Grasp: Refining Semantic Graphs into Purified Knowledge for Cross-Modal Communication

Liang Chen
Fujian Normal University
Fuzhou, China
liangchen011208@gmail.com

Xiaoding Wang
Fujian Normal University
Fuzhou, China
wangdin1982@fjnu.edu.cn

Limei Lin*
Fujian Normal University
Fuzhou, China
linlimei@fjnu.edu.cn

Dajin Wang
Montclair State University
Montclair, USA
wangd@montclair.edu

Zhiquan Liu
Jinan University
Guangzhou, China
zqliu@jnu.edu.cn

Jie Wu
Temple University
Philadelphia, USA
jiewu@temple.edu

Abstract

The explosive growth of multimodal web data demands communication that transmits meaning rather than raw bits. Existing semantic-communication systems often fail under noise, missing modalities, and distribution shifts because they optimize surface features instead of modality-invariant knowledge. We present *Grasp*, a knowledge-centric framework for cross-modal communication. Grasp segments streams into semantic blocks and builds a graph over them; a lightweight Graph Neural Networks (GNN) produces schedulable, importance-weighted representations. At its core is *knowledge purification*: we minimize a conditional mutual information upper bound to perform a three-way disentanglement—strongly related, weakly related, and task-irrelevant components—so that only essential semantics are transmitted while non-essential factors are suppressed. To maintain synchrony, we introduce one-to-two temporal contrastive learning to achieve triple alignment of video, audio, and text despite sampling asynchrony. For efficient transmission, Grasp uses a cross-modal shared vector-quantization codebook—a discrete *knowledge codebook*—updated by multimodal attention. At the receiver, a soft-recovery mechanism leverages this shared knowledge to robustly reconstruct semantics under low signal-to-noise ratio (SNR) or missing modalities, yielding graceful degradation. Across web tasks—including cross-modal retrieval and missing-modality inference—Grasp improves knowledge consistency, semantic fidelity, and downstream performance over strong baselines while maintaining low latency. These results show that communication structured around *purified knowledge* is key to building robust, semantic-aware systems for the modern web.

CCS Concepts

• **Theory of computation → Semantics and reasoning**; • **Computing methodologies → Machine learning**.

*Corresponding author.

WWW '26, Dubai, United Arab Emirates.

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792238

Keywords

Cross-Modal Communication; Semantic Graphs; Multimodal Learning; Semantic Communication

ACM Reference Format:
Liang Chen, Xiaoding Wang, Limei Lin, Dajin Wang, Zhiquan Liu, and Jie Wu. 2026. Grasp: Refining Semantic Graphs into Purified Knowledge for Cross-Modal Communication. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774904.3792238

1 Introduction

The exponential growth of multimodal web data—spanning video, audio, and text—demands communication paradigms that prioritize the transmission of *meaning* over raw bits. While semantic communication aims to address this fundamental shift by encoding and reconstructing task-relevant meaning directly [3], current approaches face significant limitations in handling the complexities of real-world multimodal environments. As applications proliferate across vehicle-to-everything coordination, telemedicine, and Internet of Things systems [6, 8], communication frameworks must operate under stringent bandwidth constraints while maintaining semantic coherence amid channel noise, packet loss, and the intrinsic challenges of multimodal data: sampling asynchrony, ambiguous semantic boundaries, and mismatched granularity. The central challenge escalates under low SNR, missing modalities, or distribution drift, where conventional approaches focused on minimizing reconstruction error often fail to preserve semantic validity and robustness [18, 21].

The root limitation of current methods lies in their focus on surface-level feature alignment rather than extracting and communicating the underlying, *modality-invariant knowledge* that guarantees cross-modal coherence. This semantics-knowledge gap manifests in three fundamental challenges for real-world deployment. First, under the rate-distortion-semantics trade-off [15], existing pipelines remain frame-level reconstruction-centric and lack *transmissible knowledge units* [23] explicitly conditioned on channel constraints. The absence of structured, compact knowledge representations hinders efficient scheduling and compression within fixed bandwidth and latency budgets. Second, shared semantics often remain entangled with modality-specific features, where under noise, asynchrony, or missing data, this entanglement leads to alignment drift and semantic incoherence [14]. This highlights the

critical need to disentangle *what is essential* (the knowledge) from *how it is presented* (the modality). Third, discrete representations (e.g., vector-quantized codebooks) often lack adaptability to maintain semantic integrity under channel impairments, where at low SNR or with missing modalities, bias amplification during dequantization undermines reliable knowledge recovery [30]. Collectively, these challenges impede the simultaneous achievement of low communication cost, high semantic fidelity, and strong robustness [7].

To bridge these gaps, we introduce *Grasp*, a framework that recasts multimodal semantic communication as *knowledge extraction and alignment*. Unlike prior work, Grasp distills and transmits *purified knowledge*—modality-invariant core semantics that remain robust under diverse web conditions. We first build a semantic graph from raw multimodal streams: kernel-based self-representation with spectral regularization [25] segments streams into coherent blocks and links them by semantic relations. A lightweight GNN then extracts enhanced features and estimates block importance, yielding prioritized knowledge units for subsequent processing.

The core innovation of Grasp lies in its knowledge purification stage, where we employ mutual information minimization to explicitly disentangle modality-invariant core knowledge from modality-specific features. Drawing on conditional information theory [4], we derive a decoupling objective based on a conditional CLUB bound, effectively isolating the essential semantics from modality-specific variations. This purification process is further enhanced by a novel one-to-two temporal contrastive criterion that maintains knowledge synchrony across modalities despite sampling asynchrony, ensuring temporal coherence in the purified knowledge representations.

For efficient transmission, the purified knowledge undergoes discretization through a cross-modally shared vector quantization codebook, functioning as a discrete *knowledge codebook*. Updated via EMA and constrained by a cross-modal commitment term, this codebook ensures consistent representation of purified knowledge across different modalities. Finally, at the receiver, a soft knowledge recovery mechanism leverages cross-modal attention to robustly reconstruct semantics under low SNR or missing modalities, enabling graceful degradation while preserving task performance. This integrated pipeline represents a significant departure from conventional semantic communication approaches, as it prioritizes the extraction and communication of purified knowledge rather than attempting to reconstruct surface-level features.

Our work makes four key contributions toward advancing multimodal semantic communication, as follows.

(1)We introduce Grasp as a knowledge-centric framework that leverages graph-structured representations for semantic organization, providing a principled approach to extract and prioritize transmissible knowledge units.

(2)We develop a knowledge purification mechanism that minimizes a conditional mutual information upper bound to perform a three-way disentanglement—strongly related, weakly related, and task-irrelevant components—thereby ensuring robust triple modal alignment.

(3)We design a unified knowledge processing pipeline that integrates graph-based semantic modeling, information-theoretic purification, shared knowledge coding, and attention based recovery into an end-to-end system.

(4)Through extensive experiments across multiple web tasks, we demonstrate that Grasp significantly advances the semantic quality vs. bit-cost frontier while maintaining low latency, with detailed ablations substantiating the value of each component.

Paper Organization. The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 details the Grasp framework. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes with discussions and future directions.

2 Related Work

This section reviews key advancements and challenges in semantic communication, multimodal learning, and knowledge representation, highlighting gaps that motivate our proposed framework.

Semantic Communication and End-to-End Learning. Modern semantic communication, driven by end-to-end learning, has shifted the goal from bit-level reconstruction to semantic fidelity [1, 10]. Yet, these methods remain largely limited to single-modality inputs and idealized channels. This renders them fragile in real-world web and packet networks, which suffer from bandwidth volatility, delay, and loss [24]. Crucially, they lack a structured representation for semantics that can be efficiently scheduled and transmitted—a gap that nascent work on semantic-aware rate control and transport-layer integration has started to address [27].

Multimodal Representation and Alignment. Learning consistent representations across heterogeneous modalities is a cornerstone of multimodal understanding. A second strand focuses on cross-modal representation learning that extracts factors genuinely shared across heterogeneous signals [9]. Contrastive learning, cross-attention, and temporal context modeling have significantly advanced cross-modal alignment under clean, synchronized conditions [11]. Yet, in practical web environments, modalities are often asynchronously sampled, suffer from intermittent availability, and are subject to dynamic distribution shifts. This leads to temporal drift and semantic misalignment, undermining the consistency of the shared semantics [12, 26]. Current methods typically address alignment as a post-hoc or offline process, rather than explicitly enforcing temporal synchrony and factor disentanglement as an integral part of the communication pipeline.

Information-Theoretic Disentanglement and Purification. An information theoretic perspective provides a principled foundation for disentangling shared and private factors. Mutual information estimation and minimization have been employed to improve discriminability and suppress redundancy [2]. However, many approaches focus on marginal mutual information or contrastive lower bounds, lacking explicit conditional control to isolate modality-invariant core knowledge from modality-specific variations [13]. This shortfall becomes critical under channel noise and quantization, where unconstrained specific features can corrupt the transmitted semantics, leading to semantic drift. A rigorous, conditional purification strategy is therefore needed to ensure the stability and purity of the communicated knowledge.

Discrete Representation and Structured Semantic Scheduling. Vector quantization and codebook learning offer a pathway to efficient,

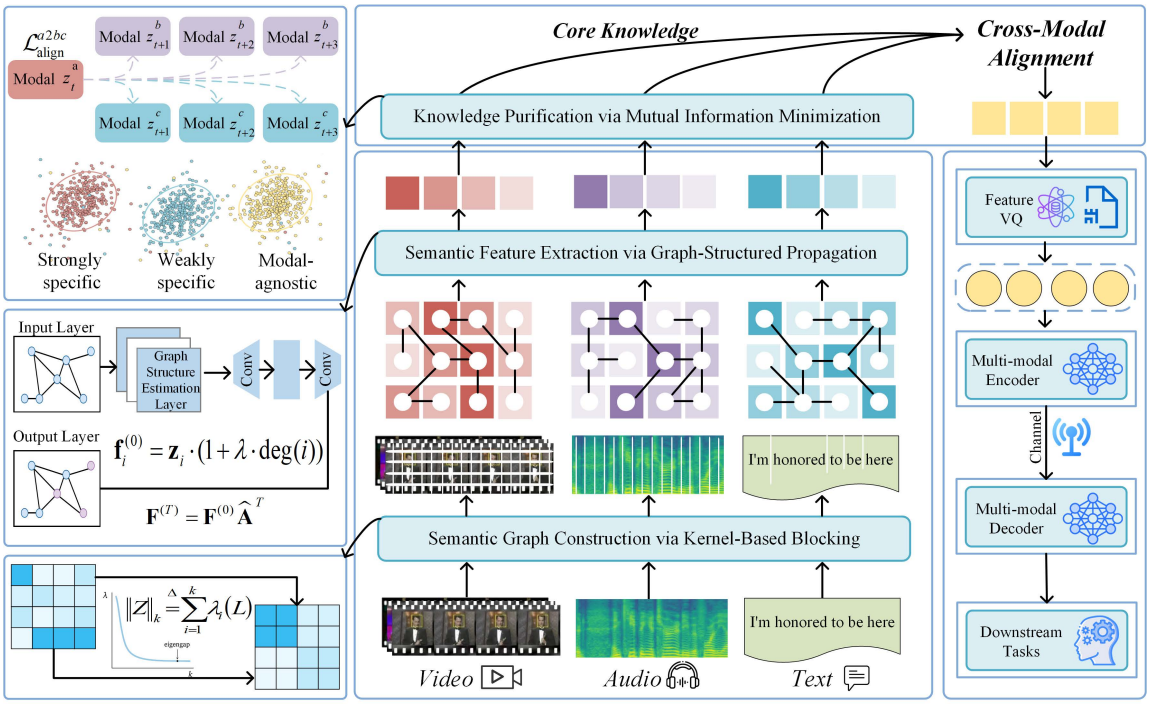

Figure 1: The Grasp Framework. An overview of our knowledge-centric pipeline: from multimodal input, to semantic graph construction, knowledge purification, efficient coding via a shared codebook, and robust semantic recovery at the receiver.

discrete semantic transmission. Stabilization techniques like momentum updates and commitment loss have improved the compactness and training stability of discrete representations [17]. In multimodal settings, key challenges remain: maintaining a consistent, shared discrete space across modalities; mitigating early codebook assignment bias; and enabling robust, soft recovery at the receiver under missing data or corruption [16, 20]. Concurrently, structured modeling—using graphs and self-representation methods—has proven effective for capturing semantic affinities and importance at a block level [22, 28]. However, these techniques have seldom been integrated into an end-to-end communication system that connects semantic structure with network-aware scheduling and rate adaptation [5].

Summary and Motivation. Prior work has advanced semantic communication, multimodal alignment, discrete representation, and structured modeling. However, a key gap remains: there is still no unified, knowledge-centric pipeline that *purifies* and structures semantics into schedulable units, transmits them through a shared discrete codebook robust to network dynamics, and supports reliable recovery under realistic web conditions. This motivates an end-to-end chain that organizes block-level semantics, stabilizes cross-modal consistency and limits semantic leakage before transmission, maps meaning into a shared discrete space, and reconstructs with soft information at the receiver—pushing the rate–semantics frontier toward real packet networks [19]. Grasp bridges this gap by refining raw multimodal data into purified, structured knowledge for consistent and efficient cross-modal communication.

3 Methodology

As shown in Figure 1, we propose *Grasp*, a novel framework that reformulates multimodal semantic communication as a process of *knowledge extraction, purification, and alignment*. It converts raw multimodal inputs (video, audio, text) into structured and purified knowledge representations, tailored for robust transmission over packet-switched networks. Grasp tackles four core challenges: (i) extracting knowledge units under varying bandwidth and latency, (ii) ensuring robustness to SNR fluctuations and packet loss, (iii) achieving cross-modal synchrony despite temporal misalignment, and (iv) optimizing the rate-knowledge trade-off end-to-end.

3.1 Semantic Graph Construction via Kernel-Based Blocking

We introduce Semantic Graph Construction via Kernel-Based Blocking (KBB) as the foundational step for constructing structured semantic representations from raw multimodal sequences. This module transforms continuous streams into semantically coherent blocks and estimates their relative importance, forming the initial semantic graph that subsequent knowledge purification stages will refine.

3.1.1 Kernel-Space Self-Representation for Semantic Blocking. Our approach begins by processing each modality stream independently to extract semantically meaningful units. Along the temporal axis, we detect content changes under duration constraints to obtain variable-length candidate segments. For each segment, a pretrained encoder extracts features, which are aggregated via mean pooling

or attention mechanisms to form block embeddings $\{x_i\}_{i=1}^N$ (representing audio windows, video snippets, or text spans). We preserve temporal extents $[s_i, e_i]$ for downstream synchronization.

Rather than applying direct clustering, we model semantic relationships through *kernel-space self-representation*, constructing the kernel matrix $K_{ij} = k(x_i, x_j)$ using a Gaussian RBF kernel as the similarity backbone, as follows.

$$K(x_i, x_j) \;=\; \exp\Big(-\|x_i - x_j\|^2/2\sigma^2\Big). \tag{1}$$

In the reproducing kernel Hilbert space (RKHS), we formulate each block as a linear combination of other blocks, capturing semantic dependencies.

$$\Phi(x_i) \;\approx\; \sum\nolimits_j Z_{ij}\,\Phi(x_j), \tag{2}$$

where $\Phi(\cdot)$ denotes the nonlinear feature map, $Z \in \mathbb{R}^{N\times N}$ is the self-representation coefficient matrix, and $K_{ij} = \langle \Phi(x_i), \Phi(x_j)\rangle$ is the kernel matrix. To ensure interpretability and support subsequent graph construction, we impose structural constraints on Z: symmetry $Z = Z^\top$ (yielding undirected semantic relationships), nonnegativity $Z \geq 0$ (maintaining semantic coherence), zero diagonal $\mathrm{diag}(Z) = 0$ (eliminating self-loops), and column normalization $\mathbf{1}_N^\top Z = \mathbf{1}_N^\top$. Under symmetry, this implies $Z\mathbf{1}_N = \mathbf{1}_N$, making Z doubly stochastic. Each semantic block is represented as a convex combination of other blocks ($\sum_j Z_{ij} = 1, Z_{ij} \geq 0$). These constraints enhance stability and allow Z to directly serve as a within-modality weighted adjacency matrix for graph construction.

3.1.2 Spectral Regularization for Semantic Group Discovery. We transform the self-representation matrix Z into an affinity structure that explicitly reveals latent semantic groupings, enabling the formation of stable semantic blocks for downstream knowledge processing. To enable spectral analysis, we first symmetrize Z by $W = (Z + Z^\top)/2$. To capture intra-block consistency and inter-block separation—key properties for knowledge unit formation—we define the degree matrix $D = \mathrm{diag}(W\mathbf{1}_N)$ and graph Laplacian $L = D - W$. Here $D_{ii} = \sum_j W_{ij}$ represents node i's total semantic connectivity. The Laplacian quadratic form is as follows.

$$x^\top L x = \tfrac{1}{2}\sum\nolimits_{i,j} W_{ij}\,(x_i - x_j)^2, \tag{3}$$

penalizes large differences between strongly connected nodes, naturally encouraging within-group consistency and across-group separation—essential for identifying coherent knowledge units.

Let the eigenvalues of L be ordered as $0 = \lambda_1(L) \leq \lambda_2(L) \leq \cdots \leq \lambda_N(L)$. We employ the spectral regularizer as follows.

$$\|Z\|_k \triangleq \sum\nolimits_{i=1}^k \lambda_i(L). \tag{4}$$

Minimizing this quantity drives the smallest k eigenvalues toward zero, pushing the graph toward k nearly disconnected components—equivalent to a near block-diagonal structure in W that aligns with k semantic groups, directly enforcing the block structure needed for knowledge extraction.

For numerical stability and scale invariance, we apply symmetric degree normalization $G = D^{-1/2} W D^{-1/2}$. Spectral clustering on W (or L) then yields semantic block groups, with within- and between-group affinities aggregated to form the block-level semantic graph for downstream knowledge purification.

Following the kernel-learning formulation, we construct a data-dependent positive semi-definite kernel matrix K (with diagonal shift $\xi > 0$) from G, as follows.

$$K_{ij} = \begin{cases} \exp\big(-2\max(G) + G_{ij}\big), & i \neq j, \\ \sum_{q\neq i} \exp\big(-2\max(G) + G_{iq}\big) + \xi, & i = j. \end{cases} \tag{5}$$

With K fixed, we estimate the optimal self-representation Z by

$$\min_Z \;\tfrac{1}{2}\,\mathrm{Tr}\big(K + Z^\top K Z\big) \;-\; \alpha\,\mathrm{Tr}(KZ) \;+\; \gamma\sum\nolimits_{i=1}^k \lambda_i(L). \tag{6}$$

The first two terms enforce self-representation consistency in the kernel space while preserving local semantic similarity, and the spectral term shapes the affinity toward a block-diagonal structure via Laplacian spectrum manipulation. After obtaining optimal Z, we form $W = (Z + Z^\top)/2$, perform spectral clustering to assign semantic cluster labels, and merge temporally adjacent segments with identical labels to produce the final semantic blocks—the fundamental units for subsequent knowledge purification.

3.2 Semantic Feature Extraction via Graph-Structured Propagation

The foundation of our knowledge purification framework lies in effectively extracting semantically rich features from multimodal data. We transform raw segments into structured semantic representations through graph-based feature propagation, capturing both intrinsic content value and relational semantics.

3.2.1 Graph-Based Semantic Feature Encoding. Given segment-level undirected affinity matrix $\mathbf{W}$ and pre-partitioned semantic blocks $\{\mathcal{B}_i\}_{i=1}^M$, we construct a semantic graph to enable structured feature extraction. Let $\mathcal{I}_i \subset \{1, \ldots, M\}$ denote the index set of segments within block $\mathcal{B}_i$. We compute the block-level adjacency matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$ that captures inter-block semantic relationships.

$$\mathbf{A}_{ij} = \Big(\sum\nolimits_{p\in\mathcal{I}_i}\sum\nolimits_{q\in\mathcal{I}_j} \mathbf{W}_{pq}\Big)/(|\mathcal{I}_i|\cdot|\mathcal{I}_j|), \qquad \mathbf{A}_{ii} = 0. \tag{7}$$

This defines our semantic feature graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ where the vertex set $\mathcal{V} = \{\mathcal{B}_i\}_{i=1}^M$ represent semantic blocks, the edge set $\mathcal{E} = \{(i,j) \mid A_{ij} > 0\}$ encode their semantic affinities, and $A = (A_{ij})$. The graph structure serves as a relational inductive bias for feature extraction, ensuring that semantically related blocks influence each other's representations.

3.2.2 Iterative Semantic Feature Enhancement. We employ an iterative propagation mechanism to enhance semantic features through graph neighborhoods. First, we normalize the adjacency matrix to obtain feature propagation weights.

$$\widehat{\mathbf{A}}_{ij} = \exp(\mathbf{A}_{ij}/\tau)/\sum\nolimits_{k=1}^M \exp(\mathbf{A}_{ik}/\tau), \tag{8}$$

where τ controls the selectivity of semantic influence.

Each block's semantic feature is initialized with its embedded representation $\mathbf{z}_i$, augmented by structural importance, as follows.

$$\mathbf{f}_i^{(0)} = \mathbf{z}_i \cdot (1 + \lambda \cdot \deg(i)), \tag{9}$$

where $\deg(i)$ is the node degree and λ balances intrinsic semantics versus structural context. Semantic features are then refined through T iterations of neighborhood aggregation.

$$\mathbf{F}^{(t)} = \mathbf{F}^{(t-1)}\widehat{\mathbf{A}}, \quad t = 1, \ldots, T \;\;\Rightarrow\;\; \mathbf{F}^{(T)} = \mathbf{F}^{(0)}\widehat{\mathbf{A}}^T. \tag{10}$$

This propagation mechanism performs *semantic feature enrichment* by iteratively blending features across semantically related blocks. The process enhances discriminative semantic patterns while suppressing noise, resulting in features that capture both local content and global semantic context.

3.2.3 Semantic Salience Scoring and Feature Prioritization. To quantify the semantic importance of each block, we compute salience scores from the enhanced features.

$$s_i = \|\mathbf{f}_i^{(T)}\|_2 / \max(1, \sqrt{d_i}), \tag{11}$$

The semantic salience scores are normalized into transmission priorities.

$$p_i = \exp(s_i) / \sum_{j=1}^{M} \exp(s_j). \tag{12}$$

These priorities p_i are further used as a score-guided routing signal to obtain an initial triplet decomposition of the enhanced block features $\mathbf{f}_i^{(T)}$. Specifically, each block is assigned (or weighted) into three preliminary components: $\mathbf{x}'_i$ denoting modality-invariant core candidates, $\mathbf{y}'_i$ capturing strongly modality-specific yet semantically relevant information, and $\mathbf{z}'_i$ representing weakly modality-specific features and noise.

3.3 Knowledge Purification via Mutual Information Minimization

We introduce the Mutual Information Minimization and Cross-Modal Alignment (MICA) module, which operates on the semantic block representations constructed in the previous stage. MICA addresses the core challenge of distilling modality-invariant *purified knowledge* from modality-specific representations through two synergistic components: (1) Conditional Mutual Information Minimization to explicitly disentangle shared semantics from modality-specific factors, and (2) Anchored Joint Alignment to maintain temporal synchrony of semantic trajectories across modalities despite sampling asynchrony.

3.3.1 Knowledge Purification via Conditional MI Minimization. At the heart of our knowledge purification approach is the explicit separation of semantic representations into distinct components that capture different aspects of the multimodal data. For each modality $m \in \mathcal{M} = \{a, b, c\}$, we further refine the score-routed preliminary triplet features $(\mathbf{x}'^m_i, \mathbf{y}'^m_i, \mathbf{z}'^m_i)$ into three complementary components, as follows.

$$x_i^m = g_{\text{inv}}(\mathbf{x}'^m_i), \quad y_i^m = g_{\text{str}}(\mathbf{y}'^m_i), \quad z_i^m = g_{\text{weak}}(\mathbf{z}'^m_i), \tag{13}$$

where x represents the *modal-agnostic core knowledge*—the purified semantic content that should be invariant across modalities; y captures *strongly modal-specific characteristics* that are distinctive to each modality but semantically relevant; and z contains *weakly modal-specific features* and noise that should be discarded during knowledge transmission.

To achieve effective knowledge purification, we minimize an upper bound of the conditional mutual information between the shared and specific components. This optimization explicitly encourages the separation of core knowledge from modality-specific artifacts. Using a learnable estimator $q_\theta(y \mid x, z)$, we define

$$\begin{aligned} I_{\text{cmi}}(x; y \mid z) = \mathbb{E}_{p(x,y,z)} \left[\log q_\theta(y \mid x, z)\right] \\ - \mathbb{E}_{p(z)\,p(x|z)\,p(y|z)} \left[\log q_\theta(y \mid x, z)\right]. \end{aligned} \tag{14}$$

The first term maximizes the conditional log-likelihood on matched triplets, guiding the estimator toward the true conditional distribution of modality-specific features given the shared knowledge and weak specifics. The second term evaluates the same quantity under a conditional independence assumption, serving as a contrastive baseline that encourages statistical independence between x and y when conditioned on z. Aggregating across all modalities yields the complete purification objective.

$$\mathcal{L}_{\text{purify}} = \sum\nolimits_{m \in \mathcal{M}} I_{\text{cmi}}(x^m; y^m \mid z^m). \tag{15}$$

By minimizing the difference between the joint expectation and the conditional independence baseline, we systematically reduce the conditional mutual information $I(x; y \mid z)$, thereby purifying the shared component x and isolating the modality-invariant core knowledge essential for robust cross-modal communication.

3.3.2 Temporal Knowledge Synchronization via Anchored Joint Alignment. To maintain semantic coherence across modalities despite temporal sampling asynchrony, we introduce an *anchored joint alignment* mechanism that implements the "one-to-two" temporal contrastive learning scheme mentioned in the abstract. This approach ensures that the purified knowledge remains synchronized across modalities over time, addressing the critical challenge of dynamic semantic alignment in real-world web data.

The alignment operates through an *anchor modality* that provides temporal context, while the remaining modalities are constrained to maintain joint consistency with this anchor across multiple future steps. Let modality a serve as the anchor with contextual representation, as follows.

$$h_t^a = \text{ContextAgg}\left(\{x_\tau^a\}_{\tau \le t}\right), \tag{16}$$

which aggregates the purified knowledge components up to time t, capturing the evolving semantic context.

For each future horizon $k = 1, \ldots, K$, we employ step-dependent projection matrices W_k^a to map the anchor context to an appropriate future-alignment space. The target representations from the other modalities are denoted as z_{t+k}^b and z_{t+k}^c. To enable effective contrastive learning, we define the in-batch negative pair set.

$$Z_{bc}(t, k) = \{(z_j^b, z_j^c) \mid j \in \mathcal{B},\ j \ne t\}, \tag{17}$$

which contains mismatched temporal pairs from the same batch.

The anchored joint alignment objective is then formulated as

$$\mathcal{L}_{\text{align}}^{a2bc} = \frac{-1}{K} \sum_{k=1}^{K} \log \left[\frac{\exp\left(z_{t+k}^b W_k^a h_t^a + z_{t+k}^c W_k^a h_t^a\right)}{\sum_{(z_j^b, z_j^c) \in Z_{bc}(t,k)} \exp\left(z_j^b W_k^a h_t^a + z_j^c W_k^a h_t^a\right)} \right]. \tag{18}$$

This formulation possesses several key properties that make it particularly suitable for knowledge synchronization:

(1) **Joint Consistency:** The numerator computes the joint score of the positive pair (z_{t+k}^b, z_{t+k}^c) conditioned on the anchor h_t^a, while the denominator aggregates joint scores of all negative pairs. The additive structure of the joint score ($\exp(u + v) = \exp(u)\exp(v)$) corresponds to a product of per-modality likelihoods under the anchor context, enforcing genuine multi-modal consistency rather than mere pairwise alignment.

(2) **Temporal Robustness:** By aligning across multiple future horizons ($k = 1, \ldots, K$), the method maintains semantic synchrony despite temporal misalignments and varying sampling rates common in web multimedia.

(3) **Flexible Configuration:** The framework naturally adapts to different scenarios—setting $K = 1$ and $k = 0$ with $h_t^a = x_t^a$ yields same-time alignment, while with only two modalities the objective reduces to standard pairwise cross-modal alignment.

The complete MICA objective combines both components is

$$\mathcal{L}_{\text{MICA}} = \mathcal{L}_{\text{purify}} + \lambda \mathcal{L}_{\text{align}}, \tag{19}$$

where λ balances the knowledge purification and temporal alignment objectives. Together, these mechanisms ensure that the system extracts and maintains synchronized, purified knowledge representations—the foundation for robust cross-modal communication in dynamic web environments.

3.4 Knowledge Codec and Cross-Modal Enhancement

Building upon the purified knowledge representations from previous stages, we now introduce the discrete knowledge coding and enhancement mechanism that enables efficient and robust cross-modal communication. This component implements the shared vector quantization codebook described in the abstract, functioning as a discrete knowledge repository that facilitates semantic-consistent transmission and graceful degradation under channel impairments.

3.4.1 Shared Knowledge Codebook with Cross-Modal Commitment. The purified block-level knowledge representations undergo discretization into semantic codes through a cross-modally shared vector quantization (VQ) codebook. This codebook serves as the fundamental *knowledge vocabulary* for all modalities, ensuring that identical semantic concepts map to consistent discrete representations regardless of their originating modality.

For each semantic block, the nearest-neighbor encoding in the shared knowledge space is formulated as

$$q(x_i^m) = e_k \quad \text{where} \quad k = \arg\min_j \|\phi^m(x_i^m) - e_j\|_2, \tag{20}$$

where $\phi^m(\cdot)$ denotes the modality-specific projection head, and e_j are the codeword embeddings in the shared knowledge codebook.

To enforce semantic consistency across modalities, we introduce a *cross-modal commitment loss* that regularizes the same semantic content to converge to consistent discrete representations as

$$\begin{aligned}\mathcal{L}_{\text{commit}}^a = \beta \left\|\phi^a(x_i^a) - \text{sg}[e_i^a]\right\|_2^2 + \frac{\beta}{4}\left\|\phi^a(x_i^a) - \text{sg}[e_i^b]\right\|_2^2 \\ + \frac{\beta}{4}\left\|\phi^a(x_i^a) - \text{sg}[e_i^c]\right\|_2^2,\end{aligned} \tag{21}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, e_i^a, e_i^b, e_i^c are the codeword representations of the same semantic content across three modalities, and β controls the commitment strength. This loss ensures that the purified knowledge components from different modalities representing the same underlying semantics are mapped to proximate regions in the discrete code space.

The knowledge codebook dynamically adapts to multimodal distributions through an exponential moving average (EMA) update mechanism that accumulates cross-modal statistics, as follows.

$$N_i^{(t)} = \gamma N_i^{(t-1)} + (1-\gamma)\left[n_i^{a(t)} + n_i^{b(t)} + n_i^{c(t)}\right], e_i^{(t)} = o_i^{(t)}/N_i^{(t)}, \tag{22}$$

where $\gamma \in (0,1)$ is the decay rate, $n_i^{m(t)}$ counts assignments to codeword e_i from modality m at step t, and $o_i^{(t)}$ aggregates the feature contributions.

Crucially, the codebook update integrates cross-modal knowledge through attention-guided feature enhancement, as follows.

$$\begin{aligned}\mathbf{o}_i^{(t)} =& \gamma\, \mathbf{o}_i^{(t-1)} + \frac{1-\gamma}{2}\Bigg[\sum\nolimits_{j=1}^{n_i^{a(t)}}\left(\mathbf{z}_{i,j}^{a(t)} + \mathbf{r}_{i,j}^{bc(t)}\right) \\ &+ \sum\nolimits_{j=1}^{n_i^{b(t)}}\left(\mathbf{z}_{i,j}^{b(t)} + \mathbf{r}_{i,j}^{ac(t)}\right) + \sum\nolimits_{j=1}^{n_i^{c(t)}}\left(\mathbf{z}_{i,j}^{c(t)} + \mathbf{r}_{i,j}^{ab(t)}\right)\Bigg].\end{aligned} \tag{23}$$

Here, $\mathbf{z}_{i,j}^{m(t)}$ represents the j-th knowledge representation from modality m assigned to codeword e_i, while $\mathbf{r}^{\cdot\cdot}$ denotes complementary features derived through cross-modal attention. This mechanism preserves a unified knowledge repository while incorporating multimodal evidence, thereby stabilizing the discrete semantic space and mitigating assignment biases during early training stages.

3.4.2 Cross-Modal Knowledge Enhancement for Robust Recovery. At the receiver side, we implement a soft recovery mechanism that leverages the shared knowledge base to reconstruct semantics under challenging conditions such as low SNR or missing modalities. This process begins with soft demodulation and channel decoding that produce codeword confidence scores $s_{i,j}^m$ for modality m, block i, and codeword index $j \in \{1, \ldots, M\}$.

The posterior distribution over the knowledge codewords and the corresponding soft reconstruction are computed as

$$\pi_{i,j}^m = \exp(s_{i,j}^m)/\sum\nolimits_{\ell=1}^{M}\exp(s_{i,\ell}^m), \qquad \hat{z}_i^{\,m} = \sum\nolimits_{j=1}^{M}\pi_{i,j}^m e_j, \tag{24}$$

where e_j denotes the j-th codeword embedding from the shared knowledge codebook. This soft dequantization provides a probabilistic reconstruction that gracefully handles uncertainty in the received signals.

We then enhance the recovered knowledge through *cross-modal knowledge attention*, which refines the target modality representation by incorporating complementary evidence from auxiliary modalities. Let the scaled dot-product attention be defined as

$$\text{Attn}(Q, K, V) = \text{softmax}\left(QK^\top/\sqrt{d}\right)V. \tag{25}$$

The knowledge refinement process for the modality a using evidence from the modalities b and c continues as follows.

$$r_i^{bc} = \text{Attn}\left(Q = \hat{z}_i^{\,a},\ K = [\hat{z}_i^{\,b}, \hat{z}_i^{\,c}],\ V = [\hat{z}_i^{\,b}, \hat{z}_i^{\,c}]\right), \tag{26}$$

where r_i^{bc} represents the residual knowledge collected from the auxiliary modalities. The target representation is then augmented through $\tilde{z}_i^{\,a} = \hat{z}_i^{\,a} + r_i^{bc}$.

This cross-modal knowledge enhancement mechanism embodies the soft recovery principle outlined in the abstract: it leverages the shared knowledge base to robustly reconstruct semantics by selectively attending to complementary evidence from other modalities. When certain modalities are corrupted or missing, the attention

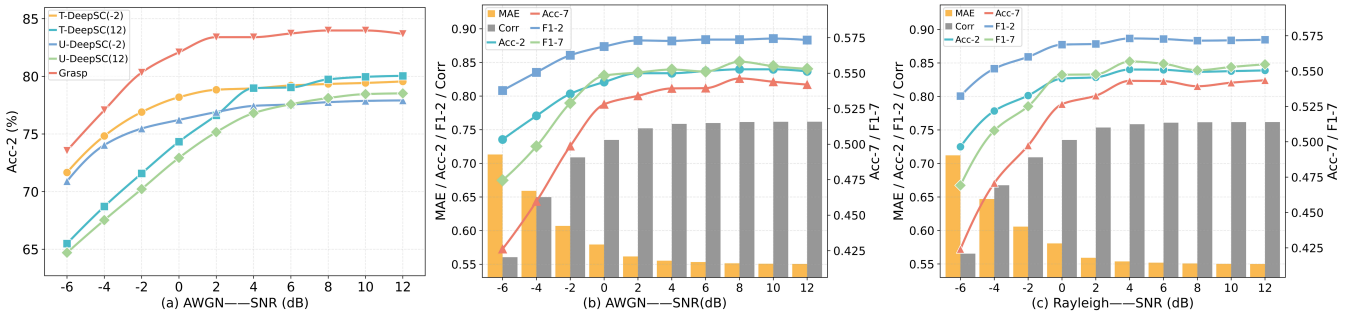

Figure 2: Experimental results on the CMU-MOSEI dataset. Grasp is compared with representative baselines under different SNR conditions, and additional evaluations are reported under AWGN and Rayleigh fading channels.

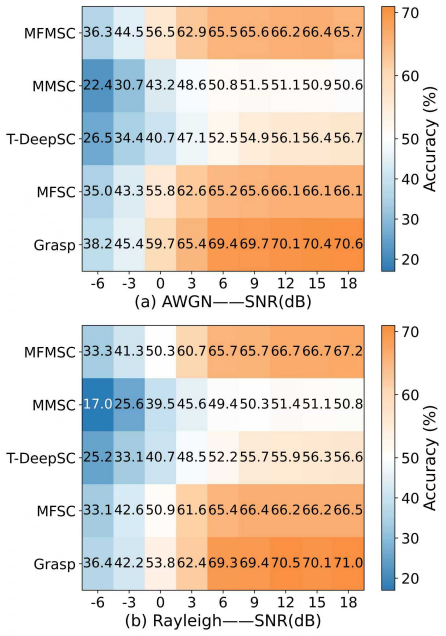

Figure 3: Comparison of Grasp with existing methods on the VQA-v2 dataset under different SNR conditions.

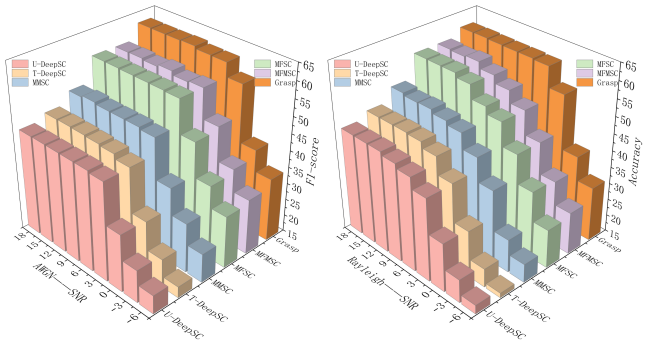

Figure 4: Performance comparison of Grasp and baseline methods on the MM-IMDB dataset under AWGN and Rayleigh channels.

mechanism automatically reweights the available evidence, ensuring that the system maintains semantic fidelity through intelligent knowledge fusion rather than relying on fragile single-modality reconstructions.

4 Experiments

To evaluate Grasp, we conduct experiments on three multimodal benchmarks (see Appendix A): **CMU-MOSEI** for video-audio-text sentiment analysis, **VQA-v2** for visual question answering, and **MM-IMDB** for genre classification. This diverse setup tests robustness across modalities and tasks under practical constraints like channel interference. (More experimental setup see Appendix B).

Results on CMU-MOSEI. Figure 2 compares the performance of Grasp with T-DeepSC and U-DeepSC [29] under different channel conditions on the CMU-MOSEI dataset, which involves video, audio, and text for multimodal sentiment analysis. Grasp consistently outperforms the baselines across all SNR levels, achieving an Acc-2 of 73.55% at −6 dB, outperforming T-DeepSC and U-DeepSC by 1.9% and 2.6%, respectively. At 8 dB, the Acc-2 rises to 83.98%, demonstrating strong robustness under noisy conditions.

Additionally, regression and seven-class evaluations show steady improvements in MAE and F1-7 scores. MAE decreases from 0.7132 at −6 dB to 0.5144 at 8 dB, while the F1-7 score improves by over 0.08, confirming stable cross-modal sentiment understanding.

Under the Rayleigh fading channel, Grasp remains robust, with Acc-2 rising from 72.5% at −6 dB to 83.86% at 8 dB. Despite slightly lower performance compared to AWGN, the consistent upward trend verifies Grasp's noise tolerance and adaptability.

Results on VQA-v2. As shown in Figure 3, Grasp achieves the best performance across all SNR conditions on the VQA-v2 dataset. At −6 dB, it attains 38.21% and 36.41% accuracy for two test configurations, surpassing MFSC [31] by around 3%. When SNR rises to 0 dB, Grasp reaches 59.71% and 53.83%, outperforming T-DeepSC by 19% and 13%, respectively. Even under high SNR (12–18 dB), it maintains the highest scores (∼ 70%), illustrating stable upper-bound performance. The performance gap widens under lower SNR, highlighting Grasp's robustness to channel noise, which stems from its

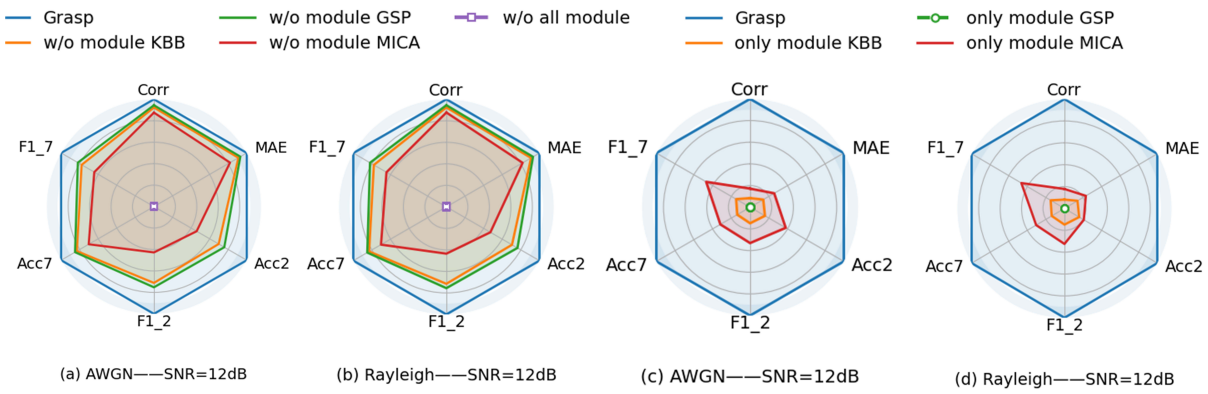

Figure 5: Ablation study of Grasp framework under AWGN and Rayleigh fading channels at an SNR of 12 dB. (a)–(b) show the performance when removing specific modules , while (c)–(d) present results when retaining only a single module .

ability to extract and transmit purified, modality-invariant semantics. These results demonstrate that Grasp's semantic refinement and cross-modal alignment mechanisms are highly effective in mitigating semantic degradation and ensuring robust vision–language reasoning.

Results on MM-IMDB. Figure 4 reports results under both AWGN and Rayleigh channels. Under AWGN, Grasp achieves an F1-score of 34.02 at −6 dB, outperforming T-DeepSC (18.35) and U-DeepSC (20.60) by over 13–15 points. As the SNR increases, its F1-score rises to 61.97 at 18 dB, maintaining a clear lead over MFMSC (57.48)[31]. Under Rayleigh fading, although performance slightly drops, Grasp still achieves 31.31 at −6 dB and 60.67 at 18 dB, remaining consistently superior. These results verify that Grasp effectively preserves essential cross-modal semantics and resists channel-induced degradation across modalities.

Across CMU-MOSEI, VQA-v2 and MM-IMDB, Grasp consistently achieves state-of-the-art results in both the AWGN and Rayleigh channels. It demonstrates significant robustness at low SNR, steady performance improvement with increasing channel quality, and superior cross-modal generalization. These results confirm that Grasp can effectively extract and transmit purified semantic representations across heterogeneous multimodal tasks and challenging communication environments.

Ablation Study. To further verify the contribution of each key component in Grasp, we conduct a comprehensive ablation study by progressively removing or isolating the three major modules: (1) **Kernel-Based Blocking (KBB)** for semantic graph construction, (2) **Graph-Structured Propagation (GSP)** for relational semantic enhancement, and (3) **Mutual Information-Based Knowledge Purification (MICA)** for modality-invariant knowledge extraction. The results are summarized in Figure 5 (more detailed in Appendix C), which presents a radar chart comparing multiple evaluation metrics across configurations.

As shown, removing **MICA** causes the largest performance degradation across all metrics, particularly in correlation and F1-score, confirming that knowledge purification is crucial for disentangling modality-invariant semantics and suppressing redundant or noisy factors. When **KBB** is excluded, the model exhibits a notable drop in accuracy and semantic consistency, indicating that kernel-based graph construction effectively captures structured dependencies and provides stable semantic grouping. The absence of **GSP** results in reduced overall coherence and weaker robustness under channel variations, verifying that graph propagation significantly enhances relational reasoning among semantic blocks.

When all three modules are combined, Grasp achieves balanced and superior performance across all metrics, as reflected by the radar chart's uniformly expanded contour. This demonstrates that the three components are complementary: KBB ensures structural organization, GSP strengthens contextual feature propagation, and MICA purifies and aligns the extracted knowledge. Together, they form a theoretically coherent and empirically validated foundation for robust multimodal semantic communication.

5 Conclusion

This paper presented *Grasp*, a tri-modal communication framework that transmits purified knowledge rather than raw data. Grasp structures inputs as semantic graphs and performs knowledge purification as a three-way disentanglement—strongly related, weakly related, and task-irrelevant—to isolate modality-invariant semantics. A one-to-two temporal contrastive objective enforces triple alignment across video, audio, and text under sampling asynchrony, while a shared vector-quantized *knowledge codebook* enables efficient transmission. A soft-recovery receiver preserves semantics at low SNR or with missing modalities. Across web tasks, Grasp improves semantic fidelity, robustness, and downstream performance, supporting knowledge-centric design for reliable semantic communication; future work will scale to richer modalities and dynamic settings.

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

[1] Sergio Barbarossa, Danilo Comminiello, Eleonora Grassucci, Francesco Pezone, Stefania Sardellitti, and Paolo Di Lorenzo. 2023. Semantic Communications Based on Adaptive Generative Models and Information Bottleneck. *IEEE Communications Magazine* 61, 11 (2023), 36–41. https://doi.org/10.1109/MCOM.005.2200829

[2] Ivan Butakov, Aleksander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, and Alexey A. Frolov. 2024. Mutual information estimation via normalizing flows. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Article 99, 31 pages. https://doi.org/10.52202/079017-0099

[3] Christina Chaccour, Walid Saad, Mérouane Debbah, Zhu Han, and Harold Vincent Poor. 2025. Less Data, More Knowledge: Building Next-Generation Semantic Communication Networks. *IEEE Communications Surveys Tutorials* 27, 1 (2025), 37–76. https://doi.org/10.1109/COMST.2024.3412852

[4] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 1779–1788. https://proceedings.mlr.press/v119/cheng20b.html

[5] Qifan Fu, Huiqiang Xie, Zhijin Qin, Gregory Slabaugh, and Xiaoming Tao. 2023. Vector Quantized Semantic Communication System. *IEEE Wireless Communications Letters* 12, 6 (2023), 982–986. https://doi.org/10.1109/LWC.2023.3255221

[6] Shaolong Guo, Yuntao Wang, Ning Zhang, Zhou Su, Tom Hao Luan, Zhiyi Tian, and Xuemin Shen. 2025. A Survey on Semantic Communication Networks: Architecture, Security, and Privacy. *IEEE Communications Surveys Tutorials* 27, 5 (2025), 2860–2894. https://doi.org/10.1109/COMST.2024.3516819

[7] Qiyu Hu, Guangyi Zhang, Zhijin Qin, Yunlong Cai, Guanding Yu, and Geoffrey Ye Li. 2023. Robust Semantic Communications With Masked VQ-VAE Enabled Codebook. *IEEE Transactions on Wireless Communications* 22, 12 (2023), 8707–8722. https://doi.org/10.1109/TWC.2023.3265201

[8] Feibo Jiang, Li Dong, Yubo Peng, Kezhi Wang, Kun Yang, Cunhua Pan, and Xiaohu You. 2025. Large AI Model Empowered Multimodal Semantic Communications. *IEEE Communications Magazine* 63, 1 (2025), 76–82. https://doi.org/10.1109/MCOM.001.2300575

[9] Vijay John and Yasutomo Kawanishi. 2022. A Multimodal Sensor Fusion Framework Robust to Missing Modalities for Person Recognition. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia* (Tokyo, Japan) *(MMAsia '22)*. Association for Computing Machinery, New York, NY, USA, Article 28, 5 pages. https://doi.org/10.1145/3551626.3564965

[10] David Burth Kurka and Deniz Gündüz. 2021. Bandwidth-Agile Image Transmission With Deep Joint Source-Channel Coding. *IEEE Transactions on Wireless Communications* 20, 12 (2021), 8081–8095. https://doi.org/10.1109/TWC.2021.3090048

[11] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *Comput. Surveys* 56, 10, Article 264 (2024), 42 pages. https://doi.org/10.1145/3656580

[12] Ronghao Lin and Haifeng Hu. 2023. MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis. *Transactions of the Association for Computational Linguistics* 11 (2023), 1686–1702. https://doi.org/10.1162/tacl_a_00628

[13] Chuanhong Liu, Caili Guo, Yang Yang, Wanli Ni, and Tony Quee Sen Quek. 2024. OFDM-Based Digital Semantic Communication With Importance Awareness. *IEEE Transactions on Communications* 72, 10 (2024), 6301–6315. https://doi.org/10.1109/TCOMM.2024.3397862

[14] Zhicheng Liu, Ali Braytee, Ali Anaissi, Guifu Zhang, Lingyun Qin, and Junaid Akram. 2024. Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1841–1848. https://doi.org/10.1145/3589335.3651971

[15] Xueyan Niu, Bo Bai, Nian Guo, Weixi Zhang, and Wei Han. 2025. Rate–Distortion–Perception Trade-Off in Information Theory, Generative Models, and Intelligent Communications. *Entropy* 27, 4, Article 373 (2025), 19 pages. https://doi.org/10.3390/e27040373

[16] Xiang Peng, Zhijin Qin, Xiaoming Tao, Jianhua Lu, and Lajos Hanzo. 2024. A Robust Semantic Text Communication System. *IEEE Transactions on Wireless Communications* 23, 9 (2024), 11372–11385. https://doi.org/10.1109/TWC.2024.3381950

[17] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. 2025. TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 2545–2555. https://doi.org/10.1109/CVPR52734.2025.00243

[18] Md Kaykobad Reza, Ashley Prater-Bennette, and Muhammad Salman Asif. 2025. Robust Multimodal Learning With Missing Modalities via Parameter-Efficient Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 2 (2025), 742–754. https://doi.org/10.1109/TPAMI.2024.3476487

[19] Sadaf Salehkalaibar, Buu Phan, Ashish Khisti, and Wei Yu. 2023. Rate-Distortion-Perception Tradeoff Based on the Conditional Perception Measure. In *2023 Biennial Symposium on Communications (BSC)*. 31–37. https://doi.org/10.1109/BSC57238.2023.10201822

[20] Sejin Seo, Jihong Park, Seung-Woo Ko, Jinho Choi, Mehdi Bennis, and Seong-Lyun Kim. 2023. Toward Semantic Communication Protocols: A Probabilistic Logic Perspective. *IEEE Journal on Selected Areas in Communications* 41, 8 (2023), 2670–2686. https://doi.org/10.1109/JSAC.2023.3288268

[21] Jiawei Shao, Yuyi Mao, and Jun Zhang. 2022. Learning Task-Oriented Communication for Edge Inference: An Information Bottleneck Approach. *IEEE Journal on Selected Areas in Communications* 40, 1 (2022), 197–211. https://doi.org/10.1109/JSAC.2021.3126087

[22] Shiqi Sun, Zhijin Qin, Huiqiang Xie, and Xiaoming Tao. 2023. Task-Oriented Explainable Semantic Communications Based on Structured Scene Graphs. In *Proceedings of the 2023 IEEE Global Communications Conference (GLOBECOM)*. IEEE, Kuala Lumpur, Malaysia, 3222–3227. https://doi.org/10.1109/GLOBECOM54140.2023.10436793

[23] Weida Wang, Xinyi Tong, Xinchun Yu, and Shao-Lun Huang. 2024. On the rate–distortion–perception–semantics tradeoff in low-rate regime for lossy compression. *Journal of the Franklin Institute* 361, 11, Article 106873 (2024). https://doi.org/10.1016/j.jfranklin.2024.106873

[24] Tong Wu, Zhiyong Chen, Dazhi He, Liang Qian, Yin Xu, Meixia Tao, and Wenjun Zhang. 2024. CDDM: Channel Denoising Diffusion Models for Wireless Semantic Communications. *IEEE Transactions on Wireless Communications* 23, 9 (2024), 11168–11183. https://doi.org/10.1109/TWC.2024.3379244

[25] Kunpeng Xu, Lifei Chen, and Shengrui Wang. 2025. Towards Robust Nonlinear Subspace Clustering: A Kernel Learning Approach. *IEEE Transactions on Artificial Intelligence* (2025), 1–13. https://doi.org/10.1109/TAI.2025.3578585

[26] Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. SeqCare: Sequential Training with External Medical Knowledge Graph for Diagnosis Prediction in Healthcare Data. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2819–2830. https://doi.org/10.1145/3543507.3583543

[27] Ke Yang, Sixian Wang, Jincheng Dai, Xiaoqi Qin, Kai Niu, and Ping Zhang. 2025. SwinJSCC: Taming Swin Transformer for Deep Joint Source-Channel Coding. *IEEE Transactions on Cognitive Communications and Networking* 11, 1 (2025), 90–104. https://doi.org/10.1109/TCCN.2024.3424842

[28] Bowen Zhang, Zhijin Qin, and Geoffrey Ye Li. 2023. Semantic Communications With Variable-Length Coding for Extended Reality. *IEEE Journal of Selected Topics in Signal Processing* 17, 5 (2023), 1038–1051. https://doi.org/10.1109/JSTSP.2023.3300509

[29] Guangyi Zhang, Qiyu Hu, Zhijin Qin, Yunlong Cai, Guanding Yu, and Xiaoming Tao. 2024. A Unified Multi-Task Semantic Communication System for Multimodal Data. *IEEE Transactions on Communications* 72, 7 (2024), 4101–4116. https://doi.org/10.1109/TCOMM.2024.3364990

[30] Chuang Zhao, Hui Tang, Jiheng Zhang, and Xiaomeng Li. 2025. Unveiling Discrete Clues: Superior Healthcare Predictions for Rare Diseases. In *Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 1747–1758. https://doi.org/10.1145/3696410.3714831

[31] Zengle Zhu, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. 2025. Synesthesia of Machines (SoM)-Enabled Multi-Task Semantic Communication System. *IEEE Transactions on Mobile Computing* (2025), 1–16. https://doi.org/10.1109/TMC.2025.3625786

A Dataset and Task Settings

To ensure comprehensive evaluation and reproducibility, we provide additional details on the datasets and task setups used in our experiments. The proposed Grasp framework is evaluated on three representative multimodal benchmark datasets, each covering a distinct semantic communication scenario: , **visual question answering** (VQA-v2), **multimodal sentiment analysis** (CMU-MOSEI) and **multimodal genre classification** (MM-IMDB). All datasets are preprocessed according to their official protocols, and all input features are normalized to ensure cross-modal consistency.

A.1 VQA-v2

The VQA-v2 dataset is a large-scale visual question answering benchmark comprising 265,016 images and more than 1.1 million

Table 1: Ablation results of Grasp under AWGN and Rayleigh channels.

Variant	SNR (dB)	AWGN Channel						Rayleigh Channel					
		CC	MAE	Acc_2	$F1_2$	Acc_7	$F1_7$	CC	MAE	Acc_2	$F1_2$	Acc_7	$F1_7$
Grasp	-6	0.5605	0.7132	0.7355	0.8083	0.4262	0.4746	0.5654	0.7121	0.7250	0.8006	0.4241	0.4692
	0	0.7348	0.5794	0.8209	0.8737	0.5281	0.5483	0.7351	0.5807	0.8267	0.8774	0.5266	0.5474
	6	0.7600	0.5534	0.8372	0.8840	0.5395	0.5512	0.7608	0.5520	0.8398	0.8857	0.5429	0.5552
	12	0.7619	0.5505	0.8368	0.8833	0.5420	0.5531	0.7618	0.5500	0.8389	0.8847	0.5435	0.5548
w/o KBB	-6	0.4873	0.7570	0.6769	0.7420	0.3890	0.4320	0.4837	0.7504	0.6704	0.7408	0.3827	0.4338
	0	0.6737	0.6152	0.7623	0.8181	0.4956	0.5123	0.6767	0.6107	0.7647	0.8164	0.4948	0.5103
	6	0.6990	0.5892	0.7786	0.8284	0.5070	0.5152	0.6974	0.5868	0.7747	0.8267	0.5079	0.5174
	12	0.7009	0.5863	0.7782	0.8277	0.5095	0.5171	0.7005	0.5857	0.7802	0.8290	0.5109	0.5187
w/o GSP	-6	0.5021	0.7340	0.6710	0.7530	0.3957	0.4476	0.5014	0.7325	0.6989	0.7520	0.3971	0.4425
	0	0.6866	0.6010	0.7651	0.8247	0.4967	0.5146	0.6829	0.6015	0.7657	0.8216	0.4946	0.5182
	6	0.7169	0.5724	0.7802	0.8297	0.5034	0.5167	0.7177	0.5778	0.7899	0.8356	0.5133	0.5238
	12	0.7177	0.5780	0.7899	0.8356	0.5133	0.5238	0.7176	0.5775	0.7919	0.8369	0.5147	0.5254
w/o MICA	-6	0.4544	0.7806	0.6257	0.7207	0.4657	0.4736	0.4565	0.7823	0.6247	0.7243	0.4651	0.4708
	0	0.6241	0.6537	0.7186	0.7957	0.4774	0.4866	0.6246	0.6546	0.7171	0.7946	0.4763	0.4875
	6	0.6628	0.6215	0.7342	0.7741	0.4864	0.4965	0.6634	0.6275	0.7312	0.7758	0.4873	0.4935
	12	0.6629	0.6221	0.7322	0.7729	0.4851	0.4950	0.6628	0.6215	0.7340	0.7741	0.4864	0.4965
w/o all modules	-6	-0.0432	1.1366	0.6151	0.6611	0.2808	0.3127	-0.0415	1.1204	0.6107	0.6510	0.2781	0.3148
	0	-0.0365	1.0417	0.6414	0.6895	0.3501	0.3899	-0.0351	1.0573	0.6490	0.6871	0.3538	0.3736
	6	-0.0360	0.9477	0.6431	0.6905	0.3506	0.3905	-0.0344	0.9498	0.6438	0.6931	0.3510	0.3989
	12	-0.0360	0.9477	0.6431	0.6905	0.3506	0.3905	-0.0450	0.9455	0.6399	0.6889	0.3516	0.3912

human-annotated question–answer pairs. Each image–question pair requires cross-modal reasoning between visual and textual modalities to infer the correct answer among multiple candidates. The VQA-v2 task evaluates the ability of Grasp to preserve cross-modal semantic integrity and reasoning capability under noisy transmission conditions.

A.2 CMU-MOSEI

CMU-MOSEI is a large-scale multimodal sentiment analysis dataset containing 23,453 annotated video segments from over 1,000 speakers, covering diverse topics and emotional expressions. Each sample includes synchronized video, audio, and text modalities, annotated on both binary (Acc-2) and seven-class (Acc-7) sentiment scales. This dataset is used to evaluate the robustness of Grasp in fine-grained affective reasoning under varying signal-to-noise ratio (SNR) conditions.

A.3 MM-IMDB

MM-IMDB is a multimodal movie genre classification dataset that combines textual plot summaries with poster images for multilabel classification. Each sample consists of a text–image pair associated with one or more of 23 movie genres. This dataset assesses the capability of Grasp to extract and transmit complementary semantic cues across heterogeneous modalities, highlighting its strength in structured multimodal knowledge fusion.

A.4 Channel and Noise Settings

In all experiments, semantic embeddings are transmitted through simulated wireless channels, including additive white Gaussian noise (AWGN) and Rayleigh fading. Multiple signal-to-noise ratio (SNR) levels are applied to evaluate model robustness under diverse channel conditions.

B Experimental Setup

B.1 Training Configuration

Grasp is implemented in PyTorch and trained with the Adam optimizer (learning rate 5×10^{-5}) for 50 epochs with a batch size of 32. We apply global ℓ_2-norm gradient clipping (threshold 4.0) and a validation-based, patience-10 early-stopping criterion: training is terminated if the validation loss shows no improvement for 10 consecutive epochs. Audio and video features are normalized prior to model input to mitigate cross-modal scale differences. The system supports both AWGN and Rayleigh channels; unless otherwise noted, AWGN is used by default. We vary SNR within $[-6, 12]$ dB and use 12 dB as the default operating point. Loss weights are tuned on the validation set, and all training/channel hyperparameters are configurable via command-line flags.

B.2 Compute Resources

All experiments were conducted on a dedicated high-performance machine equipped with an NVIDIA RTX 5090 GPU (32 GB), an Intel Xeon Platinum 8470Q CPU with 25 cores, and 90 GB of RAM, running Ubuntu 20.04. Each training epoch of the *Grasp* framework typically required about ~20 minutes, depending on the dataset and modality configuration.

C Detailed Ablation Results

As shown in Table 1, we conduct detailed ablation experiments under both AWGN and Rayleigh channels. Removing any of the core components (*KBB*, *GSP*, or *MICA*) leads to a noticeable degradation across all metrics, particularly under low SNR conditions. Among them, the exclusion of the *MICA* module causes the most significant performance drop, highlighting its essential role in cross-modal semantic alignment. Even under severe channel noise, Grasp maintains stable correlation and accuracy, demonstrating the robustness and complementarity of its modular design.