EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Knowledge Graph Fusion for Cross-Modal Semantic Communication

Yanrong Yang[1,*], Tianxiang Zhong[2], and Mengting Chen[3]

[1]Guangdong University of Technology, Guangzhou, 510006, China.
[2]Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham, UK.
[3]Guangdong R&D Center for Technological Economy, Guangzhou, Guangdong 510030, China.

Abstract

This paper proposes a knowledge graph-enhanced multi-source fusion (KG-MSF) scheme, a novel cross-modal semantic communication system to robustly fuse visual and textual data for tasks such as visual question answering (VQA) over wireless channels. The proposed KG-MSF scheme integrates knowledge graph reasoning into a multi-stage fusion and encoding pipeline, utilizing bidirectional cross-attention between modalities and structured semantic triplets to enhance semantic preservation and resilience to channel impairments. Specifically, image objects and question tokens are first aligned via cross-modal attention, then enriched with shallow and deep semantic triplets extracted through knowledge graphs, which are subsequently fused and transmitted using joint source-channel coding. Extensive simulation results are provided to demonstrate that the proposed KG-MSF scheme significantly outperforms the competing ones under both AWGN and Rayleigh fading channels, indicating KG-MSF's superior semantic robustness and efficient cross-modal reasoning in wireless environments.

Received on 09 July 2025; accepted on 11 December 2025; published on 18 December 2025

Keywords: Knowledge graph, cross-modal, semantic communication, performance evaluation.

doi:10.4108/eetsis.9216

1. Introduction

In recent years, knowledge graph (KG) has been extensively analyzed across a range of application domains, including natural language processing, cybersecurity, and engineering management [1–3]. For example, KG-enhanced text understanding has been evaluated, where the effectiveness of multi-hop reasoning for improving inference accuracy was emphasized, while scalability and computational efficiency were enhanced [4]. Additionally, the performance trade-off between different multi-modal fusion architectures was systematically compared, revealing that graph-based feature integration significantly improves cross-modal semantic alignment [5, 6]. In cybersecurity applications, KG-driven entity recognition systems were assessed and shown to achieve superior precision and recall, particularly in handling rare-event detection scenarios. Moreover, entity concept extraction from knowledge graphs for multimodal fake news detection was evaluated, with notable improvements observed in F1 scores when compared to traditional deep learning baselines [7, 8]. Across these analyses, it was consistently demonstrated that integrating knowledge graphs results in significant gains in semantic precision, robustness to noisy or incomplete inputs, and task-specific generalization [9]. At the same time, researchers have investigated the graph scalability, dynamic updating, noise resilience, and real-time inference latency in the field of KG. Dynamic aspects, such as incremental learning, relational noise tolerance, and graph sparsification techniques, have been investigated to sustain high inference efficiency without compromising semantic fidelity, indicating strong potential for future deployment of KGs in performance-critical artificial intelligence (AI) systems.

Semantic communication (SemCom) system has been proposed with increasing attention to the efficiency in

*Corresponding author. Email: yanrongyang123@hotmail.com

transmitting meaningful information over noisy communication channels [10]. In this system, communication efficiency has been enhanced by leveraging semantic-based methods that prioritize message content over raw data transmission [11, 12]. Key performance indicators, including error rates, latency, and throughput, have been extensively evaluated in various communication models. Notably, the use of machine learning-based approaches has shown substantial promise in improving performance by adapting to dynamic channel conditions and optimizing the transmission of semantically relevant data [13]. Performance analyses in wireless and optical communication networks have demonstrated that SemCom can outperform traditional methods by reducing communication cost and enhancing the robustness of information transmission under severe channel impairments [14, 15]. The application of compression technique into semantic communication has achieved a significant reduction in the communication cost, while maintaining high levels of information fidelity [16]. Further, the development of intelligent, reconfigurable systems such as RIS-enabled SemCom has been investigated, where the performance evaluation has revealed improvement in signal quality and reliability, especially in resource-constrained environments [17, 18]. However, challenges related to real-time processing, resource allocation, and scalability remain, requiring continued research into optimizing the balance between semantic richness and system efficiency.

Cross-modal learning has been proposed to optimize the fusion of data from different modalities, such as visual, auditory, and textual information, to enhance task-specific performance in various AI applications [19], where the efficiency and effectiveness of algorithms in integrating and interpreting these heterogeneous data types has been widely investigated. In particular, cross-modal deep learning models, such as those using transformer architectures, have been extensively explored, with key performance metrics including accuracy, computational efficiency, and robustness to noise [20, 21]. Moreover, it has been demonstrated that multi-modal fusion strategies can significantly improve model performance in tasks like visual question answering (VQA), speech recognition, and object detection [22]. Notably, the integration of cross-modal embeddings into SemCom system can help enhance both the semantic alignment and the robustness of AI systems, with substantial improvement in the generalization over single-modality systems [23]. Additionally, the scalability of cross-modal models has been examined, through aligning unstructured data, such as matching images with corresponding text or speech with video frames. Performance benchmarks have revealed that the cross-modal attention mechanisms and generative model, can outperform traditional models, particularly in environments with incomplete or noisy data [24]. In further, the cross-modal learning has been investigated in the imbalanced datasets, where the computational complexity and model interpretability were analyzed [25, 26].

Motivated by the above literature review, this work introduces a knowledge graph-enhanced multi-source fusion (KG-MSF), a novel semantic communication system for robust cross-modal information transmission over wireless channels, with a focus on visual question answering tasks. The proposed KG-MSF scheme extracts semantic features from both images and questions, aligns them using a bidirectional cross-attention mechanism, and enhances the fused representation through structured shallow and deep-level triplets derived from a knowledge graph. This enriched semantic representation is then transmitted using a joint source-channel coding strategy to preserve information integrity under channel impairments. Simulation results are provided to demonstrate the advantages of the proposed KG-MSF scheme over the competing ones. In particular, under Rayleigh fading at 6 dB SNR, KG-MSF reaches 0.3878 accuracy. These results validate that KG-MSF effectively enhances semantic robustness and cross-modal reasoning, offering consistent performance gains across diverse wireless environments.

2. System Model

In this paper, we consider a semantic communication system tailored for cross-modal data sources, where modality-specific semantic representations are extracted, fused, encoded, transmitted, decoded, and finally used for downstream tasks such as visual question answering. The objective of the considered semantic system is to preserve semantic integrity across heterogeneous modalities while ensuring robustness against wireless channel degradation. Let $\mathcal{D} = \{D^1, D^2, \ldots, D^M\}$ denote a collection of input data modalities. Here, each D^i corresponds to the input data of modality i (e.g., image, audio, or text), and M is the total number of modalities considered. To capture the semantic essence of each modality, we define a semantic encoder F_s^i specific to the i-th modality. The extracted semantic representation from modality D^i is given by,

$$\mathbf{S}^i = F_s^i(D^i), \tag{1}$$

where $\mathbf{S}^i = [\mathbf{s}_1, \ldots, \mathbf{s}_{n_i}] \in \mathbb{R}^{L_s^i \times n_i}$ consists of n_i semantic vectors $\mathbf{s}_j \in \mathbb{R}^{L_s^i}$. These vectors can encode embeddings, object entities, relationships, or temporal features, depending on the modality.

To unify the semantic knowledge from heterogeneous modalities, we introduce a fusion function F_{KG} that utilizes a knowledge graph to enhance cross-modal

semantic alignment. The KG allows explicit modeling of relationships among entities across modalities using triplets of the form (h, r, t), where h and t are entities and r is the semantic relation. The fused semantic representation $\mathbf{S}^f$ is computed as:

$$\mathbf{S}^f = F_{\text{KG}}(\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^M), \tag{2}$$

where $\mathbf{S}^f \in \mathbb{R}^{L_f \times n_f}$ and each row or tensor component in $\mathbf{S}^f$ encodes a knowledge-enriched feature vector. The fusion process not only consolidates cross-modal features but also introduces auxiliary information g_{KG}, such as prior entity distributions or contextual embeddings from the KG.

After that, the fused semantic features $\mathbf{S}^f$ are encoded into transmittable symbols using a joint source-channel (JSC) encoder F_C. Unlike traditional separation-based systems, JSC coding simultaneously compresses the source and adds redundancy to resist the channel impairment,

$$\mathbf{X} = F_C(\mathbf{S}^f), \tag{3}$$

where $\mathbf{X} \in \mathbb{C}^{L_x \times 1}$ is a complex-valued vector that satisfies an average power constraint, given by,

$$\mathbb{E}\left[\|\mathbf{X}\|_2^2\right] \leq P. \tag{4}$$

This constraint ensures the transmitted symbols conform to practical hardware and spectral efficiency limit.

During the transmission, the encoded signal $\mathbf{X}$ is perturbed by a fading channel and additive white Gaussian noise (AWGN). The received signal $\mathbf{Y}$ at the receiver is given by,

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{B}, \tag{5}$$

where $\mathbf{H} \in \mathbb{C}^{L_y \times L_x}$ is the channel matrix with i.i.d. elements subject to $\mathcal{CN}(0, 1)$ under Rayleigh fading, and $\mathbf{B} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ is AWGN noise. The instantaneous signal-to-noise ratio (SNR) at the receiver is given by,

$$\text{SNR} = \frac{\|\mathbf{H}\mathbf{X}\|^2}{\sigma^2}. \tag{6}$$

Assuming perfect channel state information (CSI) at the receiver, the transmitted signal $\mathbf{X}$ can be recovered using least squares (LS) channel equalization:

$$\hat{\mathbf{X}} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{Y}, \tag{7}$$

where $\mathbf{H}^H$ denotes the Hermitian transpose of $\mathbf{H}$. Subsequently, the inverse JSC decoder F_C^{-1} is used to reconstruct the semantic features:

$$\hat{\mathbf{S}}^f = F_C^{-1}(\hat{\mathbf{X}}), \tag{8}$$

where $\hat{\mathbf{S}}^f \in \mathbb{R}^{L_f \times n_f}$ and the decoder is often parameterized as a neural network trained end-to-end with F_C to preserve semantic consistency despite channel distortion. Finally, the recovered semantic features are fed into a task-specific inference model. For the VQA task, this function F_A maps the semantic space to the answer space, given by,

$$\mathcal{A} = F_A(\hat{\mathbf{S}}^f), \tag{9}$$

where $\mathcal{A}$ denotes the predicted answer. This stage may involve attention mechanisms, transformer-based reasoning, or logical inference over the semantic triplets and fused information.

3. Proposed Method

In this paper, we propose a knowledge graph-enhanced cross-modal semantic fusion (KG-MSF) scheme to enable efficient and robust fusion of visual and textual information for downstream semantic tasks such as visual question answering. This system consists of two primary stages, where one stage is semantic information extraction from images and questions, and the other one is semantic fusion using structured triplets that incorporate knowledge graph reasoning. These components work in concert to bridge low-level perceptual features and high-level abstract semantics under the influence of knowledge-driven interactions.

In the semantic extraction module, image inputs I_i are processed using a pre-trained object detector to identify a set of visual entities or objects $\{o_1, o_2, \ldots, o_e\}$. For each detected object o_i, two following types of features are extracted: the spatial position feature $\mathbf{f}_i^p \in \mathbb{R}^{d_p}$, and the region of interest (RoI) feature $\mathbf{f}_i^{\text{roi}} \in \mathbb{R}^{d_r}$. These features are projected into a common semantic space using learnable transformations and then combined via layer normalization to form the object embedding, given by,

$$\mathbf{o}_i = \frac{1}{2}\left[\text{LN}(W_p\mathbf{f}_i^p + b_p) + \text{LN}(W_{\text{roi}}\mathbf{f}_i^{\text{roi}} + b_{\text{roi}})\right]. \tag{10}$$

This formulation ensures that both spatial and semantic content are represented in a normalized and comparable form, facilitating the integration with linguistic features.

In parallel, the corresponding question T_i is tokenized into a sequence of words $\{q_1, q_2, \ldots, q_t\}$, where each word token q_j is embedded using two components: a pre-trained word embedding and a learned positional index embedding. These are summed and then normalized to produce the final token representation:

$$\mathbf{q}_j = \text{LN}(\text{WordEmbed}(q_j) + \text{IdxEmbed}(v_j)). \tag{11}$$

This ensures that the question representation captures both semantic and syntactic structures essential for

reasoning, particularly for tasks where word order and relational structure are important.

To effectively align and integrate the visual and textual modalities, the proposed scheme employs a bidirectional cross-attention mechanism. In this module, each image object attends to relevant words in the question, and vice versa. Specifically, for each image-to-text alignment, attention scores are computed as,

$$\tilde{\mathbf{q}}_{i\leftarrow j}^{(l)} = \text{softmax}\left(\left(W_Q \mathbf{q}_i^{(l-1)}\right)^T \left(W_K \mathbf{o}_j^{(l-1)}\right)\right) W_V \mathbf{o}_j^{(l-1)}, \tag{12}$$

while the reciprocal question-to-image alignment is given by,

$$\tilde{\mathbf{o}}_{i\leftarrow j}^{(l)} = \text{softmax}\left(\left(W_Q \mathbf{o}_i^{(l-1)}\right)^T \left(W_K \mathbf{q}_j^{(l-1)}\right)\right) W_V \mathbf{q}_j^{(l-1)}. \tag{13}$$

Here, W_Q, W_K, and W_V are the weight matrices for the query, key, and value components of the Transformer, and l denotes the layer index in the cross-attention block. These equations enable the model to dynamically utilize and integrate relevant visual and linguistic cues.

The fused semantic representation is further enhanced through multi-layer triplet reasoning inspired by knowledge graphs, where shallow-level triplets, such as (o_i, a, v), are constructed to capture observable attribute-object associations, and deep-level triplets, such as (q_i, r, o_j), which encode relational reasoning patterns derived from the co-dependency between questions and visual elements. A knowledge-graph-based fusion function aggregates these structured triplets, given by,

$$\mathbf{S}^f = F_{\text{KG}}^{\text{triplet}}(\tilde{\mathbf{o}}_{i\leftarrow j}, \tilde{\mathbf{q}}_{i\leftarrow j}, \text{Graph}(h, r, t)), \tag{14}$$

where $\text{Graph}(h, r, t)$ represents learned or inferred semantic relations, and $F_{\text{KG}}^{\text{triplet}}$ models their interactions. This enables both direct grounding of semantic elements and higher-order inference across modalities, laying a strong foundation for robust semantic communication in downstream tasks. The shallow-level triplets capture direct correlations between image objects and question-relevant attributes. These are constructed using object representations $\tilde{o}$ containing contextual cues from the question, with attributes derived from the visual features of the objects. Specifically, the head is set as the object embedding $\tilde{o}$, the tail corresponds to the attribute information extracted via a pre-trained module (e.g., attribute classifiers), and the relation is denoted as cls, encapsulating the overall contextual interaction between the question and image. These triplets are grounded in visual semantics and are primarily used to encode observable correlations.

In contrast, the deep-level triplets encapsulate abstract relational reasoning that emerges from logical inference over question-object interactions. These triplets are intended to model the deeper semantic alignment between the question and the most semantically relevant image object. The head of the deep-level triplet, denoted as head_d, is identified by computing a similarity-based attention matrix between the image I and question T, given by,

$$W_C = \left(W_C^T T\right)^T \left(W_C^O I\right), \tag{15}$$

where W_C^T and W_C^O are learnable parameter matrices. The most relevant object to the question is selected via a weighted maximization strategy, given by,

$$\text{head}_d = \sum_{i=1}^{e} F_d\left(\max_j W_C^{ij}\right) \mathbf{o}_i, \tag{16}$$

where F_d is a scoring function over the relevance matrix, and $\mathbf{o}_i$ is the i-th object. The tail of the deep-level triplet is the correct answer to the question, thereby facilitating semantic reasoning from image to language.

As to the model training, the proposed scheme employs a three-stage training pipeline that enables structured reasoning and robust semantic encoding across modalities. In the first stage, the network focuses on learning semantic relations within triplets using the classical TransE objective. For a given triplet (h, r, t), a margin-based ranking loss is defined to encourage semantic consistency, given by,

$$\mathcal{L}_{\text{TransE}} = \sum_{\substack{(h_+, r, t_+)\in P_+ \\ (h_-, r, t_-)\in P_-}} \left[\gamma + d(h_+ + r, t_+) - d(h_- + r, t_-)\right]_+, \tag{17}$$

where $d(u, v)$ is the cosine distance between two vectors, $\gamma > 0$ is a margin hyperparameter, and $[x]_+ = \max(0, x)$. In addition, P_+ and P_- denote the sets of positive and negative triplets, respectively, where negative samples are generated by perturbing the head or tail of valid triplets. This stage ensures that meaningful semantic relationships are preserved and distinguishable in the embedding space.

The second stage involves training the joint source-channel encoder and decoder, which are critical for the transmission and reconstruction of semantic features over noisy wireless channels. The fused semantic representation $\mathbf{S}^f$, obtained from triplet fusion, is passed through the JSC module, and the reconstructed features $\hat{\mathbf{S}}^f$ are compared using a mean squared error objective, given by,

$$\mathcal{L}_{\text{JSC}} = \mathbb{E}\left[\left\|\hat{\mathbf{S}}^f - \mathbf{S}^f\right\|_2^2\right]. \tag{18}$$

Algorithm 1 Knowledge Graph-enhanced Multi-Source Fusion (KG-MSF)

1: **Input:** Image I, question T, knowledge graph $\mathcal{G}$, channel model $p(Y|X)$
2: **Output:** Predicted answer $\hat{A}$
3: **Semantic Extraction:** Detect objects $\{o_i\}$ in I and extract spatial f_i^p and RoI f_i^{roi} features. Form object embedding $o_i = \frac{1}{2}[\text{LN}(W_p f_i^p) + \text{LN}(W_{roi} f_i^{roi})]$. Tokenize T into $\{q_j\}$ and embed each as $q_j = \text{LN}(\text{WordEmbed}(q_j) + \text{IdxEmbed}(v_j))$.
4: **Cross-modal Alignment:** Perform bidirectional cross-attention between $\{o_i\}$ and $\{q_j\}$ to obtain aligned features $\tilde{o}_i$ and $\tilde{q}_j$.
5: **Knowledge-guided Fusion:** Construct shallow triplets (o_i, a, v) and deep triplets (q_i, r, o_j). Fuse them via $\mathcal{G}$ to form semantic representation $S_f = F_{\text{KG}}^{\text{triplet}}(\tilde{o}, \tilde{q}, \mathcal{G})$.
6: **Joint Source-Channel Transmission:** Encode S_f into symbols $X = F_C(S_f)$, transmit over channel, receive Y, and decode $\hat{S}_f = F_C^{-1}(Y)$.
7: **Answer Prediction:** Predict $\hat{A} = F_A(\hat{S}_f)$ using task-specific reasoning (e.g., VQA).

This loss function minimizes distortion introduced during transmission, encouraging semantic preservation under practical communication constraints.

In the third stage, the end-to-end joint training is conducted to simulate real-time semantic inference during practical communication scenarios. After transmission, the shallow-level triplet features and global auxiliary information $\{\hat{q}_1, \ldots, \hat{q}_t, \text{cls}\}$ are reconstructed. The pre-trained extraction module from stage one is reused to recompute head_d, and the answer is inferred from the deep-level representation space. Finally, the prediction is evaluated by using the cross-entropy (CE) loss, given by,

$$\mathcal{L}_A = -P(\mathcal{A}) \log P(\hat{\mathcal{A}}), \tag{19}$$

where $\mathcal{A}$ and $\hat{\mathcal{A}}$ represent the ground-truth and predicted answers, respectively. This step ensures end-to-end consistency and fine-tunes all components, including the triplet construction, JSC modules, and the final task-specific prediction layer. The whole procedure of the proposed knowledge graph-enhanced multi-source fusion is summarized in Algorithm 1.

4. Simulation Results and Discussions

In this part, we assess the performance of the proposed KG-MSF scheme under both AWGN and Rayleigh fading channels. The transmit SNR is swept from -10dB to 20dB to evaluate robustness under varying channel conditions. For each SNR, the answer accuracy is averaged over a large set of visual question answering samples. The dataset includes image-question-answer triplets, and object detection is performed using a pre-trained Faster R-CNN model. Word embeddings are initialized using GloVe vectors, and positional encodings are added to preserve sequence structure. The semantic encoder and decoder are both implemented as transformer-based architectures with multi-head attention layers. The knowledge graph triplets are constructed using ConceptNet, incorporating both shallow attribute-based relations and deep reasoning patterns. During the training, all models are optimized using the Adam optimizer with an initial learning rate of 1e-4, batch size of 64, and trained for 100 epochs. Joint source-channel coding is implemented via a neural encoder-decoder pair trained end-to-end to preserve semantic consistency under channel distortions. Each simulation result is averaged over at least 1000 independent trials with different channel realizations to ensure statistical reliability.

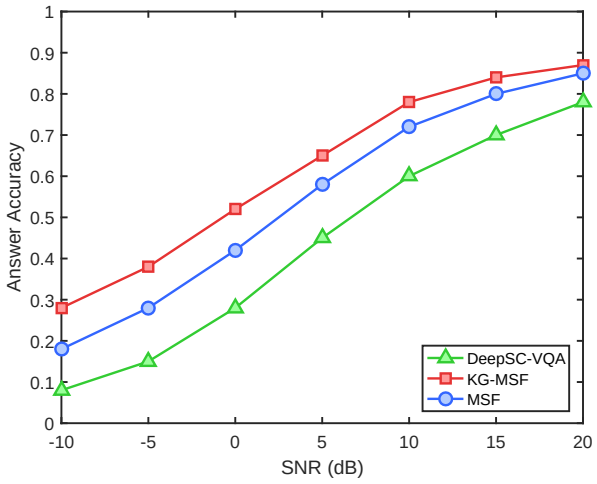

Figure 1. Answer accuracy versus SNR in AWGN channels.

Fig. 1 and Table 1 show the answer accuracy of the proposed KG-MSF versus the transmit SNR under AWGN channels, where two competing schemes of DeepSC-VQA and MSF are considered and SNR ranges from -10dB to 20dB. From this figure and table, we can find that across the entire SNR, the proposed KG-MSF scheme demonstrates consistently better performance than the two competing schemes. Specifically, at -10dB, the proposed KG-MSF scheme achieves an accuracy of 0.280, which is substantially higher than MSF's 0.180 and DeepSC-VQA's 0.080. As SNR increases, the accuracy of all schemes improves, but KG-MSF still maintains its lead. For instance, at 0dB, KG-MSF reaches 0.520, outperforming MSF at 0.420 and DeepSC-VQA at 0.280. Even at high SNRs like 15dB and 20dB, KG-MSF records 0.840 and 0.870, ahead

Table 1. Numerical answer accuracy in AWGN channels across various SNR levels.

SNR (dB)	DeepSC-VQA	KG-MSF	MSF
-10	0.080	**0.280**	0.180
-5	0.150	**0.380**	0.280
0	0.280	**0.520**	0.420
5	0.450	**0.650**	0.580
10	0.600	**0.780**	0.720
15	0.700	**0.840**	0.800
20	0.780	**0.870**	0.850

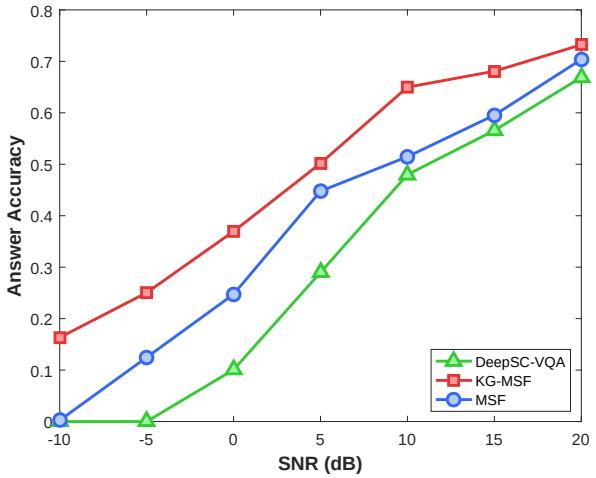

Figure 2. Answer accuracy versus SNR in Rayleigh channels.

of MSF's 0.800 and 0.850, and DeepSC-VQA's 0.700 and 0.780, respectively. These comparisons indicate that KG-MSF not only handles low-SNR conditions with greater robustness but also maintains superior semantic preservation and answer accuracy as the channel becomes cleaner. This consistent advantage across all SNR levels shows the effectiveness of KG-MSF's knowledge-guided multi-source fusion strategy in improving semantic communication performance over AWGN channels.

Fig. 2 and Table 2 present the answer accuracy versus SNR under Rayleigh fading channels, comparing the proposed KG-MSF scheme with MSF and DeepSC-VQA, where the transmit SNR varies from -10dB to 20dB. From this figure and table, we can see that the proposed KG-MSF scheme clearly outperforms the competing ones across all tested SNR levels, demonstrating strong resilience to multipath fading and channel noise. Specifically, at a very low SNR of -10dB, the proposed KG-MSF scheme achieves an accuracy of 0.1633, while MSF barely reaches 0.0025 and DeepSC-VQA performs even worse. As SNR improves, the performance gap remains significant. For example, at 0dB, the proposed KG-MSF scheme reaches 0.3696, surpassing MSF's 0.2473 and DeepSC-VQA's 0.1007. A similar advantage is evident at 6dB, where the proposed KG-MSF scheme achieves 0.3878, outperforming MSF at 0.3676 and DeepSC-VQA at 0.3047. Even at high SNRs like 15dB and 20dB, the proposed KG-MSF scheme continues to lead with accuracies of 0.6808 and 0.7323, while MSF reaches 0.5955 and 0.7037, and DeepSC-VQA trails at 0.5657 and 0.6689. These consistent gains, ranging from 5% to over 10% relative to MSF and even larger compared to DeepSC-VQA, confirm that the integration of knowledge graphs and multi-source fusion in KG-MSF enables robust semantic reasoning and enhanced cross-modal alignment, particularly under more challenging wireless conditions like Rayleigh fading.

5. Conclusion

This paper proposed an knowledge graph-enhanced multi-source fusion scheme, a new semantic communication system for robust cross-modal information transmission over wireless channels, with a focus on visual question answering applications. The proposed KG-MSF scheme extracted semantic features from both visual and textual inputs, integrated them through a bidirectional cross-attention mechanism, and further enriched the fused representation by constructing structured shallow and deep-level triplets guided by a knowledge graph. To preserve semantic integrity under noisy channel conditions, the fused features were transmitted using a joint source-channel coding strategy that balanced the compression efficiency and resilience to channel impairments. Extensive simulation results were proided to validate the superiority of the KG-MSF scheme over competing ones. For example, under Rayleigh fading conditions at 10dB SNR, the proposed KG-MSF scheme achieved an answer accuracy of 0.6498, outperforming MSF (0.5147) and DeepSC-VQA (0.4799). These findings demonstrated that KG-MSF significantly enhanced semantic robustness and cross-modal reasoning performance, delivering consistent gains across diverse wireless environments.

Table 2. Numerical answer accuracy in Rayleigh channels across various SNR levels.

SNR (dB)	DeepSC-VQA	KG-MSF	MSF
-10	0.0000	**0.1633**	0.0025
-5	0.0000	**0.2504**	0.1247
0	0.1007	**0.3696**	0.2473
5	0.2899	**0.5018**	0.4477
10	0.4799	**0.6498**	0.5147
15	0.5657	**0.6808**	0.5955
20	0.6689	**0.7323**	0.7037

5.1. Copyright

References

[1] L. Bai, X. Song, and L. Zhu, "Joint multi-feature information entity alignment for cross-lingual temporal knowledge graph with BERT," *IEEE Trans. Big Data*, vol. 11, no. 2, pp. 345–358, 2025.

[2] F. Yang, W. Chen, H. Lin, S. Wu, X. Li, Z. Li, and Y. Wang, "Task-oriented tool manipulation with robotic dexterous hands: A knowledge graph approach from fingers to functionality," *IEEE Trans. Cybern.*, vol. 55, no. 1, pp. 395–408, 2025.

[3] T. Zhang, J. Cheng, L. Miao, H. Chen, Q. Li, Q. He, J. Lyu, and L. Ma, "Multi-hop reasoning with relation based node quality evaluation for sparse medical knowledge graph," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 9, no. 2, pp. 1805–1816, 2025.

[4] H. Sun, J. Wang, J. Weng, and W. Tan, "KG-ID: knowledge graph-based intrusion detection on in-vehicle network," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 4, pp. 4988–5000, 2025.

[5] L. Liao, L. Zheng, J. Shang, X. Li, and F. Chen, "ATPF: an adaptive temporal perturbation framework for adversarial attacks on temporal knowledge graph," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 3, pp. 1091–1104, 2025.

[6] T. Song, L. Yin, Y. Liu, L. Liao, J. Luo, and Z. Xu, "Expressiveness analysis and enhancing framework for geometric knowledge graph embedding models," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 1, pp. 306–318, 2025.

[7] Z. Liu, T. Sun, X. Sun, and W. Cui, "Estimating remaining useful life of aircraft engine system via a novel graph tensor fusion network based on knowledge of physical structure and thermodynamics," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–14, 2025.

[8] S. Bi, Z. Miao, and Q. Min, "LEMON: A knowledge-enhanced, type-constrained, and grammar-guided model for question generation over knowledge graphs," *IEEE Trans. Learn. Technol.*, vol. 18, pp. 256–272, 2025.

[9] C. Mai, Y. Chang, C. Chen, and Z. Zheng, "Enhanced scalable graph neural network via knowledge distillation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 1, pp. 1258–1271, 2025.

[10] V. Hoang, V. Nguyen, R. Chang, P. Lin, R. Hwang, and T. Q. Duong, "Adversarial attacks against shared knowledge interpretation in semantic communications," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 2, pp. 1024–1040, 2025.

[11] Y. Bo, S. Shao, and M. Tao, "Deep learning-based superposition coded modulation for hierarchical semantic communications over broadcast channels," *IEEE Trans. Commun.*, vol. 73, no. 2, pp. 1186–1200, 2025.

[12] Y. Rong, G. Nan, M. Zhang, S. Chen, S. Wang, X. Zhang, N. Ma, S. Gong, Z. Yang, Q. Cui, X. Tao, and T. Q. S. Quek, "Semantic entropy can simultaneously benefit transmission efficiency and channel security of wireless semantic communications," *IEEE Trans. Inf. Forensics Secur.*, vol. 20, pp. 2067–2082, 2025.

[13] R. Cheng, Y. Sun, D. Niyato, L. Zhang, L. Zhang, and M. A. Imran, "A wireless ai-generated content (AIGC) provisioning framework empowered by semantic communication," *IEEE Trans. Mob. Comput.*, vol. 24, no. 3, pp. 2137–2150, 2025.

[14] X. Liu, H. Liang, Z. Bao, C. Dong, and X. Xu, "A semantic communication system for point cloud," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 894–910, 2025.

[15] L. Wang, W. Wu, F. Zhou, Z. Qin, and Q. Wu, "Irs-enhanced secure semantic communication networks: Cross-layer and context-awared resource allocation," *IEEE Trans. Wirel. Commun.*, vol. 24, no. 1, pp. 494–508, 2025.

[16] Z. Wan, S. Liu, Z. Xu, W. Ni, Z. Chen, and F. Wang, "Semantic communication method based on compression ratio optimization for vision tasks in the artificial intelligence of things," *IEEE Trans. Consumer Electron.*, vol. 70, no. 2, pp. 4934–4944, 2024.

[17] P. Wang, J. Li, C. Liu, X. Fan, M. Ma, and Y. Wang, "Distributed semantic communications for multimodal audio-visual parsing tasks," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 4, pp. 1707–1716, 2024.

[18] Y. Tang, N. Zhou, Q. Yu, D. Wu, C. Hou, G. Tao, and M. Chen, "Intelligent fabric enabled 6g semantic communication system for in-cabin scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1153–1162, 2023.

[19] C. Sun, D. Ming, L. Xu, S. Xie, R. Liu, and X. Ling, "A hybrid damaged building sample generation method based on cross-scale fusion generative model for destroyed building detection after earthquake," *IEEE Trans. Geosci. Remote. Sens.*, vol. 63, pp. 1–15, 2025.

[20] M. Liu, H. Liu, and T. Guo, "Cross-model cross-stream learning for self-supervised human action recognition," *IEEE Trans. Hum. Mach. Syst.*, vol. 54, no. 6, pp. 743–752,

2024.
[21] J. Wang, Y. Xie, S. Xie, and X. Chen, "Dual cross-attention transformer networks for temporal predictive modeling of industrial process," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, 2024.
[22] P. Miao, W. Su, G. Wang, X. Li, and X. Li, "Self-paced multi-grained cross-modal interaction modeling for referring expression comprehension," *IEEE Trans. Image Process.*, vol. 33, pp. 1497–1507, 2024.
[23] J. Ding, W. Li, L. Pei, M. Yang, A. Tian, and B. Yuan, "Novel pipeline integrating cross-modality and motion model for nearshore multi-object tracking in optical video surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 12 464–12 476, 2024.
[24] W. Li, H. Zhou, C. Zhang, W. Nie, X. Li, and A. Liu, "Dual-stage uncertainty modeling for unsupervised cross-domain 3d model retrieval," *IEEE Trans. Multim.*, vol. 26, pp. 8996–9007, 2024.
[25] J. Xu and L. Cao, "Copula variational LSTM for high-dimensional cross-market multivariate dependence modeling," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 11, pp. 16 233–16 247, 2024.
[26] Y. He and W. Shen, "Msit: A cross-machine fault diagnosis model for machine-level CNC spindle motors," *IEEE Trans. Reliab.*, vol. 73, no. 1, pp. 792–802, 2024.