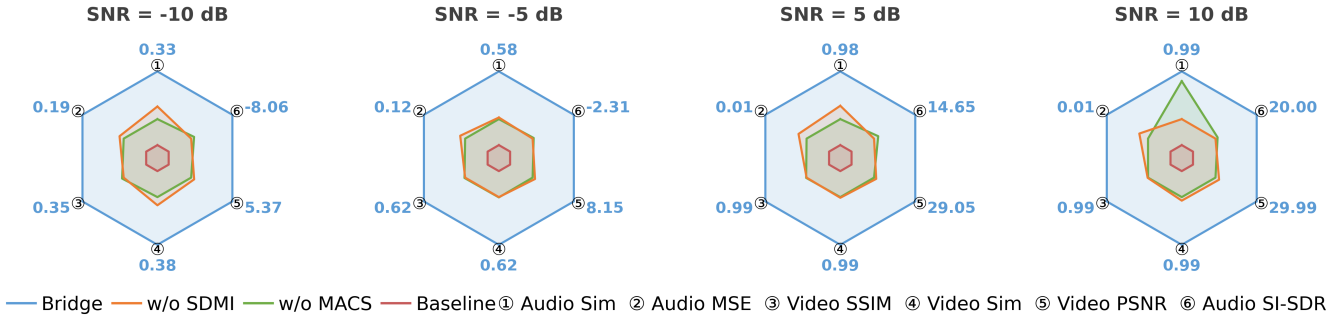

Figure 5: Ablation results comparing Bridge across multiple SNR levels.

Ablation Study. To quantify the contribution of the two core components in Bridge—*SDMI* for shared/specific disentanglement and semantic purification, and *MACS* for cross-modal contrastive semantic alignment—we conduct a controlled ablation on VGGSound under four representative SNRs (-10, -5, 5, and 10 dB). Figure 5 visualizes the results with six metrics (Audio Sim, Audio MSE, Audio SI-SDR, Video Sim, Video PSNR, and Video SSIM), where larger values are better except for MSE. In Figure 5, the farther the value is from the center, the better the performance.

Across all SNRs, the full Bridge yields the outermost radar profile, indicating the best overall multimodal reconstruction quality. Removing SDMI causes a uniform degradation in both modalities, especially in low SNR regimes: at -10 dB, Audio Sim drops from 0.223 to 0.198, SI-SDR decreases from -11.09 to -13.06 dB, and Video PSNR falls from 5.58 to 4.53 dB, suggesting that SDMI is crucial for suppressing modality-specific noise leakage and stabilizing semantic codes under severe channel corruption. Removing MACS leads to an even larger decline in cross-modal consistency, reflected by simultaneous drops in Audio/Video Sim and perceptual fidelity; e.g., at -10 dB the Video Sim decreases from 0.207 to 0.133 and SSIM from 0.213 to 0.124, highlighting MACS as the main driver for maintaining aligned semantics across audio–video streams. As SNR increases, all variants improve, but Bridge preserves a clear margin, particularly on signal-level quality: at 10 dB, Bridge reaches 15.23 dB SI-SDR and 22.50 dB PSNR, while the two ablations remain lower (e.g., 14.29/21.63 for *w/o SDMI* and 10.45/19.69 for *w/o MACS*). The baseline without these components stays inside the radar across all SNRs, confirming that Bridge's robustness stems from the effect of SDMI-based disentanglement and MACS-based cross-modal alignment.

5 Conclusion

This paper presented Bridge, a novel cross-modal learning framework that effectively bridges modal gaps in semantic communication through unified representation learning. Our key innovations include: *Disentangled Feature Learning for Unified Cross-Modal Representation*, which disentangles modality-specific and modality-invariant features to improve both intra-modal precision and inter-modal generalization; and *Improved Cross-Modal Alignment under Noisy Channel Conditions*, which aligns heterogeneous features in a shared latent space to preserve semantic consistency under noise and fading. Extensive experiments on multiple multimodal benchmarks further demonstrate that BRIDGE achieves strong robustness in low-SNR environments and preserves cross-modal semantic with stable task performance.

Acknowledgments

This work is supported by the Natural Science Foundation of Fujian Province (No. 2024J09032, No. 2025J01379, No. 2025H0043, No. 2025J02019), the Joint Funds for the Innovation of Science and Technology of Fujian Province (No. 2024Y9491).

References

[Akbari *et al.*, 2021] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: transformers for multimodal self-supervised learning from raw video, audio and text. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[Bao *et al.*, 2022] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: unified vision-language pre-training with mixture-of-modality-experts. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

[Ben Zion *et al.*, 2024] Rotem Ben Zion, Boaz Carmeli, Orr Paradise, and Yonatan Belinkov. Semantics and spatiality of emergent communication. In *Advances in Neural Information Processing Systems*, pages 110156–110196, 2024.

[Chen *et al.*, 2020] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020.

[Chen *et al.*, 2024] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. Eat: Self-supervised pre-training with efficient audio transformer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3807–3815, 2024.

[Cheng *et al.*, 2024] Shiqi Cheng, Xuefei Zhang, Yao Sun, Qimei Cui, and Xiaofeng Tao. Knowledge discrepancy oriented privacy preserving for semantic communication. *IEEE Transactions on Vehicular Technology*, 73(8):11637–11646, 2024.

[Cheng *et al.*, 2025] Lu Cheng, Hongliang Zhang, Boya Di, Dusit Niyato, and Lingyang Song. Large language models empower multimodal integrated sensing and communication. *IEEE Communications Magazine*, 63(5):190–197, 2025.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics*, pages 4171–4186, 2019.

[Elizalde *et al.*, 2023] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023.

[Hwang *et al.*, 2024] Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. EVEREST: Efficient masked video autoencoder by removing redundant spatiotemporal tokens. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20889–20907, 2024.

[Li *et al.*, 2021] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. Align before fuse: vision and language representation learning with momentum distillation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference*, pages 740–755, 2014.

[Liu *et al.*, 2024] Yinqiu Liu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Ping Zhang, and Xuemin Shen. Cross-modal generative semantic communications for mobile aigc: Joint semantic encoding and prompt engineering. *IEEE Transactions on Mobile Computing*, 23(12):14871–14888, 2024.

[Lu *et al.*, 2024] Zhilin Lu, Rongpeng Li, Kun Lu, Xianfu Chen, Ekram Hossain, Zhifeng Zhao, and Honggang Zhang. Semantics-empowered communications: A tutorial-cum-survey. *IEEE Communications Surveys Tutorials*, 26(1):41–79, 2024.

[Miao *et al.*, 2024] Xupeng Miao, Shenhan Zhu, Fangcheng Fu, Ziyu Guo, Zhi Yang, Yaofeng Tu, Zhihao Jia, and Bin Cui. X-former elucidator: Reviving efficient attention for long context language modeling. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8179–8187, 2024.

[Nguyen *et al.*, 2024] Loc X. Nguyen, Huy Q. Le, Ye Lin Tun, Pyae Sone Aung, Yan Kyaw Tun, Zhu Han, and Choong Seon Hong. An efficient federated learning framework for training semantic communication systems. *IEEE Transactions on Vehicular Technology*, 73(10):15872–15877, 2024.

[Ovalle *et al.*, 2017] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[Peng *et al.*, 2025] Xiang Peng, Zhijin Qin, Xiaoming Tao, Jianhua Lu, and Khaled B. Letaief. A robust image semantic communication system with multi-scale vision transformer. *IEEE Journal on Selected Areas in Communications*, 43(4):1278–1291, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[Shao *et al.*, 2022] Jiawei Shao, Yuyi Mao, and Jun Zhang. Learning task-oriented communication for edge inference: An information bottleneck approach. *IEEE Journal on Selected Areas in Communications*, 40(1):197–211, 2022.

[Tang *et al.*, 2024] Jianhang Tang, Jiangtian Nie, Jingpan Bai, Ji Xu, Shaobo Li, Yang Zhang, and Yanli Yuan. Uav-assisted digital-twin synchronization with tiny-machine-learning-based semantic communications. *IEEE Internet of Things Journal*, 11(17):28437–28451, 2024.

[Wang *et al.*, 2023a] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186, 2023.

[Wang *et al.*, 2023b] Yang Wang, Zhen Gao, Dezhi Zheng, Sheng Chen, Deniz Gündüz, and H. Vincent Poor. Transformer-empowered 6g intelligent networks: From massive mimo processing to semantic communication. *IEEE Wireless Communications*, 30(6):127–135, 2023.

[Weng *et al.*, 2023] Zhenzi Weng, Zhijin Qin, Xiaoming Tao, Chengkang Pan, Guangyi Liu, and Geoffrey Ye Li. Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications*, 22(9):6227–6240, 2023.

[Xia *et al.*, 2024] Le Xia, Yao Sun, Dusit Niyato, Xiaoqian Li, and Muhammad Ali Imran. Joint user association and bandwidth allocation in semantic communication networks. *IEEE Transactions on Vehicular Technology*, 73(2):2699–2711, 2024.

[Xie *et al.*, 2021] Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. Deep learning enabled semantic communication systems. *IEEE Transactions on Signal Processing*, 69:2663–2675, 2021.

[Xie *et al.*, 2022] Huiqiang Xie, Zhijin Qin, Xiaoming Tao, and Khaled B. Letaief. Task-oriented multi-user semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(9):2584–2597, 2022.

[Xie *et al.*, 2023] Songjie Xie, Shuai Ma, Ming Ding, Yuanming Shi, Mingjian Tang, and Youlong Wu. Robust information bottleneck for task-oriented communication with digital modulation. *IEEE Journal on Selected Areas in Communications*, 41(8):2577–2591, 2023.

[Yang *et al.*, 2023] Wanting Yang, Hongyang Du, Zi Qin Liew, Wei Yang Bryan Lim, Zehui Xiong, Dusit Niyato, Xuefen Chi, Xuemin Shen, and Chunyan Miao. Semantic communications for future internet: Fundamentals, applications, and challenges. *IEEE Communications Surveys Tutorials*, 25(1):213–250, 2023.

[Ye *et al.*, 2025] Peigen Ye, Jingpu Duan, Hongyang Du, and Yulan Guo. User-intent-driven semantic communication via adaptive deep understanding. *ArXiv*, abs/2508.05884, 2025.

[Zadeh *et al.*, 2018] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[Zhang *et al.*, 2024] Guangyi Zhang, Qiyu Hu, Zhijin Qin, Yunlong Cai, Guanding Yu, and Xiaoming Tao. A unified multi-task semantic communication system for multimodal data. *IEEE Transactions on Communications*, 72(7):4101–4116, 2024.

[Zhang *et al.*, 2026] Maojun Zhang, Haotian Wu, Guangxu Zhu, Richeng Jin, Xiaoming Chen, and Deniz Gündüz. Semantics-guided diffusion for deep joint source-channel coding in wireless image transmission. *IEEE Transactions on Wireless Communications*, 25:1547–1564, 2026.

[Zhao *et al.*, 2024] Bowen Zhao, Huanlai Xing, Lexi Xu, Yang Li, Li Feng, Jincheng Peng, and Zhiwen Xiao. On forecasting-oriented time series transmission: A federated semantic communication system. *IEEE Transactions on Mobile Computing*, 23(12):13728–13744, 2024.

[Zhu *et al.*, 2022] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: learning sparse generalist models with conditional moes. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

[Zhu *et al.*, 2025] Zengle Zhu, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. Synesthesia of machines (som)-enabled multi-task semantic communication system. *IEEE Transactions on Mobile Computing*, pages 1–16, 2025.