Information Processing and Management 63 (2026) 104591

Contents lists available at ScienceDirect

Information Processing and Management

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

Knowledge-based visual question classification using quaternion hypergraph consistent network

Jing Wang [a], Duantengchuan Li [b,*], Xu Du [c], Hao Li [c], Zhuang Hu [c]

[a] School of Automation, Chongqing University of Posts and Telecommunications, 400065, Chongqing, China
[b] School of Information Management, Wuhan University, 430072, Wuhan, China
[c] Faculty of Artificial Intelligence in Education, Central China Normal University, 430079, Wuhan, China

ARTICLE INFO

Keywords:
Intelligent education
Multimodal fusion
Quaternion hypergraph
Explicit-implicit features
Knowledge consistency

ABSTRACT

Visual questions are an important means to evaluate students' knowledge. Knowledge-based visual question classification can effectively excavate the knowledge intention of the question, and realize the effective organization and management of online question resources at the knowledge level. The existing methods simply regard it as a multimodal classification task, ignoring the capture of implicit knowledge information and the fine-grained interactions between multimodal and multi-granularity features. To mitigate this, we propose a quaternion hypergraph consistent network (QHCN). This approach can not only extract explicit semantic features and implicit knowledge features from text and images simultaneously, but also considers three key properties among explicit-implicit features: modality complementation, modality independence, and knowledge consistency. Specifically, a visual question is represented as a quaternion vector consisting of two modalities and four-dimensional features. To achieve multimodal complementation, the consistency of vision and language guides the construction of a quaternion hypergraph, and a quaternion convolution operator deeply fuses explicit-implicit features. To capture interdependencies between explicit-implicit features, the independence loss and knowledge consistency loss are designed to optimize hypergraph network parameters and enhance the hypergraph structure. Extensive experiments on visual question sets verify that our QHCN achieved an accuracy of 94.82% and an F1 score of 94.76%, outperforming the optimal baseline by +1.46% and +1.53%, respectively.

1. Introduction

Visual questions, composed of question text and attached images, help to bring out the knowledge intention from multiple dimensions. They are an important form of learning resources in subject education. Recently, there are numerous question-based intelligent applications in the field of AI in Education (AIEd) (Li et al., 2024a; Liu et al., 2024b), such as question answering (Li et al., 2025a; Wang et al., 2023b), cognitive diagnosis (Li et al., 2025b), knowledge tracking (Song et al., 2024), and exam paper generation (Bi et al., 2024). These applications collect millions of questions (or exercises) as learning resources to evaluate students' knowledge acquisition and provide personalized learning services (Li et al., 2024b, 2025b). The realization of these intelligent services benefits from mining the hidden knowledge intention and labeling accurate knowledge points. Hence, knowledge-based question classification task is a key way for efficiently organizing and managing large-scale question banks for intelligent educational applications.

* Corresponding author.
E-mail addresses: wangj@cqupt.edu.cn (J. Wang), dtclee1222@whu.edu.cn (D. Li).

https://doi.org/10.1016/j.ipm.2025.104591
Received 5 August 2025; Received in revised form 27 October 2025; Accepted 26 December 2025
Available online 13 January 2026

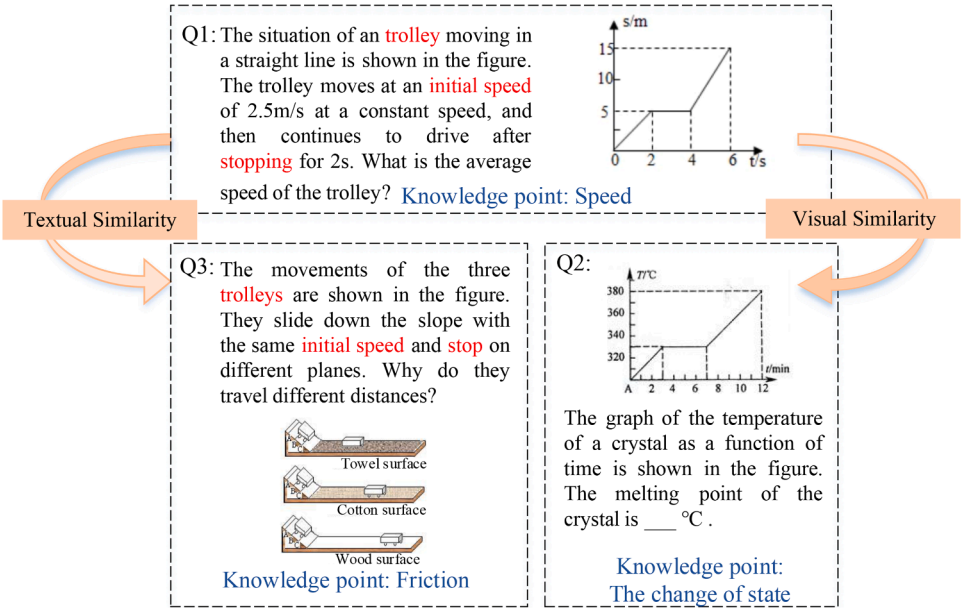


Fig. 1. Some questions with textual and visual contents. It is noticeable that multi-source information can further enrich the question representation, while the textual similarity and visual similarity arise from relying only on a single modality.

Visual questions have a better knowledge expression ability than single-modality contents, and can better help learners understand the knowledge intention of questions. For example, three examples of questions with textual and visual contents in physics as shown in Fig. 1. The explicit visual representations of questions Q_1 and Q_2 are similar, but they are annotated with "Speed" and "Temperature" knowledge points, respectively. This can be further distinguished by some keywords such as "temperature" and "speed" in the text content. Similarly, the textual content of questions Q_1 and Q_3 have semantic similarity, but can be distinguished by the attached images. In conclusion, different modalities in the same question are the embodiment of different dimensions, providing additional supplements to understand knowledge intention. Hence, how to effectively integrate the multi-source information such as textual modality and visual modality in the question is the focus of current research.

Although the existing multimodal fusion methods have achieved promising results, such as early fusion (Gadzicki et al., 2020), mid-fusion (Li et al., 2023b; Pan & Huang, 2022), and late fusion (Duan et al., 2024), these works largely focus on exploiting multimodal explicit information (*i.e.*, semantic features and visual features) (Guo et al., 2023; Li et al., 2023c), but ignore the implicit information directly related to the knowledge point from multiple modalities (*i.e.*, textual-knowledge features and visual-knowledge features). Implicit information focuses on the analysis of knowledge intentions underlying questions, complementing general semantic features. This paper refers to this as "implicit knowledge features. For instance, some keywords play a pivotal role in the question text, such as "Speed", "Temperature" and "Crystal" in Q_1 and Q_2. These keywords and their contextual information identify the corresponding knowledge scenarios and reveal knowledge intentions, which differ from the dimensions emphasized by explicit semantic features (Nguyen et al., 2021b; Udandarao et al., 2021). For the image in a question, it is easy to cause misjudgment of knowledge points only by relying on explicit visual features (Li et al., 2023a), such as the images of Q_1 and Q_2 both depict function graphs illustrating the relationship between two variables with identical trends. If only the explicit visual features of the images are extracted, the two questions could easily be misclassified as the same knowledge point. Fine-grained coordinate text and knowledge objects such as "s/m", "t/s", "$T/°C$" and "graduated cylinder" highlight the knowledge points investigated by the image, which can accurately identify questions Q_1 and Q_2 belonging to "Speed" and "Temperature", respectively.

In order to overcome the limitations of single explicit feature representation and insufficient multimodal interaction (Ji et al., 2022), this study proceeds from two aspects: on the one hand, we synchronize the explicit information (semantic features and visual features) and the implicit information (textual and visual knowledge features) in modeling the multimodal question. On the other hand, we deeply explore the multilevel interaction between the explicit features and the implicit features. In particular, we observe that there is limited redundancy between explicit semantic features and implicit knowledge features, and the two present more unique representational advantages and show significant complementarity. As shown in Fig. 2, each question is composed of two modalities and four-dimensional explicit-implicit features. Furthermore, we further dissect three properties of interactions between explicit-implicit features:

- **Modality Complementation:** The explicit-implicit features of two modalities in the same question are complementary to each other, it is necessary to integrate meaningful information from four-dimensional features, as can be seen in Fig. 2. In this way, we hope that it is possible to propagate information learned from multi-dimensional features to each other to generate a more robust question representation.

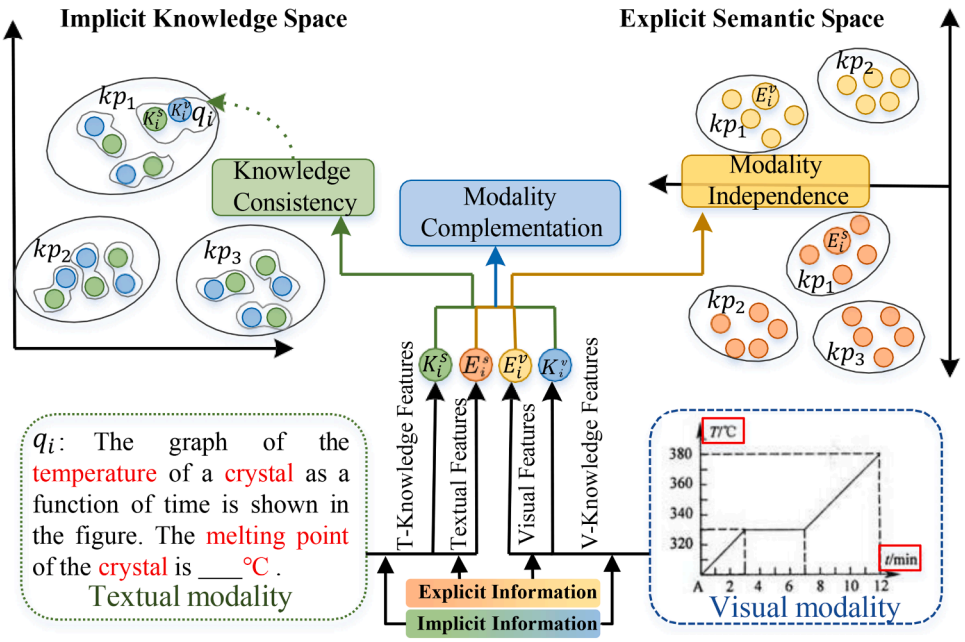


Fig. 2. Illustration of the explicit-implicit features and three key properties among these features, including modality complementation, modality independence, and knowledge consistency. Where "T-" and "V-" stand for "Textual-" and "Visual-", respectively. "kp_i" refers to the i th knowledge point, and "$K_i^s, E_i^s, K_i^v, E_i^v$" denote the four-dimensional explicit-implicit features of question q_i.

- **Modality Independence:** Each modality has its independent knowledge pointer, just like the word "temperature" is represented by a "thermometer" in the image. As the multimodal fusion proceeds, the transformed explicit feature of different modalities tend to become closer in the vector space, which impairs the unique statistical properties of different modal spaces. Hence, it is necessary to preserve the mutual independence between explicit features of different modalities while fusing multi-dimensional features.
- **Knowledge Consistency:** The implicit knowledge information in the textual modality is mainly reflected in some keywords and their contextual information, and the knowledge information in the visual modality is conveyed through fine-grained knowledge entities. Although there are different forms of knowledge representations in these two modalities, knowledge information should be consistent in terms of explaining the knowledge points of the same question. To sum up, a good feature fusion scheme should extract and integrate explicit-implicit information from multiple modalities while preserving their mutual independence and knowledge consistency.

To achieve this goal, we propose a quaternion hypergraph consistency network (QHCN), which simultaneously considers explicit-implicit features and integrates three properties in the quaternion space. This fusion scheme consists of three modules, namely, the modality complementation module, modality independence module, and knowledge consistency module. In the modality complementation module, each question is represented by a quaternion vector to capture both explicit and implicit features of text and images (Parcollet et al., 2020). Due to the differences in the attributes of explicit and implicit features, the relationships between questions exhibit high complexity. Inspired by the hypergraph structure (Bai et al., 2021), the consistency of textual hypergraph and visual hypergraph guides the construction of a quaternion hypergraph, where a hyperedge connects the questions that both textual and visual features are similar. This not only effectively captures the higher-order interactions across modalities, but also enhances the unity and discriminability of multimodal representations. At the same time, instead of the dot product, the Hamiltonian product (Hamilton, 1844) is used to explore cross-modality cross-feature interactions in quaternion space. In the modality independence module, we introduce an independence loss, which is a local regularizer of optimizing hypergraph network parameters to maintain the mutual independence between explicit features of different modalities. In the knowledge consistency module, a novel concept of knowledge consistency loss is proposed, which constrains dual-source knowledge by constantly narrowing the distance between implicit knowledge features in different modalities (Zhao et al., 2025). The learning of three properties between explicit-implicit features is synergistically implemented in our QHCN. The final output serves as a joint visual question representation to predict the corresponding knowledge point. Comprehensive experimental results conducted on the visual question dataset demonstrate that our approach outperforms other comparison methods by effectively integrating three fine-grained properties among explicit-implicit features in a multimodal fusion scheme.

In summarize, the major contributions are listed below:

- A quaternion hypergraph consistent network is proposed to integrate three fine-grained properties between explicit-implicit features in a multimodal fusion scheme. This approach maintains their mutual independence and knowledge consistency while fusing explicit-implicit features in the quaternion space.
- Independence loss, as a local regularizer, is designed to make textual features and visual features maintain modality-specific and distinct from each other. That can enrich the question representation by integrating multimodal information.
- Combined with the knowledge characteristics in education, knowledge consistency loss is proposed to constrain the textual-knowledge features and visual-knowledge features. That can optimize quaternion hypergraph network parameters and enhance the hypergraph structure.
- Compared with some competitive methods, extensive experiments verify that QHCN, considering the three properties of the interaction between the explicit-implicit features, achieves optimal classification performance on the real-world visaul question dataset.

The remaining of this article is organized as follows. Section 2 provides an overview of related work, including visual question classification, quaternion representation, and hypergraph learning. Section 3 elaborates on the QHCN model, which consists of three modules, modality complementation module, modality independence module, and knowledge consistency module. Extensive experiments and analyses are summarized in Section 4. Conclusions and suggestions for future work are discussed in Section 5.

2. Related work

2.1. Visual question classification

Visual question classification aims to predict the correct knowledge point according to the knowledge understanding of the attached image and question text. As it has been pointed out by some research (Silva et al., 2018; Sun et al., 2019; Zhao et al., 2021), multi-source data provide more plentiful information than single-modality, and its in-depth fusion can help us understand the knowledge intention of questions more comprehensively. Therefore, it has been a significant challenge for designing a robust yet accurate knowledge-based question classification method that combines the knowledge characteristic in the educational domain. Next, we introduce the multimodal classification algorithms from two aspects: multimodal feature representation and the fusion network.

The most typical multimodal feature representation applies deep learning algorithms to learn the global textual and visual features from question text and image, such as BERT (Liao et al., 2024), ResNet (Nizarudeen & Shanmughavel, 2024), and ViT (Shao et al., 2024), where each modality is represented by a one-dimensional vector. These methods only extract the global appearance features, while ignoring fine-grained representations of regions or words. To extract more discriminative visual and textual features, a region-based approach combined with an attention mechanism is applied to highlight important image regions and question words (Huang & Gong, 2024; Jin et al., 2024). For example, Huang and Gong (2024) proposes a dual-attention framework incorporating self-attention and directed attention, aiming to explore intramodal and intermodal interactions. Jin et al. (2024) proposes multi-element hypergraph networks to filter noise and highlight key information such as word position and sentence. All these models focus on learning explicit semantic representations over images and sentences and fail to simultaneously dig deep into implicit knowledge features that can reflect knowledge intention.

Another important research direction is to design a multimodal fusion module to learn the interaction between textual modality and visual modality. Fusion strategies mainly classified into three popular types according to the fusion form (Dixit & Satapathy, 2024; Duan et al., 2024), including early fusion (Gadzicki et al., 2020), mid-fusion (Pan & Huang, 2022) and late fusion (Dixit & Satapathy, 2024). Among them, early fusion and late fusion limit the improvement of classification accuracy without capturing the interaction information between multimodal features. Given the shortcomings of the above two fusion methods, intermediate fusion achieves dense multimodal interaction in the middle of the neural network, which effectively fuses clues from image and question text, which has become the current mainstream multimodal fusion method, such as ViLBERT (Lu et al., 2019), and ViLT (Kim et al., 2021). Most recently, impressive results on popular fusion networks are obtained not only considering the complementarity among multimodal features, but also preserving the independent properties of each modality (Han et al., 2021). However, these models fail to incorporate knowledge properties from the educational domain into multimodal fusion strategies and they are also lack of interpretability.

To summarize, there are two major limitations in the existing approaches, one of which is the over-reliance on explicit semantic features in text and visual modalities, while ignoring the necessity of extracting implicit knowledge features. Secondly, although they focus on the shallow complementarity between multimodal features, they fail to effectively model their deeper and more complex interaction mechanisms. It is worth emphasizing that the redundant information between explicit semantic features and implicit knowledge features is limited, which is more reflected in their unique representational advantages. Therefore, effective fusion of these two types of features is crucial for constructing a more comprehensive multimodal comprehension model.

2.2. Quaternion representation

Quaternions is firstly introduced by (Hamilton, 1844) with more advantages in handling multi-dimensional features and spatial rotations compared with real-valued numbers, which is widely applied in many applications, such as knowledge graph (Guo et al., 2025), 3D scenes (Wang et al., 2022), motion tracking (Ogri et al., 2024), speech recognition (Guizzo et al., 2023) and so on. Quaternions can be represented as the four-dimensional vector space with bases 1, i, j, and k, which consist of one real part and three

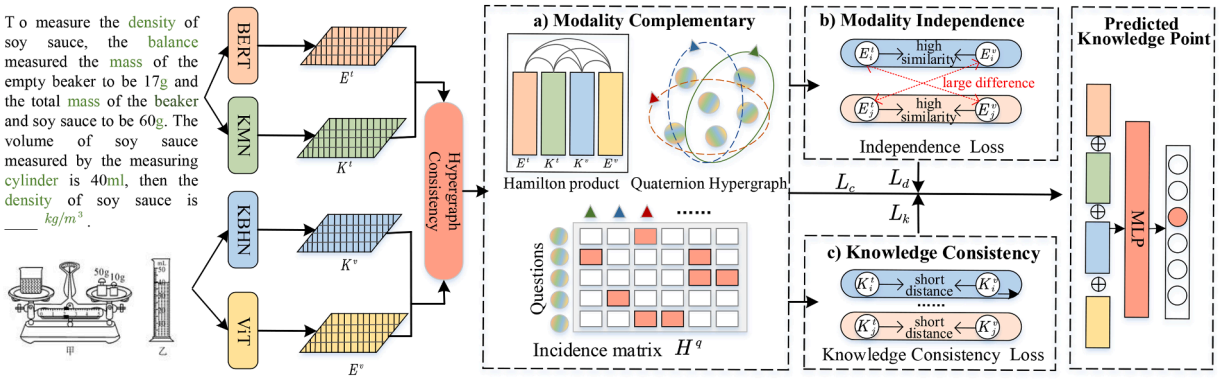


Fig. 3. The overall architecture of QHCN. The input includes the question text and attached image of each visual question. First, the explicit-implicit features extracted by respective pre-trained models for visual questions. Then, the three key properties between explicit-implicit features, such as modality complementary, modality independence, and knowledge consistency, are integrated into QHCN. Finally, the question representation integrating multimodal information is used to predict knowledge points.

imaginary parts. The specific formula is defined as:

$$\mathrm{H} = \{a + bi + cj + dk, \quad a, b, c, d \in R\}, \tag{1}$$

where a is the real part, (i, j, k) denotes the three imaginary axes, and (b, c, d) denotes the three imaginary components. Quaternion is constrained by the following formula:

$$i^2 = j^2 = k^2 = ijk = -1. \tag{2}$$

The Hamilton product can be utilized to capture the potential interaction relationships of multi-dimensional features in the quaternion space, but it lacks the noncommutative multiplication rules. Consequently, the Hamilton product $\otimes$ of two quaternion vectors $q_1 = \{a_1 + b_1 i + c_1 j + d_1 k\}$ and $q_2 = \{a_2 + b_2 i + c_2 j + d_2 k\}$ in the quaternion space is defined as:

$$\begin{aligned} q_1 \otimes q_2 = &\left(a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2\right) \\ &+ \left(a_1 b_2 + b_1 a_2 + c_1 d_2 - d_1 c_2\right)\mathrm{i} \\ &+ \left(a_1 c_2 - b_1 d_2 + c_1 a_2 + d_1 b_2\right)\mathrm{j} \\ &+ \left(a_1 d_2 + b_1 c_2 - c_1 b_2 + d_1 a_2\right)\mathrm{k}. \end{aligned} \tag{3}$$

Lately, the research of quaternion in the neural network has attracted widespread attention because of the ability of quaternion neurons to deal with multi-dimensional features. Quaternion multi-layer perceptron (QMLP) (Minemoto et al., 2017) outperforms standard MLP for document understanding. Quaternion recurrent neural networks (QRNNs) (Schwung et al., 2021) capture strong external relations and internal structural dependencies with the quaternion algebra. Compared with real-valued RNNs, it improves model performance while reducing a large number of parameters. Quaternion convolutional neural networks (Zhu et al., 2018) redesign the novel convolution layer and fully-connected layer in the quaternion space, which outperform the real-valued CNNs with the same structures. Quaternion graph neural networks (Nguyen et al., 2021a) learn graph representations within the quaternion space instead of the Euclidean space for some downstream tasks such as graph classification and node classification, thereby achieving some state-of-the-art results.

2.3. Hypergraph learning

Different from the simple graph that models pair-wise relations, a hyperedge in Hypergraph (Zhou et al., 2006) connect multiple nodes simultaneously, which can model high-order relations in complex data. Hypergraph learning is first proposed in (Zhou et al., 2006), as a label propagation method, that aims to learn vertice feature representations by minimizing the label differences among vertices that share the same hyperedge. Given this advantage, it is widely employed in many tasks, such as disease diagnosis (Luo et al., 2024; Shao et al., 2020), academic performance prediction (Li et al., 2022), and image analysis (Bloch & Bretto, 2024).

Inspired by graph-based deep learning methods, such as GCN (Wu et al., 2022) and GAT (Yao et al., 2022), HGNN (Ji et al., 2022; Liu et al., 2024a) is the first network that designs vertex convolution and hypergraph convolution to model high-order data correlation for representation learning. Considering the variability of the hypergraph structure, the dynamic hypergraph neural networks framework (DHGNN) (Jiang et al., 2019) realizes the dynamic update of the hypergraph structure based on learned vertex features, consisting of dynamic hypergraph construction (DHG) and hypergraph convolution (HGC). To highlight the importance of different hyperedges, hypergraph attention (Bai et al., 2021) is proposed to assign adaptive weights to hyperedges. Furthermore, hypergraph learning has attracted great attention and achieved prominent performance in multimodal fusion (Huang et al., 2024; Li et al., 2024c; Zhang et al., 2024). For example, Li et al. (2024c) present a dual adversarial bipartite lattice hypergraph to ensure

that figures are accurately associated with their descriptive spoken word segments. Huang et al. (2024) design a multimodal dynamic hypergraph network (MDH) to explore intra-modal and inter-modal interactions. These studies show that hypergraphs are not only capable of capturing high-order correlations in multimodal data but also deal with the heterogeneity of multimodal data, showcasing significant potential in modeling the relationships between multimodal data.

In view of this, higher-order correlations between image and text modalities between questions can be modeled with the help of hypergraphs. In addition, since the question is split into two modalities and four-dimensional features by considering the explicit and implicit features, how to improve the hypergraph structure to model the complex correlation between multidimensional features needs to be further explored.

3. The QHCN model

The overall architecture of QHCN is shown in Fig. 3. It innovatively considers both explicit-implicit information in textual and visual contents of questions, and analyzes fine-grained relationships between explicit-implicit features. The overall architecture consists of three components: modality complementation module, modality independence module, and knowledge consistency module. First, in addition to considering the explicit semantic information (*i.e.*, textual features and visual features) in each visual question, we also extract the implicit knowledge information in different modalities (*i.e.*, textual-knowledge features and visual-knowledge features). By this means, the explicit-implicit features are embedded in the quaternion space. Second, we construct a quaternion hypergraph based on the consistency of textual hypergraph and visual hypergraph. The hypergraph learns the interaction between explicit and implicit features through the Hamilton product. In addition to the mutual complementation, it also includes the mutual independence between textual features and visual features, while the knowledge consistency between textual-knowledge features and visual-knowledge features. Finally, the rich visual question representations are learned by incorporating these three properties into a quaternion hypergraph consistency network.

3.1. Explicit-implicit features representation

To guarantee a more comprehensive question representation, we innovatively consider both explicit and implicit information for visual questions. Each question consists of two modalities, including textual (t) and visual (v) modalities, and four-dimensional features, such as textual features E^t and textual-knowledge features K^t from textual modality, and visual features E^v and visual-knowledge features K^v from visual modality. Given the distinct representational advantages and significant complementary potential of both explicit semantic features and implicit knowledge features, it is imperative to employ feature extraction methods tailored to their respective distribution and structural characteristics. The implicit knowledge information is mainly reflected in some knowledge objects in the image, such as “pulley”, “balance” and “thermometer”, etc., as well as the knowledge attribute words of question text such as “constant speed”, “m/s” and the surrounding contextual information. Hence, given a question q_i with text and image, it can be represented by a quaternion vector $q_i = \left\{\mathrm{E}_i^t, \mathrm{E}_i^v, \mathrm{K}_i^v, \mathrm{K}_i^i\right\}$. Our goal is to predict the knowledge point kp_i by given multimodal input. Next, we will introduce the extraction of explicit-implicit features in detail separately.

Textual Features Representation. The question text is represented as a sequence of words $q_i^t = \left\{w_1, \ldots, w_t\right\}$, we use the pre-trained language model, BERT as the text encoder to extract the high-level textual features through stacked self-attention and MLP blocks. The input word sequence is fed into the BERT model after adding two special tokens, $\lceil CLS \rceil$ and $\lceil SEP \rceil$, marking the beginning and end of the question text. The entire network structure is fine-tuned based on the downstream tasks of question classification by adding the MLP layer and softmax layer, $\lceil CLS \rceil$ can aggregate the information of text sequence as the final textual features $\mathrm{E}^t \in \mathbb{R}^{d_t}$.

Visual Features Representation. Given the question image q_i^v, the pre-trained CNNs or Vision Transformer (ViT) (Shao et al., 2024) are fine-tuned on all question images to extract visual features. Recent work on computer vision has also shown that ViT achieves excellent results compared to CNN-like architectures (Dou et al., 2022). Hence, ViT as an example is used to describe the visual feature extraction process. Each image $q_i^v \in \mathbb{R}^{H \times W \times C}$ is reshaped into a sequence of 2D patches $v_p \in \mathbb{R}^{N \times (P^2 \times C)}$ that as an input to Transformer. (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the image patch size, P is 16 in the original paper, and N is the number of image patches that determines the input sequence length, calculated as $N = HW/P^2$. Through trainable linear projection, the image patches are mapped into a fixed-dimensional vector to generate the patch embedding, which together with positional encodings are regarded as the input to the Transformer. During the fine-tuning, we prepend a learnable embedding $\lceil class \rceil$ token to the sequence of embedded patches, which aggregates all patch information as the visual features $\mathrm{E}^v \in \mathbb{R}^{d_v}$.

Textual-Knowledge Features Representation. Different from explicit semantic features, implicit knowledge features related to the question text are highly contextualized, and their key information is often hidden in the attributes of specific keywords and their complex interactions. To accurately mine and model such implicit knowledge features, which are crucial for deeper understanding of the question, we have designed the Knowledge Mapping Network (KMN) in our preliminary work (Wang et al., 2023a). First, an independent knowledge attribute graph is constructed to generate the knowledge feature of the question text based on knowledge attribute words and contextual information. Second, a latent knowledge space is established to simulate the cognitive structure of learners. Finally, the knowledge representation is mapped onto scene basis vectors to generate textual-knowledge features $\mathrm{K}^t \in \mathbb{R}^{d_{kt}}$.

Visual-Knowledge Features Representation. The implicit knowledge features in images are mainly reflected in the fine-grained knowledge objects and coordinate texts, abbreviated as visual-knowledge features. The explicit visual features and implicit knowledge features in the image are inconsistent, that is, two images with similar visual representations do not necessarily annotate the same knowledge point. To this end, we present a knowledge-aware bi-hypergraph network (KBHN) in previous work (Li et al., 2023a). This

approach first applies the pre-trained Faster R-CNN and OCR systems to identify knowledge entities, including a set of knowledge objects and text regions. To avoid introducing noise, we select the top p region proposals with high confidence, and then their associated region features are performed in the mean-pooled operation as an initial knowledge representation. Based on typical images, that is, the collocation rules between knowledge objects and coordinate text, a knowledge hypergraph is constructed to learn the high-order correlation between images, to continuously update the visual-knowledge features $\mathrm{K}^v \in \mathbb{R}^{d_{kv}}$.

The four-dimensional feature vector, including textual features, textual-knowledge features, visual features, and visual-knowledge features, takes into account both explicit and implicit information for each modality.

3.2. Modality complementation module

In a multimodal fusion process, the modality pairs exchange respective information to "complement" the missing cues. To this end, meaningful information from multiple modalities is integrated into a better joint representation of visual questions by analyzing inter-modality interactions. First, we convert a question with two modalities and four-dimensional features into a quaternion vector, which is the hypercomplex space with bases 1, i, j, and k. In the quaternion space, the information transfer between explicit-implicit features is achieved via the Hamilton product. Then, the quaternion hypergraph is constructed based on the consistency of textual features distribution and visual features distribution, in which the vertices denote the four-dimensional explicit-implicit features of visual questions, and the hyperedges denote the consistency of explicit-implicit features of different modalities. Finally, the quaternion hypergraph convolution operator is applied to continuously fuse and update the quaternion vector by modeling the similarities between questions.

3.2.1. Question quaternion

The explicit-implicit features include textual features and knowledge features in the textual modality, and visual features and knowledge features in the visual modality. These four-dimensional features are the output features of their respective pre-trained models. A fully connected layer is utilized to uniformly transform the output feature into an h-dimension, then $\mathrm{E}^t, \mathrm{E}^v,\ \mathrm{K}^t,\ \mathrm{K}^v \in \mathbb{R}^{d_h}$ denote the updated explicit-implicit features after the process of the projection matrix, h is the dimension of the feature vector.

$$\begin{aligned} \mathrm{E}^t &= \mathrm{Linear}\left(\mathrm{E}^t; \theta_{et}\right), \qquad \mathrm{E}^v = \mathrm{Linear}\left(\mathrm{E}^v; \theta_{vv}\right). \\ \mathrm{K}^t &= \mathrm{Linear}\left(\mathrm{K}^t; \theta_{kt}\right), \quad \mathrm{K}^v = \mathrm{Linear}\left(\mathrm{K}^v; \theta_{kv}\right). \end{aligned} \tag{4}$$

The explicit-implicit features are utilized as the four components of the quaternion vector. Since the real-part coefficient of a quaternion equally influences all output components (both real and imaginary), they play a dominant and controlling role in feature fusion, enabling information to propagate throughout the entire quaternion representation. To ensure that the fusion space possesses a stable global semantic anchor, the modal feature with the most complete semantics, highest information density, and most stable distribution must be selected as the real part. Accordingly, the explicit textual semantic feature, being the most information-rich and semantically expressive carrier, is uniquely suited to serve as the real component of the quaternion (Pham et al., 2019; Tsai et al., 2019). Therefore, E^t is regarded as the real component of the quaternion q, $\mathrm{E}^v,\ \mathrm{K}^t,\ \mathrm{K}^v$ as the three imaginary components. The order of these three imaginary components will be further verified in the experimental section. The question quaternion is expressed as:

$$\mathrm{q} = \mathrm{E}^t + \mathrm{E}^v i + \mathrm{K}^v j + \mathrm{K}^t k, \tag{5}$$

where i, j, k are imaginary components, and they satisfy the quaternion rules: $i^2 = j^2 = k^2 = ijk = -1$. The interdependencies and interactions between these components are encoded naturally during training via the Hamilton product $\bigotimes$.

3.2.2. Quaternion hypergraph construction

To model the high-order connectivity information of different questions, a textual hypergraph $\mathcal{G}_t = \left\{\mathcal{V}_v, \mathcal{E}_v, \mathbf{W}_v\right\}$ and a visual hypergraph $\mathcal{G}_v = \left\{\mathcal{V}_v, \mathcal{E}_v, \mathbf{W}_v\right\}$ are constructed based on the feature similarity of textual modality and visual modality via the KNN, respectively. $\mathcal{V}_t = \left\{v_1^t, v_2^t, \ldots, v_n^t\right\}$ denotes the question set with text, and $\mathcal{V}_v = \left\{v_1^v, v_2^v, \ldots, v_n^v\right\}$ denotes the question set with image, n denotes the number of visual questions. Each vertex v_i^t and v_i^v is initialized with explicit-implicit features of textual modality (*i.e.*, textual features and textual-knowledge features) and explicit-implicit features of visual modality (*i.e.*, visual features and visual-knowledge features), respectively. $\mathcal{E}_t = \left\{e_1^t, e_2^t, \ldots, e_n^t\right\}$ and $\mathcal{E}_v = \left\{e_1^v, e_2^v, \ldots, e_n^v\right\}$ are a set of hyperedges in two hypergraphs. Taking the textual hypergraph as an example, each question v_i^t acts as the center and the K nearest neighbor questions are connected with question v_i^t by a hyperedge e_i^t by calculating the similarity of explicit-implicit features $\mathrm{E}^t \| \mathrm{K}^t$. Similarly, a hyperedge e_i^v is constructed based on the similarity of $\mathrm{E}^v \| \mathrm{K}^v$ in the visual hypergraph. W_t and W_v denote a diagonal matrix of hyperedges weights. The hypergraph incidence matrix $\mathbf{H}^t \in \mathbb{R}^{n \times n}$ and $\mathbf{H}^v \in \mathbb{R}^{n \times n}$ establish the correlation between vertices and hyperedges in the textual hypergraph and visual hypergraph, which is formulated as:

$$\begin{aligned} \mathbf{H}_{ij}^t &= \begin{cases} w_{ij}^t & v_i^t \in e_j^t \\ 0 & v_i^t \notin e_j^t \end{cases}, \\ \mathbf{H}_{ij}^v &= \begin{cases} w_{ij}^v & v_i^v \in e_j^v \\ 0 & v_i^v \notin e_j^v \end{cases}, \end{aligned} \tag{6}$$

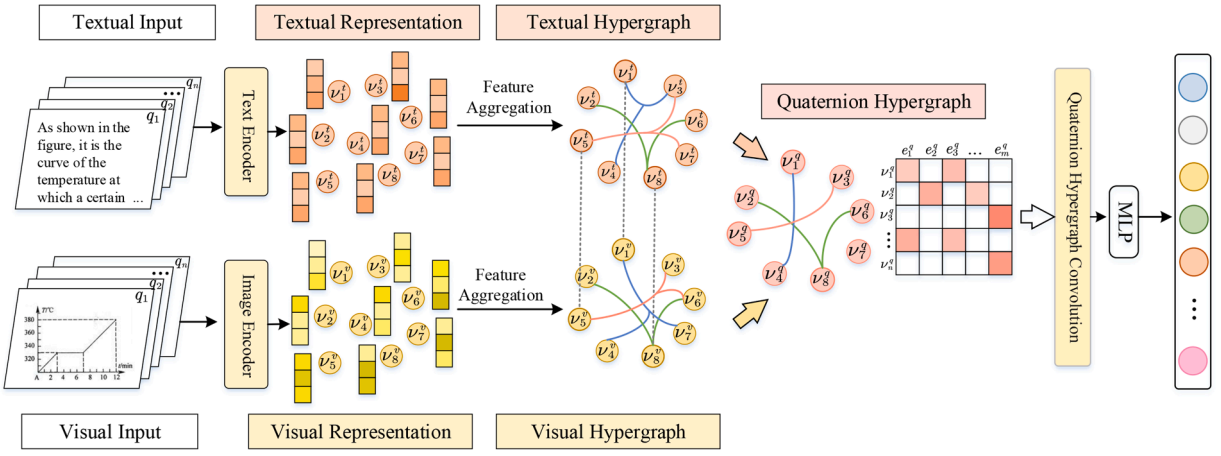


Fig. 4. Schematic of the proposed quaternion hypergraph that considers feature similarity in both textual space and visual space.

where $w_{ij}^t = \frac{\text{dist}\left(v_i^t, v_j^t\right)}{\sum_{u\in\mathcal{T}} \text{dist}\left(v_j^t, v_u^t\right)}$ and $w_{ij}^v = \frac{\text{dist}\left(v_i^v, v_j^v\right)}{\sum_{u\in\mathcal{V}} \text{dist}\left(v_j^v, v_u^v\right)}$. $dist()$ calculates the Euclidean distance between v_i^t and v_j^t. $\mathcal{T}$ is the set of K question nearest to v_j^t. $\mathcal{V}$ is the set of K question nearest to v_j^v.

If some questions are similar in both textual and visual feature space, then they have a high probability of belonging to the same knowledge point, so they can be connected by a hyperedge. To this end, a quaternion hypergraph is constructed to consider the correlations of textual and visual modalities simultaneously, denoted as $\mathcal{G}_q = \left\{\mathcal{V}_q, \mathcal{E}_q, \mathbf{W}_q\right\}$, as shown in Fig. 4. $\mathcal{V}_q$ is similar to $\mathcal{V}_t$ and $\mathcal{V}_v$, denoting a set of vertices that correspond to visual questions, except that the quaternion vector is regarded as the feature vector of each vertex. $\mathcal{E}_q = \mathcal{E}_t \cap \mathcal{E}_v$ means that the hyperedge in the quaternion hypergraph connects vertices with similar explicit-implicit features in both textual space and visual space. For example, $Edge\left(e_j^t\right) = \left\{v_i^t\right\}_{i=1:K}$ and $Edge\left(e_j^v\right) = \left\{v_i^v\right\}_{i=1:K}$ denote the vertex set contained in the hyperedge e_j^t and e_j^v, respectively. In the quaternion hypergraph, the vertex set contained in the hyperedge e_j^q is denoted as $\left\{v_i^q\right\}_{i=1:P} = \left\{v_i^t\right\}_{i=1:K} \cap \left\{v_i^v\right\}_{i=1:K}$. P is the number of vertices in $Edge\left(e_j^t\right)$, which varies with the "centroid" question, $P \le K$. Based on the above definition, the topological structure of the quaternion hypergraph can be represented by an incidence matrix $\mathbf{H}^q \in \mathbb{R}^{n\times m}$, with entries in $\mathbf{H}^q$ are defined as:

$$\mathbf{H}_{ij}^q = \begin{cases} \frac{\mathrm{W}_{ij}^t + \mathrm{W}_{ij}^v}{2} & v_i^q \in Edge\left(e_j^q\right) \\ 0 & v_i^q \notin Edge\left(e_j^q\right) \end{cases}, \tag{7}$$

where $\mathbf{H}_{ij}^q$ denotes the correlation between the vertex v_i^q and hyperedge e_j^q. To consider the self-connection of vertices, the final incidence matrix of quaternion hypergraph is defined as $\widehat{\mathbf{H}^q} = \mathbf{H}^q + I$, I is the identity matrix. Based on the incidence matrix and the hyperedge weight matrix W_q, the vertex degree matrix D_q and hyperedge degree matrix B_q is calculated as:

$$\mathbf{D}_q\left(v_i^q\right) = \sum_{j=1}^{m} \mathcal{W}_q \widehat{\mathrm{H}_{ij}^q}, \tag{8}$$

$$\mathbf{B}_q\left(e_j^q\right) = \sum_{i=1}^{n} \widehat{\mathrm{H}_{ij}^q}, \tag{9}$$

where $\mathbf{D}^q \in \mathbb{R}^{n\times n}$, $\mathbf{B}^q \in \mathbb{R}^{m\times m}$ and $\mathbf{W}^q \in \mathbb{R}^{m\times m}$ are the diagonal matrices. $\mathbf{W}_q = I$ means that each hyperedge is assigned the same weight. n and m denote the number of vertices and hyperedges, respectively. Here, $n = m$, which is due to each question as a central vertex constructing a hyperedge, connecting neighbor nodes that are both textually and visually similar.

3.2.3. *Quaternion hypergraph convolutional layer*

The quaternion hypergraph convolution operator (Guo et al., 2023) is applied to capture the high-level interaction among visual questions. Different from traditional hypergraph neural networks, the quaternion vector $\mathrm{X}^q \in H^{n\times 4\times d_h}$ regard as the input of the convolutional layer, and quaternion filter matrix $\theta^q = \theta^t + \theta^v i + \theta^{kv} j + \theta^{kt} k$ is used to continuously update the feature vector of visual questions in the quaternion space. The quaternion hypergraph convolutional layer is defined as follows:

$$\begin{aligned} \mathbf{X}^{q(l+1)} &= QHConv\left(\mathbf{X}^{q(l)}, \widehat{\mathbf{H}^q}, \theta^{q(l)}\right) \\ &= \sigma\left(\mathbf{D}_q^{-\frac{1}{2}} \widehat{\mathbf{H}^q} \mathbf{W}_q \mathbf{B}_q^{-1} \widehat{\mathbf{H}^q}^T \mathbf{D}_q^{-\frac{1}{2}} \mathbf{X}^{q(l)} \otimes \theta^{q(l)}\right), \end{aligned} \tag{10}$$

where $\sigma(\cdot)$ is the nonlinear activation function (such as ReLU and tanh). $\mathbf{X}^{q(l)}$ is the quaternion vector of visual questions at the l layer, $\theta^{q(l)}$ is the quaternion weight matrix at the l layer. $\otimes$ denote the Hamilton product, which instead of the dot product is utilized to calculate the multiplication of two quaternions $\mathbf{X}^{q(l)}$ and $\theta^{q(l)}$. $\mathbf{X}^{q(l+1)}$ is the updated quaternion vector at the $l+1$ layer.

The Hamilton product (Gaudet & Maida, 2018) can realize the interaction between explicit-implicit features in the hypercomplex vector space by calculating the multiplication of question quaternion $X^q = \left\langle E^t, E^v, K^v, K^t \right\rangle$ and weight quaternion $\theta^q = \left\langle \theta^t, \theta^v, \theta^{kv}, \theta^{kt} \right\rangle$. Note that the quaternion weight matrix consists of four components, θ^t, θ^v, θ^{kv}, and $\theta^{kt} \in \mathbb{R}^{n \times \theta_h}$, which are jointly shared across the four-dimensional feature vector of the question quaternion, E^t, E^v, K^v, and $K^t \in \mathbb{R}^{n \times d_h}$. Here, "shared" does not mean that all components use identical parameters. Instead, the quaternion formulation maintains a set of four interdependent weight components that represent the real and imaginary parts of a unified quaternion transformation. This allows each feature dimension (explicit/implicit, textual/visual) to interact through rotational and cross-channel coupling operations, enabling individualized feature transformations within a consistent quaternion space. Such a design achieves expressive cross-modal representation while substantially reducing redundant parameters compared with independent Euclidean-space weights (Parcollet et al., 2020).

$$\begin{aligned} \mathrm{X}^q \otimes \theta^q = & \left(\mathrm{E}^t\theta^t - \mathrm{E}^v\theta^v - \mathrm{K}^v\theta^{kv} - \mathrm{K}^t\theta^{kt}\right) \\ & + \left(\mathrm{E}^t\theta^v + \mathrm{E}^v\theta^t + \mathrm{K}^v\theta^{kt} - \mathrm{K}^t\theta^{kv}\right)\mathrm{i} \\ & + \left(\mathrm{E}^t\theta^{kv} - \mathrm{E}^v\theta^{kt} + \mathrm{K}^v\theta^t + \mathrm{K}^t\theta^v\right)\mathrm{j} \\ & + \left(\mathrm{E}^t\theta^{kt} + \mathrm{E}^v\theta^{kv} - \mathrm{K}^v\theta^v + \mathrm{K}^t\theta^t\right)\mathrm{k}. \end{aligned} \tag{11}$$

With the Hamilton product and hypergraph convolution, the quaternion hypergraph convolution not only considers the high-order correlation between visual questions but also integrates the interaction between explicit-implicit features of each question. The entire convolution process consists of the following three stages: fusion within explicit-implicit features, vertex convolution, and hyperedge convolution. Firstly, the interaction between explicit-implicit features is integrated into the visual question representations through the Hamilton product in the quaternion space. Secondly, vertex features with visual similarity and textual similarity is aggregated to form hyperedge features by the incidence matrix $\widehat{\mathrm{H}^q}^T \in \mathbb{R}^{m \times n}$, defined as "vertex convolution". Thirdly, the vertex features are updated by pre-multiplying $\widehat{\mathrm{H}^q}$ to aggregate information from multiple hyperedges. Overall, quaternion hypergraph convolutional network can model high-order correlations between questions with both visual similarity and textual similarity, and achieve explicit-implicit features fusion.

3.3. Modality independence module

Each modality has its own feature embedding spaces that maintain the unique properties of their respective modality. However, as fusion proceeds, multimodal features tend to become closer in the hidden space through the learning of multiple network layers, which impairs the independent properties of different modalities, called feature space collapse (Han et al., 2021). Hence, maintaining modality-specific information while considering the interactions between multimodal features is the key to multimodal fusion. To this end, we introduce the independence loss similar to the one in Nie et al. (2020), as a local regularizer that can maximize the separability of the unimodal representations to preserve the mutual independence of each modality in the process of multimodal fusion. Specifically, semantic features E^t and visual features E^v are the explicit semantic features of textual modality and visual modality, which contain rich modality-specific information. The independent loss function is designed to effectively preserve the independence of E^t and E^v by maximizing the similarity between E_i^t and E_i^v from the same question q_i, as well as the dissimilarity between E_i^t and E_j^v from different questions.

In practice, the independence loss is operated on a dataset of n questions (instances) $\mathrm{Q} = \left\{q_i\right\}_{i=1}^n$ where each question q_i is the question text q_i^t with an attached image q_i^v. First, the l_2 norm is employed to normalize the semantic features $E^t = \left\{E_1^t, E_2^t, \ldots, E_n^t\right\}$ and visual features $E^v = \left\{E_1^v, E_2^v, \ldots, E_n^v\right\}$, denoted as $\xi\left(E_i^t\right)$ and $\xi\left(E_i^v\right)$. Then, similar to Wang et al. (2019), we define the similarity of the two modalities as $sim = \left\langle \xi\left(E_i^t\right), \xi\left(E_j^v\right) \right\rangle$, where $\langle , \rangle$ denotes dot product, resulting in an similarity matrix $\mathbf{S}$ which represent the similarity of E^t and E^v among different questions, respectively. Finally, an independence loss is calculated to preserve the independence of each modality, the central idea is to encourage modality pairs in the same question to be closer, and push modality pairs of different questions apart. Our independence loss is formulated as:

$$L_d = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{e^{S_{ii}/\tau}}{\sum_{i \neq j} \left(e^{S_{ij}/\tau} + e^{S_{ji}/\tau}\right)}, \tag{12}$$

where S_{ii} denotes the similarity of $\xi\left(E_i^t\right)$ and $\xi\left(E_i^v\right)$ in the same question q_i, S_{ij} is the similarity between $\xi\left(E_i^t\right)$ and $\xi\left(E_j^v\right)$, and S_{ji} refers to the similarity between $\xi\left(E_j^t\right)$ and $\xi\left(E_i^v\right)$. Independence loss expects to maximize $e^{S_{ii}}$ and minimize $e^{S_{ij}}$ and $e^{S_{ji}}$. That is, we need to make $\xi\left(E_i^t\right)$ and $\xi\left(E_i^v\right)$ as consistent as possible, and $\xi\left(E_i^t\right)$ and $\xi\left(E_j^v\right)$ as inconsistent.

However, we observe an important limitation in our proposed independence loss according to Morgado et al. (2021). Specifically, some questions are semantically related to q_i, which makes $e^{S_{ij}}$ and $e^{S_{ji}}$ have large values. This contradicts the goal of independence loss, which make semantic features more similar to visual features of the same question as opposed to visual features of other questions and vice versa. To alleviate this issue, we define a semantic relevance score srs_{ij} for q_i and q_j. If two questions are similar in both textual and visual feature space, then they are more likely to be semantically related, and these questions are not included in the

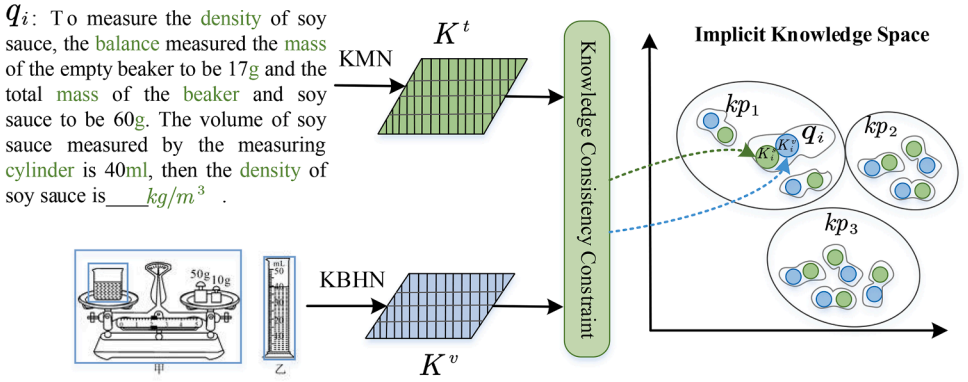


Fig. 5. Illustration of the knowledge consistency between implicit knowledge features of different modalities in the same question.

formulation as other questions. The semantic relevance score is defined as:

$$srs_{ij} = \min\left(\xi\left(E_i^t\right)^T \xi\left(E_j^t\right), \xi\left(E_i^v\right)^T \xi\left(E_j^v\right)\right). \tag{13}$$

A larger srs_{ij} indicates that question q_i and q_j are similar in both semantic space and visual space. $\mathcal{F}_i = \text{TopN}\left(srs_{ij}\right)$ denotes that there are N questions with high semantic relevance to question q_i, we need to remove these N questions from the dataset before calculating the independence loss. The independence loss is updated as:

$$L_d = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{e^{S_{ii}/\tau}}{\sum_{i\neq j}^{Q\setminus\mathcal{F}_i} \left(e^{S_{ij/\tau}} + e^{S_{ji/\tau}}\right)}, \tag{14}$$

where $Q\setminus\mathcal{F}_i$ indicates that the question set with high semantic relevance is removed from the full training dataset. τ is a temperature hyper-parameter that controls the strength of penalties. The independence loss guides the training process by maximizing the separation of semantic features and visual features.

3.4. Knowledge consistency module

Although different modalities have their individual modality-specific information, they are the descriptions of one knowledge point with the same knowledge intention. As shown in Fig. 5, the knowledge objects "graduated cylinder" and "balance" in the images and knowledge attribute words "density", "balance" and "kg/m^3", etc, are both the understanding of the knowledge point of "Density", but the form of expression is different. Hence, the implicit knowledge features K_i^t and K_i^v of the same question q_i should be consistent in terms of depicting the same knowledge point kq_i. In other words, K_i^t and K_i^v are close in the knowledge space. This property enriches question representations by optimizing the knowledge similarity between implicit information of different modalities.

To capture the knowledge consistency between K^t and K^v, we propose a knowledge consistency loss to measure the correlation of implicit knowledge features by Euclidean distances in implicit knowledge space. The consistency loss, as a local regularizer, aims to learn a cross-modal similarity metric and imposes constraints on the training process by minimizing the distance of K_i^t and K_i^v in the same question q_i. First, l_2 norm is employed to normalize the textual-knowledge features and visual-knowledge features, denoted as $\xi\left(K_i^t\right)$ and $\xi\left(K_i^v\right)$. Then, we leverage Euclidean distances to measure the distances between $\xi\left(K_i^t\right)$ and $\xi\left(K_i^v\right)$, denoted as $dist\left(\mathrm{K}_i^t,\ \mathrm{K}_i^v\right)$. The distance is calculated as follows:

$$dist\left(\mathrm{K}_i^t,\ \mathrm{K}_i^v\right) = \sqrt{\sum_{k=1}^{h} \left(K_{i,k}^t - K_{i,k}^v\right)^2}, \tag{15}$$

where $K_i^v = \left(\mathrm{K}_{i,1}^v,\ \mathrm{K}_{i,1}^v, \ldots, \mathrm{K}_{i,h}^v\right)$, $K_i^t = \left(\mathrm{K}_{i,1}^t,\ \mathrm{K}_{i,1}^t, \ldots, \mathrm{K}_{i,h}^t\right)$ are visual-knowledge features and textual-knowledge features of the question q_i respectively. Based on the relative distance of knowledge feature vectors in different modalities, we design a knowledge consistency loss, which makes K^t and K^v in the same question more and more similar. The formulation of the knowledge consistency loss function is denoted as follows:

$$L_k = \sum_{i}^{n} \log dist\left(\mathrm{K}_i^t,\ \mathrm{K}_i^v\right). \tag{16}$$

Knowledge consistency loss, as a regularization term, aims to minimize Eq. (16), that is, to minimize Eq. (15), which requires minimizing $dist\left(\mathrm{K}_i^t,\ \mathrm{K}_i^v\right)$ to supervise the learning process. That is to say, the textual-knowledge features and visual-knowledge

features of each question are as similar as possible. In this way, the knowledge consistency loss is applied to optimize the parameters of the quaternion hypergraph neural network by minimizing the distance of knowledge features of different modalities, to realize the rapid convergence of the network.

3.5. Training objectives

In the training step, all the parameters of QHCN are learned and optimized by the above three modules simultaneously. We summarize the overall procedure of QHCN in Algorithm 1. First, in each iteration, explicit-implicit features, including textual features, textual-knowledge features, visual features, and visual-knowledge features, are updated in the modality complementation module through the Hamiltonian product and quaternion hypergraph convolution operation. Then, the updated explicit-implicit features are applied to calculate the independence loss and knowledge consistency loss, which can further optimize the parameters of QHCN. So far, parameter learning has been completed during the backpropagation process. Finally, after multiple training iterations, explicit-implicit features with rich information are generated for each question. The joint feature vector after concatenating these explicit-implicit features is fed to multiple fully-connected (FC) layers to form the final representation of visual questions $X^{q\prime}$. The function is formulated as:

$$X^{q'} = QHCN\left(E_s, E_v, K_s, K_s; \theta^q, \theta^l\right), \quad X^{q'} \in \mathbb{R}^{d_c}, \tag{17}$$

where θ^q and θ^l are the parameters of the convolutional layers and fully-connected layers, respectively, d_c is the dimension of question representation. The final representation $X_q^f \in \mathbb{R}^{d_c}$ is fed into the softmax layer to obtain the confidence score for each knowledge point. Cross-entropy loss, a commonly used loss function in multi-classification tasks, was used to train our model. The specific definitions are as follows:

$$\hat{y}_i = \mathrm{softmax}\left(W^f X^{q'} + b^f\right), \tag{18}$$

$$L_c = -\sum_{q \in Q} \sum_{i=1}^{K} y_i \log \hat{y}_i, \tag{19}$$

where K refers to the number of knowledge points, Q represents the training set. $\hat{y}_i$ is the predicted label distribution and y_i is the ground-truth label distribution.

In addition to the cross-entropy loss, independence loss learns from the difference of explicit semantic features, but knowledge consistency loss learns from the similarity of implicit knowledge features. Therefore, the loss function of this work consists of three parts: cross-entropy loss, independence loss, and knowledge consistency loss. Among them, the Cross-Entropy loss is regarded as the primary loss function to facilitate the training for the classification task. The independence loss and knowledge consistency loss are used as auxiliary losses to further optimize the parameters and speed up the training process. The total loss is the weighted sum of the three losses:

$$L = L_c + \lambda_1 L_d + \lambda_2 L_k, \quad \lambda_1 + \lambda_2 = 1, \tag{20}$$

where L_c, L_d and L_k represents the cross-entropy loss, independence loss, and knowledge consistency loss, respectively. λ_1 and λ_2 are the corresponding loss coefficient of independence loss and knowledge consistency loss, respectively. The parameters of the entire network are automatically learned by minimizing the total loss over the training set.

4. Experiments

4.1. Visual question sets

Dataset Description. The data used in this study is a real-world dataset of junior high school grades 7 to 8 excavated from publicly available teaching resources in physics disciplines, such as textbooks, supplementary books, quiz questions, etc. It is obtained by regrouping and organizing on the basis of two previously published papers (Li et al., 2023a; Wang et al., 2023a). The dataset contains 16,077 questions of 46 knowledge points, such as liquid pressure, levers, plan mirror imaging, friction, the reflection of light, and sound, etc. Each question is composed of the title text and the attached image, and marks the knowledge points under investigation. For the question text, after preprocessing and padding operations, the number of words in each question is fixed at 150, and the number of keywords is 25. For the images, each image is in JPG format, rescaled to 224 × 224 pixels due to the inconsistent size of the images. The dataset is divided into train and test sets with roughly an 8 : 2 ratio.

Annotation Process. To guarantee annotation quality and consistency, a rigorous two-stage annotation protocol was implemented by three domain experts in physics. In the initial phase, each annotator independently labeled all 16,077 questions. Subsequently, questions with inter-annotator disagreements underwent consensus discussions among all experts to establish definitive knowledge point classifications. This systematic approach ensures the creation of a reliable, generalizable physics visual question dataset for robust training and validation of QHCN.

Specific Examples. To quickly understand the visual question dataset, we randomly select question examples from five knowledge points, as shown in Fig. 6. The left column shows the knowledge point for the same row of questions, each question includes the question text and an image. It can be seen from the representative questions that the text and image of each question have common knowledge intention, and are different manifestations of the same knowledge point. Moreover, visual questions contain richer

Algorithm 1 Quaternion hypergraph consistency network.

Input: The visual question set: $Q = \left\{\left(q_i^t, q_i^v\right)\right\}_{i=1}^{n}$
Output: The knowledge point for each question kp_i

Step 1. Pretraining
Textual features $\mathrm{E}_i^t = \mathrm{Bert}\left(w_{i,1}, w_{i,2}, \ldots, w_{i,t}\right)$
Textual-knowledge features $\mathrm{K}_i^t = KMN\left(a_{i,1}, a_{i,2}, \ldots, a_{i,kt}\right)$
Visual features $\mathrm{E}_i^v = ViT\left(i_{i,1}, i_{i,2}, \ldots, i_{i,v}\right)$
Visual-knowledge features $\mathrm{K}_i^v = KBHN\left(o_{i,1}, o_{i,2}, \ldots, o_{i,kv}\right)$

Step 2. Quaternion Hypergraph Construction
Textual hypergraph $H^t = \mathrm{KNN}\left(E^t \| K^t\right)$
Visual hypergraph $H^v = \mathrm{KNN}(E^v \| K^v)$
Quaternion hypergraph $H^k = H^t \cap H^v$
Vertex quaternion $X^q = \left\langle E^t, E^v, K^v, K^t \right\rangle$, Weight quaternion $\theta^q = \left\langle \theta^t, \theta^v, \theta^{kv}, \theta^{kt} \right\rangle$
$\mathrm{X}^{q(l+1)} \leftarrow QH\,\mathrm{Conv}\left(\mathrm{X}^{q(l)}, \widehat{\mathrm{H}}^q, \theta^{q(l)}\right)$ using Eq.(10)

Step 3. Joint Training
$\mathrm{X}^{q(l+1)} = \left\langle E^{t(l+1)}, K^{t(l+1)}, E^{v(l+1)}, K^{v(l+1)} \right\rangle$
$\widehat{\mathrm{H}^t} \leftarrow \mathrm{X}^{q(l+1)}$
$L_c \leftarrow \mathrm{X}^{q(l+1)}$
$L_d \leftarrow \mathrm{E}^{t(l+1)}, \mathrm{E}^{v(l+1)}$
$L_k \leftarrow \mathrm{K}^{t(l+1)}, \mathrm{K}^{v(l+1)}$
$L \leftarrow L_c, L_k, \mathrm{L}_d, \lambda_1, \lambda_2$

information than single-modality contents. For example, q_1 and q_3 have textual similarity, and q_2 and q_5 have visual similarity, if only relying on a single modality, it is easy to cause misjudgment. Therefore, different modalities in the same question are complementary to each other, it is necessary to integrate meaningful information from multiple modalities and consider intermodality interactions.

4.2. Experiment settings

4.2.1. Implementation details

For word segmentation of the question text, PkuSeg based on a physical lexicon is used to preserve the integrity of subject-specific nouns. For example, “uniform rectilinear motion”, “mirror reflection”, “standard atmospheric pressure”, etc., these subject-specific nouns serve as a whole words rather than multiple words. The question text is embedded in a 768-dimensional latent semantic space by a pre-trained BERT. In the image processing, fine-tune ResNet-50 or ViT to adapt the question classification task. A series of input patches with 16×16 is regarded as the input of ViT, and the output *[class]* token is the visual features of the image. The four-dimensional features of visual questions, including visual features, textual features, visual-knowledge features, and textual-knowledge features, are reduced to 1024-dimension by the respective MLP layers and constitute the quaternion vector of each question. Other parameters are as follows: $d_t = 768$, $d_v = 768$, $d_{kt} = 1024$, $d_{kv} = 2048$, $d_h = 1024$, $d_c = 256$.

The architecture of the quaternion hypergraph consistency network is two HGCN layers, two FC layers, and a softmax classification layer. The whole model is trained with the Adam optimizer for 500 epochs with a learning rate of 1e-4 and weight decay of 1e-5. The probability of dropout (Srivastava et al., 2014) is set to 0.5 to avoid overfitting during training. Following Wu et al. (2018), we set the temperature hyper-parameter τ to 0.07 in independence loss. λ_1 and λ_2 are hyperparameters that control the relative weights of independence loss and knowledge consistency loss, and their values are discussed in the experimental section. The experimental environment for implementing QHCN is an Nvidia GeForce GTX 2080 Ti GPU with 11G memory. Two evaluation metrics, Accuracy and macro-F1, are applied to verify the effectiveness of QHCN.

4.2.2. Baselines

To validate the effectiveness of QHCN, we conduct a comparative analysis with the following baseline models on our visual question set. First, we evaluate mainstream unimodal models, including Att-BLSTM and BERT for the textual modality, and ResNet-50 and ViT for the visual modality. Second, we consider a comparison with several state-of-the-art multimodal fusion models reported in the literature, such as ViLBERT, $Pythia_G$, and ViLT. Third, we propose two model variants (*i.e.* $QHCN_{RL}$ and $QHCN_{VB}$) to validate the superiority of our method, $QHCN_{RL}$ is based on ResNet-50 and Att-BLSTM, $QHCN_{VB}$ uses ViT and BERT as text encoder and image encoder, respectively. To ensure the reliability of the experimental results, all models are reimplemented on our visual question dataset.

Unimodal models:

- Att-BLSTM (Yao et al., 2022). The network architecture combines BLSTM and neural attention mechanism, which can automatically focus on important words in the question text while considering the word order of the question text.

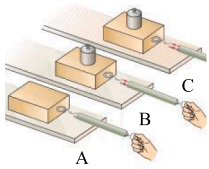
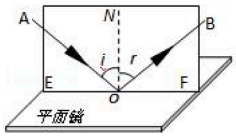
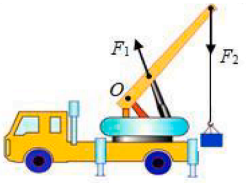
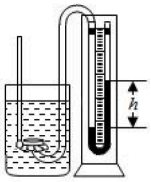
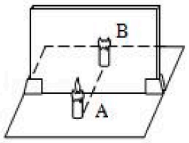
Knowledge point	Text Modality	Image Modality
Friction	q_1: In the experiment to explore the influencing factors of sliding friction, the dynamometer pulls the block to move on the board. The board in C is rougher than the board in A and B. Please draw the force diagram of the weight in the figure.	
Reflection of Light	q_2: ENF is a plane connected by two pieces of cardboard and placed perpendicular to the plane mirror. To explore the relationship between the angle of reflection and the angle of incidence, the operation that should be performed during the experiment is ____.	
Levers	q_3: As shown in the figure, this is a schematic diagram of a hydraulic truck crane. The force of the hydraulic rod on the boom is at point A and the direction can be adjusted. Draw a schematic diagram of the force acting on the boom by the hydraulic rod at this time.	
Liquid Pressure	q_4: As shown in the figure, the red ink injected into the U-shaped tube, if the density of the red ink is close to that of water, and h=15cm, the pressure at the rubber film is about _____Pa. (g=10N/kg)	
Plane Mirror Imaging	q_5: To explore the "relationship between the image and the object in the plane mirror imaging", a lit candle A is placed in front of the glass plate, and an unlit candle B moves behind the glass plate until the candle B coincides with the image of the candle A, indicating the size of the image with the object _____.	

Fig. 6. Some question examples are composed of question text and attached images.

- BERT (Liao et al., 2024). Bidirectional Encoder Representations from Transformers is a pretrained language representation model. It applies a new masked language model (MLM) to generate deep bidirectional language representations. We fine-tuned the Bert on all question texts.
- S-KMN (Wang et al., 2023a). This model employs dual-channel joint modeling of semantics and knowledge to enrich the representation of questions from both explicit semantic and implicit knowledge perspectives. This is our previous work, which focused solely on knowledge annotation for test question texts.
- ResNet (Nizarudeen & Shanmughavel, 2024). It addresses two challenges of deep network structures. First of all, it effectively solves the degradation problem through residuals blocks with short connections. Second, the gradient disappearance and gradient explosion problems are solved by adding BN layers.
- ViT (Shao et al., 2024). Vision transformer as a transformer architecture similar to Bert is a pre-trained model. It splits the image into a sequence of image patches to learn image feature representation. Due to the small image size used in the education domain, vitb/16 is chosen as a baseline model.
- KBHN (Li et al., 2023a). It not only extracts coarse-grained visual features by constructing a visual hypergraph, but also establishes a knowledge hypergraph based on typical images to aggregate images sharing similar knowledge information. This also aligns with our previous research work.

Table 1
The performance comparison of QHCN and all baselines on the test set.

Type	Method	Accuracy	Macro-F1
Unimodal models	Textual(Att-BLSTM) (Yao et al., 2022)	0.8892	0.8879
	Textual(Bert) (Liao et al., 2024)	0.9032	0.9022
	Textual(S-KMN) (Wang et al., 2023a)	0.9105	0.9097
	Visual(ResNet50) (Nizarudeen & Shanmughavel, 2024)	0.8853	0.8832
	Visual(ViT) (Shao et al., 2024)	0.8945	0.8903
	Visual(KBHN) (Li et al., 2023a)	0.9023	0.9011
VLP models	$HGCN_{early}$	0.9276	0.9272
	$HGCN_{late}$	0.9284	0.9279
	ViLBERT (Lu et al., 2019)	0.9308	0.9285
	$Pythia_G$ (Jiang et al., 2020)	0.9315	0.9294
	ViLT (Kim et al., 2021)	0.9336	0.9323
Multimodal Affective Analysis	DMLANet (Yadav & Vishwakarma, 2023)	0.9235	0.9217
	MDPF (Lu et al., 2025)	0.9431	0.9428
Multimodal Fake News Detection	CSFND (Peng et al., 2024)	0.9318	0.9320
	KEVL (Gao et al., 2024)	0.9362	0.9371
Visual Question Answering	VG-CALF(Lameesa et al., 2025)	0.9289	0.9285
	CMQEF (Li et al., 2024d)	0.9322	0.9324
QHCN	$QHCN_{RL}$	0.9456	0.9451
	$QHCN_{VB}$	**0.9482**	**0.9476**

Multimodal fusion models:

- $HGCN_{early}$ and $HGCN_{late}$. As for $HGCN_{early}$, textual features and visual features extracted by ViT and BERT are concatenated, and the hypergraph is constructed by KNN. The merged features are directly fed into HGCN. As for $HGCN_{late}$, textual features and visual features are fed into different HGCNs respectively and concatenated afterward.
- ViLBERT (Lu et al., 2019). This model extends the popular BERT architecture to learn the joint visual-linguistic representations for vision-and-language tasks. It transfers the pre-trained BERT and ViT to process textual and visual inputs separately and achieves the interaction of the two modalities through novel co-attentional transformer layers.
- $Pythia_G$ (Jiang et al., 2020). Grid-based convolutional features, instead of region-based features from bottom-up attention, are fused with textual representations in the co-attention model (Yu et al., 2018) implemented in Pythia (Singh et al., 2019) for VQA tasks. It remarkably speeds up training while maintaining on-par results in terms of accuracy.
- ViLT (Kim et al., 2021). To address a heavy visual embedder, Vision-and-Language Transformer is proposed to apply ViT (Shao et al., 2024) of grid-based or region-based methods to extract visual features. The processing of visual inputs and textual inputs is drastically simplified in a non-convolutional manner with transformers. It is up to tens of times faster without hurting performance on downstream tasks.
- $QHCN_{RL}$. The authors proposed the quaternion hypergraph consistency network to iteratively model three key properties between explicit-implicit features. This model applies ResNet-50 and BiLSTM as visual embedders and textual embedders, respectively.
- $QHCN_{VB}$. That applies ViT and BERT as visual embedder and textual embedder in our QHCN, respectively. It achieves the state-of-the-art performance on visual question classification tasks.

4.3. Results

4.3.1. Overall comparison

Comparison with single-modal and multi-modal approaches. To verify the effectiveness of QHCN in multimodal fusion, we conduct the experiments with our method and all the baseline models on our physics question dataset under various modality settings. The experiment results are depicted in Table 1, where $QHCN_{RL}$ and $QHCN_{VB}$ denotes two variants of our model. From these results, we can make the following observations.

For all baseline models, including general-domain unimodal architectures (BERT and ViT) and education-specific ones (S-KMN and KBHN), textual features extracted by BERT consistently outperform visual features from ViT in the single-modality setting. It indicates that textual modality carries richer information. The education-oriented models S-KMN and KBHN, which incorporate implicit knowledge from question texts and instructional images, further improve annotation accuracy over BERT and ViT, underscoring the advantage of domain-aware modeling. Moreover, multimodal fusion models such as ViLT achieve the best performance, surpassing ViT and BERT by 4.20% and 3.01% in macro-F1, respectively, demonstrating that integrating complementary textual and visual cues enables a more comprehensive understanding of educational content.

As shown in Table 1, our QHCNs outperform all the baseline models over all metrics on our dataset, whether in the unimodal models or multimodal fusion frameworks. Compared to ViLT, $QHCN_{VB}$ attains an improvement of around 1.46% and 1.53% in the accuracy and macro-F1. The major reason for this is that the existing multimodal fusion methods account only for the interaction of different modalities, while our QHCNs also maintain the independence and knowledge consistency between different modalities. What's more,

$QHCN_{VB}$ achieves some slight performance gain over $QHCN_{RL}$ due to improvements in the textual and visual embedders. This suggests that adequate characterization of different explicit semantic features helps improve the model performance.

Comparison with cross-task baseline models. To comprehensively evaluate the performance of QHCN, we select two state-of-the-art models in three common tasks, namely DMLANet and MDPF for multimodal affective analysis, CSFND and KEVL for fake news detection, as well as VG-CALF and CMQEF for visual question answering. These baseline models represent the latest advancements in multimodal tasks and adopt diverse strategies for feature extraction and interaction modeling. From the perspective of feature extraction, most baseline models (e.g., MDPF, KEVL, CSFND, and VG-CALF) rely on pretrained models such as BERT and ViT to extract textual and visual features, respectively. This strategy leverages the powerful representation capabilities of large-scale pretrained models but simultaneously increases model complexity and computational costs. In contrast, DMLANet and CMQEF adopt more traditional feature extraction methods (e.g., LSTM, Glove, and Faster R-CNN), which reduce model complexity at the cost of lower classification accuracy. From the perspective of modality interaction modeling, KEVL and CSFND primarily employ various attention mechanisms to capture the complex interactions between textual and visual modalities. These models achieve classification accuracy comparable to Vision-Language Pretraining models (VLP), which demonstrates the effectiveness of attention mechanisms in multimodal tasks. However, these models do not explicitly consider knowledge consistency between modalities, which may limit further improvements in their performance.

The model most conceptually aligned with QHCN is MDPF, which similarly enriches the multimodal feature space from a multi-level perspective and incorporates image descriptions as external knowledge. However, experimental results indicate that MDPF performs slightly worse than QHCN. We attribute this performance gap to two potential reasons. First, the value of image descriptions is limited. The image content in the subject question is often simple and information-scarce, and internal knowledge inconsistencies may further reduce the usefulness of image descriptions for performance improvement. Second, MDPF lacks explicit knowledge consistency modeling. While it focuses on cross-modal alignment between text and images, it does not explicitly consider knowledge consistency between the two modalities, which may limit its effectiveness in classification tasks. These findings highlight the importance of both multidimensional feature integration and intermodal knowledge consistency in achieving superior performance in multimodal learning.

4.3.2. Domain adaptation performance analysis

To comprehensively assess the generalization capability of QHCN across diverse multimodal tasks, we choose three representative tasks: multimodal affective analysis, multimodal fake news detection, and visual question answering, and conduct extensive experiments on the MVSA-Multiple, Twitter, and VQA v2.0 public datasets (Peng & Li, 2025). The innovation of QHCN lies in its decomposition of the dual modalities (image and text) in visual question into explicit and implicit four-dimensional features from both semantic and knowledge perspectives. Furthermore, it explores the complex interactions among multimodal features, including modality complementation, modality independence, and knowledge consistency. Further improvements are needed when applying this approach to the three domains. For multimodal affective analysis and fake news detection, the image-text pairs from the MVSA-Multiple and Twitter dataset can be utilized to extract four-dimensional features from both semantic and emotional perspectives. Subsequently, modules such as cross-modal alignment and emotional consistency can be applied to enhance performance. In contrast, for visual question answering, the VQA v2.0 dataset presents unique challenges. As a general-domain dataset with diverse and complex question types, the lack of an external knowledge base makes it difficult to identify knowledge-related attributes in the question and corresponding entities in the images. Consequently, QHCN focuses solely on extracting textual and visual features from visual questions and leverages only the modality complementarity and modality independence modules to generate answers.

The experimental results of the improved QHCN on the three public datasets (MVSA-Multiple, Twitter, and VQA v2.0) are presented in Table 2, the experimental results for all baseline models are obtained from these two reference studies (Li et al., 2024d; Lu et al., 2025). Due to the similarity between multimodal affective analysis and the original task of QHCN, the model demonstrates significant transferability to this domain. From Table 2, it is observed that enriching the data feature space with finer-grained four-dimensional features and exploring modal complementarity, modal independence, and affective consistency can effectively improve the sentiment classification accuracy by 1.2% compared to the optimal baseline model MDPF. As for the fake news detection task, although affective features can be used as important cues for fake news detection, aspects such as the authenticity of the text and image content are not considered, so the detection accuracy is slightly lower than the optimal baseline model MDPF. In the case of the visual question answering task, the unique characteristics of the VQA v2.0 dataset limit the full migration of QHCN. Specifically, the model only considers the textual and visual modalities, as well as the interactions of modality complementarity and modality independence, leading to suboptimal performance. In summary, QHCN is highly applicable to multimodal tasks with multi-level features, such as multimodal affective analysis and fake news detection. For visual question answering datasets, the introduction of an external knowledge base to extract knowledge-related features could enable the direct application of QHCN. The successful application of QHCN across multiple multimodal tasks further demonstrates its generalization capability and robustness.

4.3.3. Ablation study on QHCN

To examine the functionality of different components introduced in the overall architecture, we conduct an ablation study on the physics question dataset, which is mainly divided into three parts, 1) the selection of quaternion positions for explicit-implicit four-dimensional features, 2) the effectiveness of explicit-implicit features from multiple modalities, 3) the effectiveness of three major modules, including modality complementarity, modality independence, and knowledge consistency.

Selection of quaternion positions for explicit-implicit four-dimensional features. As shown in Table 3, the positions of different modal features within the quaternion components exert a certain influence on model performance. (1) **Uniqueness of the real part**

Table 2
Cross-dataset generalization evaluation on MVSA-multiple, Twitter and VQA v2.0.

Tasks	Datasets	Models	Accuracy	F1
Multimodal Affective Analysis	MVSA-Multiple	CoMN	0.689	0.688
		FENet	0.714	0.712
		CM-BERT	0.700	0.736
		MAG	0.751	0.743
		MVAN	0.724	0.723
		ITIN	0.735	0.735
		DMLANet	0.778	0.752
		MDPF + B&V	0.784	0.782
		$QHCN_{VB}$	**0.796**	**0.788**
Multimodal Fake News Detection	Twitter	EANN	0.648	0.639
		MAVE	0.745	0.744
		MCAN	0.796	0.786
		CAFE	0.806	0.806
		LIIMR	0.831	0.836
		MFIR	0.858	0.859
		CDD	0.874	0.856
		CSFND	0.833	0.831
		KEVL	0.826	0.824
		MDPF + B&V	**0.885**	**0.885**
		$QHCN_{VB}$	0.872	0.869
Visual Question Answering	VQA v2.0	GPT_O	0.547	–
		GPT_A	0.557	–
		$BERT_O$	0.495	–
		$BERT_A$	0.505	–
		CMQEF	**0.655**	–
		$QHCN_{VB}$	0.584	0.579

Table 3
Accuracy comparison under different quaternion component configurations.

Index	Types	Real	**i**	**j**	**k**	Accuracy
①	(E^t, E^v, K^v, K^t)	E^t	E^v	K^v	K^t	**0.9482**
②	(E^v, E^t, K^v, K^t)	E^v	E^t	K^v	K^t	0.9033
③	(E^t, K^t, E^v, K^v)	E^t	K^t	E^v	K^v	0.9418
④	(E^t, E^v, K^t, K^v)	E^t	E^v	K^t	K^v	0.9206
⑤	(E^t, K^t, K^v, E^v)	E^t	K^t	K^v	E^v	0.9352

selection. Comparing Strategy ① and Strategy ②, replacing the textual explicit feature E^t with the visual explicit feature E^v as the real part leads to a notable drop in accuracy (from 0.9482 to 0.9033). This demonstrates that assigning the less semantically rich visual feature as the real component destabilizes the reference frame of the fusion space, eliminating the scalar anchor and impairing rotational consistency. Therefore, the textual explicit feature E^t, which carries the densest and most expressive semantics, is the only theoretically appropriate choice for the real part. (2) **Ordering of visual components.** A comparison between Strategy ③ and Strategy ④ reveals that separating the two visual features E^v and K^v by inserting K^t causes performance degradation (from 0.9418 to 0.9206). This is because adjacent quaternion components interact directly through the Hamilton product. The explicit and implicit visual features (E^v and K^v) are intrinsically correlated. Their spatial adjacency facilitates a direct, coherent flow of information within visual forms, thereby enabling unified and self-consistent visual understanding. In contrast, separating them disrupts this natural coupling and weakens the model's capacity for holistic visual representation learning. (3) **Placement of the textual implicit knowledge features K^t.** With E^v and K^v kept adjacent, K^t can be positioned either at i (to link with E^t) or at k (to link with K^v). Comparing Strategies ①, ③ and ⑤ shows that positioning K^t at the k component yields the best performance (0.9482). This configuration effectively balances inter-modal interaction, intra-modal cohesion, and knowledge-level association. Strategy ③ fails to capture strong cross-modal links between E^t and E^v and neglects $K^t - K^v$ correlations, while Strategy ⑤ overlooks the expressive capacity of E^v. Therefore, the configuration (E^t, E^v, K^v, K^t) represents the optimal and theoretically justified design for quaternion-based multimodal fusion.

Effect of explicit-implicit features from multiple modalities. To verify the complementarity and non-redundancy of explicit and implicit features, this study conducted cross-modal explicit-implicit feature superposition experiments. To highlight the advantage of explicit-implicit features, we compare the performance of different combinations of four features from two modalities in the visual question. "ALL_F" denotes the question feature representation composed of explicit-implicit features, including textual-semantic(TS), textual-knowledge (TK), visual-semantic(VS), and visual-knowledge(VK). We combine four-dimensional features in pairs according to modality and explicit-implicit information to form four feature fusion strategies, namely, explicit-implicit features in textual modality

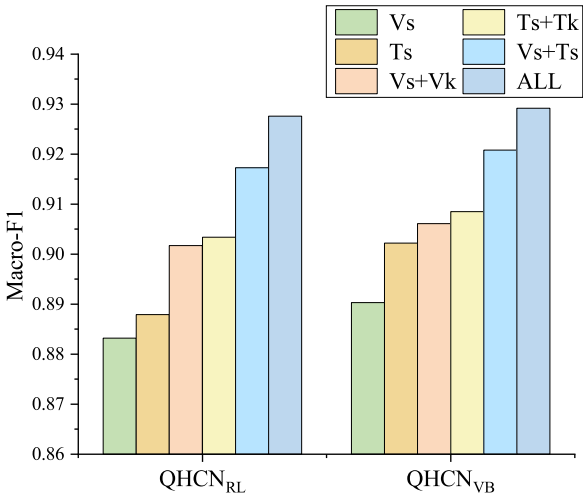


Fig. 7. Ablation study with different fusion strategies.

(TS + TK), explicit-implicit features in visual modality (VS + VK), the explicit feature combination of different modalities (TS + VS) and the four-dimensional feature fusion (ALL_F). All multi-feature fusion strategies directly concatenate them and feed them into two fully connected layers to formulate the final question representation. The results of different combinations of explicit-implicit features are listed in Fig. 7.

From Fig. 7, we can conclude that the contribution of textual representations to the entire model is higher than that of visual representations. Whether in textual modality or visual modality, the result is slightly better when explicit features are combined with implicit knowledge features, which indicates that implicit knowledge features are supplemented with richer information. The combination of ViT and BERT, adding multiple modal information can achieve a 3.05% higher macro-F1 and 1.68% higher macro-F1 score compared to the ViT and BERT, respectively. Similarly, the fusion of ResNet50 and Att-BLSTM significantly outperforms by 3.41%, 2.94% in macro-F1 compared to the ResNet50 and Att-BLSTM, respectively. This proves that the fusion of multi-source features has a large contribution to the final performance. The best performance is achieved when integrating explicit-implicit features from two modalities. No matter which dimensional feature is removed, the performance of the model degrades, which indicates that explicit-implicit features in both textual modality and visual modality are effective for visual question representation.

Effect of three major modules. To examine the functionality of each module of QHCNs, we sequentially add each one for comparison, the results are exhibited in Fig. 8. "Q-FC" denotes that each question is represented as a quaternion vector followed by the general fully connected layer. "ALL" denotes the QHCN with all modules, including modality complementarity (HGCN), independence loss (L_d), and knowledge consistency loss (L_k), which correspond to modality complementarity, modality independence, and knowledge consistency, respectively. From Fig. 8, we can observe that each component plays a significant role in improving the performance of QHCN. Q-$HGCN_{VB}$ boosts the performance over Q-FC to 1.49% accuracy and 1.38% macro-F1 score, which reveals that it is indeed helpful to learn multimodal interactions between explicit-implicit features by the Hamilton product and model the higher-order correlations between questions through the quaternion convolution operation. Adding independence loss (L_d) has an improved macro-F1 performance further by 0.25% compared to Q-$HGCN_{VB}$. On this basis, the knowledge consistency loss is added to form the entire model $QHCN_{VB}$ improved it further by 0.21%, yielding a net gain score of 1.84%. These demonstrate how well the quaternion hypergraph consistency network combined with the independence loss and knowledge consistency loss can significantly improve the performance of the network. Therefore, a good multimodal fusion scheme needs to consider the complementarity between explicit-implicit features while maintaining mutual independence and knowledge consistency.

4.3.4. Hypergraph construction and training strategy

Effect of Different hypergraph construction strategies. The hypergraph construction strategies plays a decisive role in the process of constructing a quaternion hypergraph. We study the impact of different hypergraph construction strategies on model performance. The four hypergraph construction strategies are illustrated in Fig. 9. "H_{inter}" denotes that some questions that are similar in both textual space and visual space are contained in a hyperedge. "H_{union}" indicates that as long as one of the textual features or the visual features is similar between questions, a hyperedge is constructed. "$H_{concate}$" refers to directly concatenating the incidence matrix of the textual hypergraph and the visual hypergraph, respectively, which doubles the number of hyperedges. "H_{merge}" is to first fuse the explicit-implicit features of different modalities, and then build the hyperedge based on KNN.

Fig. 10 compares the performance of four hypergraph construction strategies at $QHCN_{RL}$ and $QHCN_{VB}$. It is observed that "H_{inter}" outperforms strongly the other construction strategies in terms of accuracy and macro-F1 score by a large margin. This is because the questions on each hyperedge are consistent across explicit-implicit features, thus reducing the number of hyperedges and eliminating redundant information. Taking $QHCN_{VB}$ as an example, "H_{merge}" perform slightly worse than "H_{inter}", the accuracy is relatively decreased by 1.28%, and the macro-F1 score is relatively decreased by 1.39%. The performance of "H_{union}"n is close to

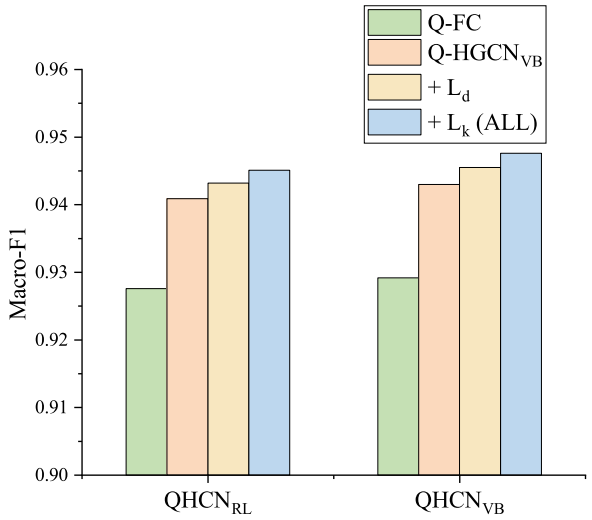


Fig. 8. Ablation study with three major modules, including modality complementarity, modality independence, and knowledge consistency.

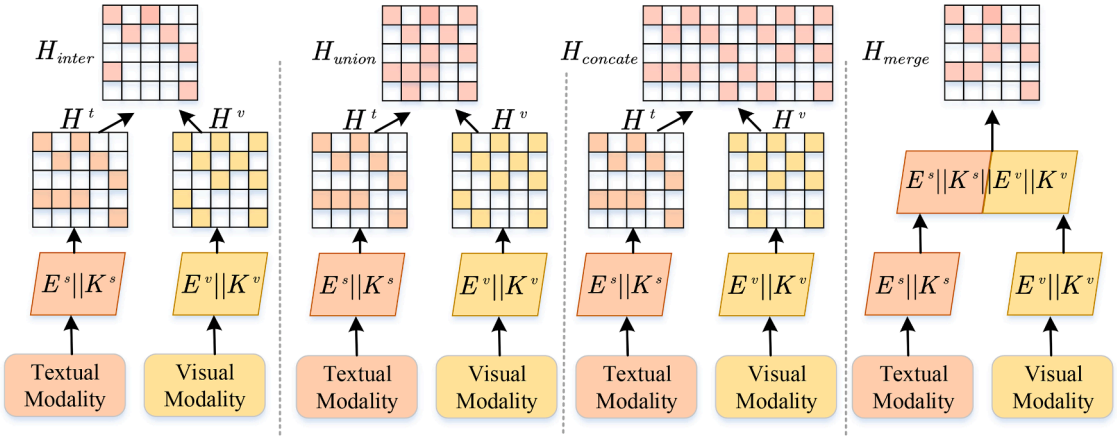


Fig. 9. Illustration of four quaternion hypergraph construction methods.

"$H_{concate}$", achieving the worst accuracy. One possible reason is that redundant hyperedges can cause harmful effects bringing in malicious noise, which damages the feature representation of visual questions and confuses the model. These results demonstrate the effectiveness of our approach in choosing "H_{inter}" to construct a quaternion hypergraph for performance gains.

Comparison of dynamic hypergraphs and static hypergraphs. To examine the effect of adaptive topology learning, we implement a dynamic hypergraph generator based on the K-nearest neighbor (KNN) mechanism, in which hyperedge connections are adaptively updated according to evolving feature representations during training. Under identical experimental settings-including a two-layer hypergraph convolutional network (QHCN), consistent feature dimensionality, and identical optimization parameters-we compared this dynamic variant with a static hypergraph constructed from a fixed feature similarity matrix. The dynamic design aims to capture evolving relationships among samples, while the static design leverages a stable, domain-specific prior.

The experimental results in Table 4 reveal that the static hypergraph achieved superior overall performance, with accuracy approximately 0.6% higher than that of the dynamic counterpart (0.9482 vs. 0.9420 in accuracy), while training efficiency improved dramatically (1.2s vs. 18s per epoch). The dynamic version also introduced nearly three times more parameters due to its additional structure-update module. These results suggest that, in our setting involving a shallow network and moderate-scale multimodal data, dynamically updating the hypergraph structure does not yield tangible benefits but instead increases computational cost and optimization difficulty.

We attribute this performance gap to two main factors. First, the shallow QHCN lacks sufficient capacity to jointly learn node representations and hypergraph structure, and thus benefits more from the strong inductive bias embedded in a predefined static topology, which guides feature propagation and ensures stable optimization. Second, dynamically updating connections can amplify feature noise, leading to unstable edge formation and a higher risk of overfitting, particularly in small-batch multimodal settings.

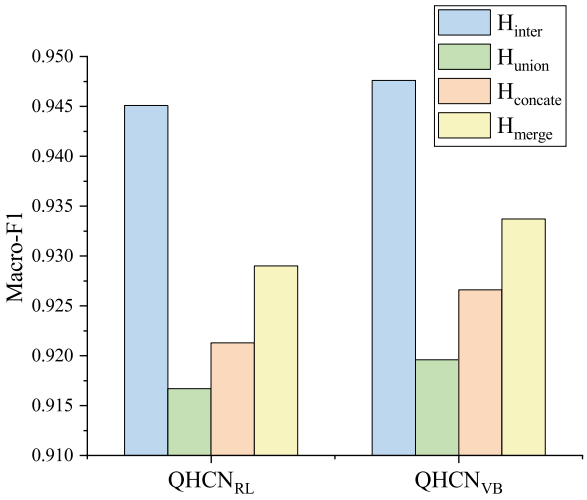


Fig. 10. Macro-F1 of four quaternion hypergraph construction methods in $QHCN_{RL}$ and $QHCN_{VB}$.

Table 4
Comparison between static and dynamic hypergraphs.

Hypergraph Type	Accuracy	F1-Score	Training Time (s/epoch)	# Parameters (M)
Static ($QHCN_{VB}$)	**0.9482**	**0.9476**	**1.2**	**1.06**
Dynamic (D–$QHCN_{VB}$)	0.9420	0.9408	18	3.18

Overall, the static hypergraph design demonstrates an effective balance between accuracy, stability, and efficiency, while dynamic strategies may prove advantageous only in deeper architectures or larger datasets.

4.3.5. Hyperparameter research

Effect of the nearest neighbor value K**.** In the feature fusion stage, the hyperparameter K determines the number of neighbors connected by each hyperedge, thereby directly controlling the density of the constructed quaternion hypergraph. To select an appropriate K for both $QHCN_{RL}$ and $QHCN_{VB}$, we conducted a grid search within the range of 10-100 and evaluated ten representative values 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. As shown in Fig. 11, $QHCN_{VB}$ achieves its highest accuracy and macro-F1 when $K = 60$, while $QHCN_{RL}$ performs best at $K = 30$. Performance degradation occurs when K deviates from these optimal values, indicating that the optimal neighborhood size is closely tied to the intrinsic feature characteristics of each backbone model.

From the perspective of feature quality, this difference arises from the representational properties of the two encoder pairs. In $QHCN_{VB}$ (ViT + BERT), both encoders employ self-attention mechanisms that enable global contextual modeling, producing highly discriminative and tightly clustered features within each class. Consequently, a larger neighborhood size ($K = 60$) helps the hypergraph capture more homogeneous samples, enrich intra-class relationships, and enhance generalization. In contrast, $QHCN_{RL}$ (ResNet + Att-BLSTM) generates features that are more sensitive to local variations and noise, and thus a smaller $K = 30$ preserves hyperedge purity by avoiding the inclusion of heterogeneous neighbors.

Effect of independence loss and knowledge loss. To evaluate different values of the hyper-parameters λ_1 and λ_2, we conduct a series of experiments to explore their different combinations on the performance of $QHCN_{VB}$. Specifically, we adopt different values of λ_1 and λ_2 in the range from 0 to 1, respectively, and they add up to 1. Fig. 12 reports the model performance with different hyper-parameters on the test set of our dataset. Note that the setting of $\lambda_2 = 0$ corresponds to applying only the cross-entropy loss and independence loss, while $\lambda_1 = 0$ applies only the cross-entropy loss and knowledge consistency loss. It is observed that our $QHCN_{VB}$ achieves the best classification performance with $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$. However, when λ_1 and λ_2 change relatively, the accuracy of knowledge point classification degrades, which indicates that the performance is very sensitive to the value of λ_1 and λ_2. Meanwhile, we observe that both the independence loss module and the knowledge consistency module contribute to the improvement of the model results, which indicates that both modality independence and knowledge consistency are effective for knowledge-based visual question classification.

4.3.6. Complexity analysis

Since QHCN employs pre-trained ViT and BERT as its visual and textual encoders respectively, we comparable baseline models (e.g. ViLT, ViLBERT, and MDPF) for complexity analysis to ensure evaluation fairness. These baseline models similarly utilize transformer architectures for multimodal feature extraction, thus maintaining methodological consistency in our comparative analysis. We conduct a quantitative evaluation of computational complexity for QHCN and comparable baseline models through both space complexity and time complexity metrics. Specifically, space complexity is measured by the total parameter count of each model, while time

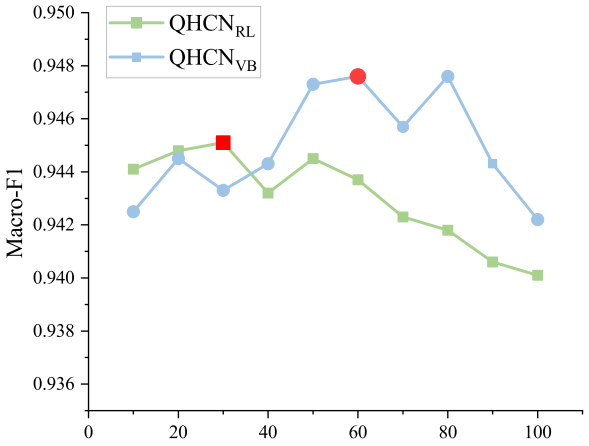


Fig. 11. Effect of the nearest neighbor value K. The red dots represent the best K value in $QHCN_{RL}$ and $QHCN_{VB}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

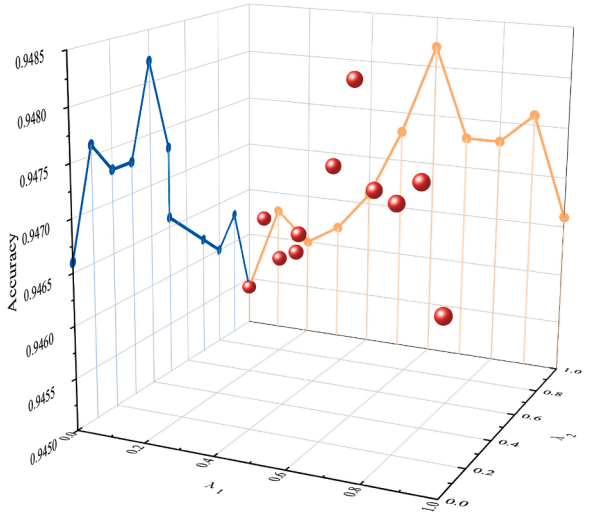


Fig. 12. Effect of the hyper-parameters λ_1 and λ_2 in $QHCN_{VB}$.

complexity is quantified using floating-point operations (FLOPs). The experimental results are presented in Table 5, where n denotes the sequence length of question texts, m represents the number of image patches, and L_B, L_V, L_T indicate the layer numbers of BERT, ViT, and Transformer components, respectively. The variables d_B, d_V, d_T correspond to the feature dimensions of their respective networks. The parameters P_R, P_{Res}, P_{GRU},P_{Att}, P_{LMM}, P_{GCN} denote the parameter counts of Faster R-CNN, ResNet, GRU, Attention, LLM, and GCN modules, while C_R, C_{Res}, C_{GRU},C_{Att}, C_{LMM}, C_{GCN}, C_{KNN} represent their corresponding computational costs measured in FLOPs.

The experimental results in Table 5 demonstrate that ViLBERT and ViLT, as general-purpose multimodal fusion models based on Transformer architectures, exhibit comparable space complexity. However, their time complexity varies significantly due to distinct modality interaction mechanisms. Specifically, ViLBERT's dual-stream design introduces higher computational overhead compared to ViLT's single-stream approach. However, neither of these models explicitly models modality independence and knowledge consistency, which limits their performance on complex multimodal tasks. In contrast, models such as HGCN, KEVL, CSFND, and VG-CALF have incorporated corresponding improvements during the multimodal fusion stage by introducing additional modules including attention mechanisms, GRUs, and GCNs, albeit at the cost of increased spatial and temporal complexity. The proposed QHCN model employs four distinct methods to extract explicit and implicit four-dimensional features, which inevitably introduces additional model complexity and computational overhead. However, this impact is mitigated through two key design innovations: (1) a lightweight graph convolutional network in the multimodal fusion stage, and (2) modality independence and knowledge consistency modules implemented via carefully designed loss functions that require minimal additional parameters. In comparison with models such as KEVL, CSFND, and VG-CALF, the proposed approach achieves superior visual question classification accuracy while maintaining

Table 5
Comparative analysis of space complexity and time complexity of different models.

Models	Space Complexity	Time complexity
HGCN	$O(L_T \cdot d_T^2 + P_{GCN})$	$O(\lvert E\rvert \cdot d_T + \lvert V\rvert \cdot d_T^2)$
ViLBERT	$O(L_T \cdot d_T^2)$	$O\left(L \cdot \left(\begin{smallmatrix} n^2 \cdot d_T + m^2 \cdot d_T + n \cdot d_T^2 \\ + m \cdot d_T^2 \end{smallmatrix}\right)\right)$
ViLT	$O(L_T \cdot d_T^2)$	$O(L \cdot ((n+m)^2 \cdot d_T + (n+m) \cdot d_T^2))$
KEVL	$O(L_B \cdot d_B^2 + L_T \cdot d_T^2 + P_R)$	$O\left(\begin{smallmatrix} L_B \cdot (n^2 \cdot d_B + n \cdot d_B^2) + \\ L_T \cdot (m^2 \cdot d_T + m \cdot d_T^2) + C_R \end{smallmatrix}\right)$
CSFND	$O(L_B \cdot d_B^2 + P_{\text{Res}} + P_{GRU} + P_{Att})$	$O\left(\begin{smallmatrix} L_B \cdot (n^2 \cdot d_B + n \cdot d_B^2) + \\ C_{\text{Res}} + C_{KNN} + C_{GRU} + C_{Att} \end{smallmatrix}\right)$
VG-CALF	$O(L_B \cdot d_B^2 + L_V \cdot d_V^2 + P_{Att})$	$O\left(\begin{smallmatrix} L_V \cdot (m^2 \cdot d_V + m \cdot d_V^2) + \\ L_B \cdot (m^2 \cdot d_B + m \cdot d_B^2) + C_{Att} \end{smallmatrix}\right)$
MDPF	$O(L_B \cdot d_B^2 + L_V \cdot d_V^2 + P_{LLM} + P_{GCN} + P_{Att})$	$O\left(\begin{smallmatrix} L_V \cdot (m^2 \cdot d_V + m \cdot d_V^2) + \\ L_B \cdot (m^2 \cdot d_B + m \cdot d_B^2) + \\ C_{LLM} + C_{GCN} + C_{Att} \end{smallmatrix}\right)$
QHCN	$O(L_B \cdot d_B^2 + L_V \cdot d_V^2 + P_{GCN} + P_B)$	$O\left(\begin{smallmatrix} L_V \cdot (m^2 \cdot d_V + m \cdot d_V^2) + \\ L_B \cdot (m^2 \cdot d_B + m \cdot d_B^2) + \\ C_{GCN} + C_R \end{smallmatrix}\right)$

comparable model complexity. QHCN also slightly improves model performance while reducing complexity compared to MDPF, which also considers multiple features.

In summary, QHCN achieves an optimal balance between training complexity and model performance. Unlike most multimodal fusion algorithms that significantly increase computational complexity by incorporating excessive modules to enhance performance, QHCN intelligently enriches the visual question feature space via multi-dimensional embeddings and employs a compact architectural design to effectively capture intricate cross-feature interactions, consequently boosting classification performance.

4.3.7. Robustness experiment

In real-world educational applications, multimodal questions often contain various noises, such as irrelevant words in text or extraneous background objects in images. To evaluate its robustness in realistic noisy scenarios, we conduct controlled experiments comparing the performance degradation of the proposed QHCN against ViLT.

For textual noise, we randomly select 10% and 20% of samples from the original test set and add common nouns/adjectives unrelated to the knowledge domain at random positions (e.g., “weather, concert, travel, discount”), avoiding synonyms of labeled terms. Similarly, for 10% and 20% of the samples randomly selected, an irrelevant object (e.g., “cup,” “backpack”) from the MS-COCO dataset is superimposed onto non-core areas of the corresponding question image, with its size and transparency adjusted to simulate real image interference. The powerful and widely adopted ViLT model is employed as the baseline for comparison across three noise conditions: textual noise only, visual noise only, and their combination. All models are trained on clean, uncorrupted data and evaluated on test sets with varying levels of noise injection, with robustness measured by the degree of performance degradation.

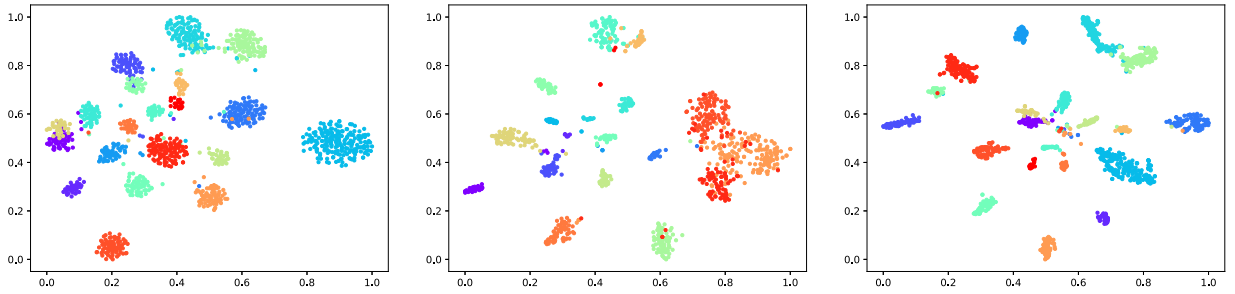
As shown in Table 6, under three noise settings and two interference intensities, the accuracy decline of the $QHCN_{VB}$ model is significantly smaller than that of the ViLT model. When subjected to $+10\%$ and $+20\%$ text noise, the performance of $QHCN_{VB}$ decreases by only 1.81% and 3.31%, respectively, whereas ViLT suffers more significant drops of 3.15% and 5.28%. Similar trends are observed under visual noise, where $QHCN_{VB}$ shows smaller accuracy declines (-1.48% and -2.96%) compared to ViLT (-2.77% and -4.87%). In addition, the results reveal that textual noise has a slightly greater impact on model performance than visual noise for both architectures. The mixed-noise setting, which combines $+10\%$ text and $+10\%$ visual perturbations, leads to the largest performance degradation (-3.87% for $QHCN_{VB}$ and -6.07% for ViLT). This can be attributed to the potential overlap of corrupted samples across modalities, amplifying noise interference and disrupting cross-modal consistency.

The underlying reason for this phenomenon may be that textual noise primarily contaminates textual explicit semantic features (E^t). However, the core of QHCN lies in its parallel extraction of textual implicit knowledge features (K^t). These features are extracted by constructing knowledge attribute graph based on a predefined knowledge attribute lexicon. This process itself functions as an efficient noise filter, capable of ignoring irrelevant words like “cup” and “backpack” while precisely capturing attributes related to core knowledge concepts (such as “density” and “leverage”). This ensures the purity and stability of the knowledge representation. Similarly, irrelevant objects superimposed in images primarily interfere with explicit visual semantic features (E^v). The robustness of QHCN stems from its extraction method for implicit visual knowledge features (K^v). These features are typically obtained by matching the input image against a series of prototypical knowledge concept images. This prototype-driven matching mechanism enables the model to focus on key visual patterns related to physical principles while actively ignoring visual elements unrelated to the knowledge prototypes (such as the cup in the background).

Experimental results demonstrate that the $QHCN_{VB}$ model exhibits significant robustness advantages over the powerful ViLT baseline when confronted with textual and visual noise. This not only validates the model’s practical value in real-world educational scenarios but also indirectly confirms the effectiveness of its design. By decoupling explicit and implicit features and leveraging knowledge-guided feature extraction, the model constructs more resilient multimodal representations that are less susceptible to surface-level noise interference.

Table 6
Performance comparison under different noise settings. The values in parentheses denote performance drops relative to the clean data.

Noise Settings	Noise Ratio	$QHCN_{VB}$	ViLT
Textual Noise	+10% Text Noise	0.9295 (−1.81%)	0.9008 (−3.15%)
	+20% Text Noise	0.9145 (−3.31%)	0.8795 (−5.28%)
Visual Noise	+10% Visual Noise	0.9328 (−1.48%)	0.9046 (−2.77%)
	+20% Visual Noise	0.9180 (−2.96%)	0.8836 (−4.87%)
Mixed Noise	+10% Text Noise +10% Visual Noise	0.9089 (−3.87%)	0.8716 (−6.07%)



(a) $HGCN_{early}$ (b) ViLT (c) $QHCN_{VB}$

Fig. 13. Visualization of question features in reduced dimensions using T-SNE with three fusion methods. (a) fusion features extracted with $HGCN_{early}$ showing the 20 clusters of the test set tend to be scattered. (b) fusion features extracted with ViLT showing the 20 clusters of the test set with better separated with less entanglements. (c) fusion features extracted with $QHCN_{VB}$, introducing the implicit knowledge features and three key properties among explicit-implicit features, showing the 20 clusters of the test set with more concentrated and well separation.

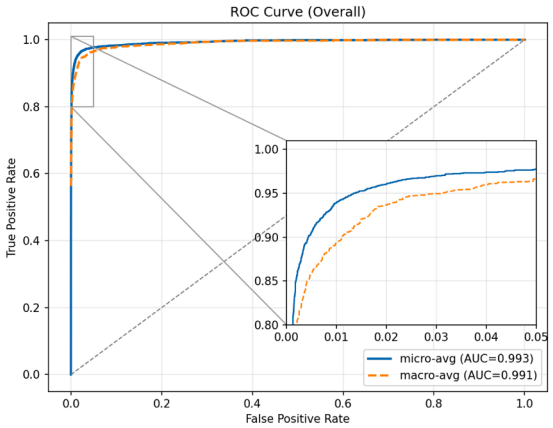
4.3.8. Visualization experiment

Visualization of Question Representation. To help qualitatively evaluate our multimodal fusion method, we visualize the question features learned by $QHCN_{VB}$ on physics question dataset with t-SNE. To clearly show that our method learns richer question representations than other models, such as the state-of-the-art ViLT and $HGCN_{early}$. We randomly select 20 of 46 knowledge points and visualize the feature distribution of questions annotating these knowledge points. As shown in Fig. 13, each color represents a knowledge point, and each point refers to the feature representation of a question reduced from a high-dimensional space to a two-dimensional space.

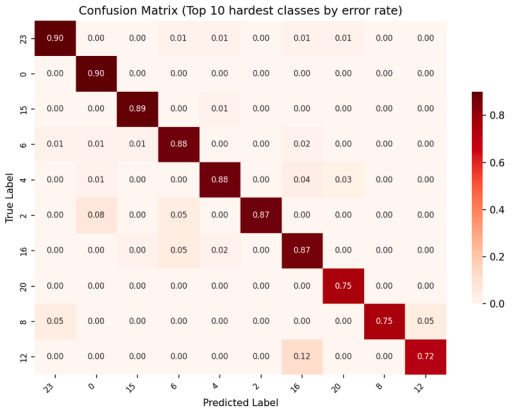
From Fig. 13, we can observe that the aggregative and discriminative of the feature representation learned by $QHCN_{VB}$ is much better than $HGCN_{early}$ and ViLT. These multimodal fusion methods fuse explicit semantic features in different modalities, which promotes the intra-class aggregation of each knowledge point to a certain extent, but some stubborn questions make the demarcation between knowledge points unclear, resulting in knowledge points confusion and misjudgment, as shown in Fig. 13(a) and (b). Compared with $HGCN_{early}$, ViLT has more concentrated question features within the same knowledge point, and the feature boundary between different knowledge points is also clearer, but there are still a few knowledge points that are not distinguished from each other. Relatively speaking, the features learned by $QHCN_{VB}$ are more discriminable with a more significant segregated area between visual questions of all knowledge points, as exhibited in Fig. 13(c). This is likely because QHCN not only considers explicit semantic features, but also learns implicit knowledge features of textual modality and visual modality in combination with educational domain. Moewovwe, three kinds of interactions between explicit-implicit features, including modality complementarity, modality independence, and knowledge consistency, are analyzed in fine-grained manner. The question representation is greatly enriched by considering multi-dimensional features and fine-grained interaction between explicit-implicit features, thus achieving better performance.

Analysis of model discrimination capability. To further validate the discriminative capability of the proposed model, we conduct an additional experiment to analyze the ROC curves and corresponding AUC scores. As illustrated in Fig. 14(a), the proposed $QHCN_{VB}$ achieves AUC = 0.993 (micro-average) and AUC = 0.991 (macro-average), demonstrating excellent classification performance. The zoomed-in subplot (False Positive Rate $\in [0, 0.05]$) highlights that both curves remain near the top-left corner, indicating the model's strong robustness under stringent false-positive constraints. These results confirm that our approach maintains high sensitivity and specificity, effectively distinguishing positive samples even under low false alarm conditions.

Visualization of confusion matrix. To provide deeper insight into the classification behavior of the proposed QHCN, we construct the confusion matrix of classification results. Because the dataset involves 46 knowledge-points, directly visualizing the complete 46 × 46 matrix would make detailed interpretation difficult. Therefore, we focused on the 10 categories with the highest error rates, which represent the most challenging fine-grained distinctions, as shown in Fig. 14(b).



(a) ROC curves and AUC scores of QHCN



(b) Confusion matrix of the top 10 hardest knowledge-point classes ranked by error rate.

Fig. 14. ROC curves and confusion matrix of QHCN.

The diagonal elements of the matrix exhibit generally high accuracy values (0.72-0.90) for most difficult classes, confirming that the model performs robustly even under hard conditions. However, several off-diagonal elements indicate localized confusion between semantically or visually correlated knowledge points. For instance, misclassifications occur between knowledge points 12 and 16 (representing “Mass and Density”) and between points 0 and 2 (representing “Buoyancy and Pressure”). Their corresponding questions frequently involve overlapping conceptual cues, such as analogous problem contexts, similar diagrammatic elements, or parallel linguistic structures, causing their feature representations to converge in the embedding space. However, since our model explicitly incorporates implicit knowledge features extracted from both textual and visual modalities, these confusions have been significantly alleviated compared with baseline multimodal encoders. The integration of hidden semantic cues helps the model better capture contextual dependencies and latent conceptual relations, thereby improving the classification accuracy of difficult categories to the 0.72-0.90 range.

5. Discussion

To comprehensively verify the effectiveness and robustness of the proposed QHCN model, a series of extensive experiments were conducted. First, we thoroughly evaluated the overall performance of QHCN by comparing it with unimodal and multimodal baseline models, including both general-domain and education-domain specialized models, as well as several cross-task baselines. The results consistently show QHCN’s exceptional performance across these comparisons. To further assess its generalization ability across different multimodal tasks, we applied QHCN to three representative tasks: multimodal sentiment analysis, multimodal fake news detection, and visual question answering, and performed experiments on three public datasets: MVSA-Multiple, Twitter, and VQA v2.0. The results confirm that QHCN effectively adapts to heterogeneous data distributions and task objectives. Ablation studies were conducted to explore the functional contributions of different modules: (1) the effectiveness of multimodal explicit-implicit feature modeling; (2) the impact of quaternion positional configuration among the four dimensions; and (3) the roles of the three major modules: modality complementarity, modality independence, and knowledge consistency. Since the quaternion hypergraph convolution network is the core of our architecture, we further investigated hypergraph construction strategies and compared static and dynamic variants. We also analyzed two key hyperparameters to study their influence on performance and stability. Moreover, quantitative assessments of computational complexity (in terms of spatial and temporal cost) indicate that QHCN achieves an optimal trade-off between training efficiency and performance. To validate the model’s robustness under noisy real-world conditions, we conducted comparative noise-injection experiments with ViLT, showing that QHCN exhibits smaller performance degradation across textual, visual, and mixed noise settings. Finally, visualization analyses, including t-SNE embeddings, ROC curves, and confusion matrices, demonstrate that the learned question features are well clustered within knowledge categories and discriminative across different concepts. These outstanding results stem from three key innovations in our model design:

(1) Semantic-knowledge decomposition for four-dimensional explicit-implicit representations.

From both semantic and knowledge perspectives, QHCN decomposes multimodal information (image and text) into explicit and implicit components, constructing a quaternion representation that unifies them into a four-dimensional feature space. The explicit dimension captures observable semantic cues, such as object appearances and textual entities, while the implicit dimension models latent cognitive and contextual factors related to domain knowledge. This decomposition allows the model to distinguish between surface-level descriptions and conceptual-level understanding, thereby improving interpretability and transferability across educational tasks.

(2) Quaternion hypergraph for high-order multimodal relational modeling.

Building upon quaternion algebra, QHCN introduces a quaternion hypergraph convolution network that enables structured interaction among multimodal features. Unlike conventional multimodal fusion networks that rely on pairwise concatenation or attention mechanisms, the quaternion hypergraph encodes high-order relationships through hyperedges, allowing the model to capture subtle correlations among text, visual regions, and their associated knowledge entities. Moreover, the quaternion formulation not only enhances expressiveness via rotational coupling between real and imaginary parts but also reduces parameter redundancy through weight sharing across four interdependent components, achieving an elegant balance between complexity and performance.

(3) In-depth exploration of complex interactions among multimodality.

To explore the complex interactions among multiple modalities, QHCN introduces three collaborative modules. The modality complementarity module enhances cross-modal cooperation by allowing key information from one modality to compensate for the deficiencies of another. The modality independence module suppresses redundant or noisy information to ensure the discriminability of modality-specific feature distributions. The knowledge consistency mechanism guarantees semantic alignment between visual and textual representations under the same knowledge concept. In particular, the knowledge consistency module plays a crucial role in stabilizing the multimodal embedding space and guiding feature propagation within the quaternion hypergraph.

Together, these structural innovations enable QHCN to unify explicit and implicit semantics, visual and textual modalities, and local and global relationships under a single quaternion hypergraph framework, providing both representational interpretability and cross-domain generalization.

6. Conclusion

In this study, we present a quaternion hypergraph consistent network (QHCN), a multimodal fusion architecture for knowledge-based visual question classification that integrates both explicit semantic features and implicit knowledge features. It provides a more effective way of utilizing the three key properties of explicit-implicit features, including modality complementation, modality independence, and knowledge consistency. Specifically, a quaternion hypergraph is constructed based on the consistency of explicit-implicit features in textual space and visual space. The Hamiltonian product and quaternion convolution operator can enable fusion and interaction of multimodal features. To further consider intramodal and intermodal complex correlations in the fusion scheme, we introduce the independence loss and knowledge consistency loss respectively in the training process. Extensive experiments and comprehensive analysis show that our model outperforms other multimodal fusion models. Our QHCN can generate more robust knowledge-based question representations by dynamically modeling correlations between multidimensional features in quaternion space.

Despite the good performance of our QHCN, we can explore more in enriching the question representation. On the one hand, we can enrich the question representation by leveraging the disciplinary knowledge base and introducing the rich expertise of thought chains (Qiu et al., 2024) and LLMs (e.g., ChatGPT) (Wang et al., 2024). Moreover, we explicitly acknowledge that extending the proposed algorithm to other scientific disciplines and datasets is a key direction for future work. We are currently planning to build the necessary domain knowledge bases for other scientific disciplines, which will serve as the central focus of our subsequent studies. This cross-disciplinary expansion will allow us to validate the generality and adaptability of our framework across diverse educational and cognitive domains.

CRediT authorship contribution statement

Jing Wang: Writing - review & editing, Writing - original draft, Software, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Duantengchuan Li:** Writing - review & editing, Writing - original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization; **Xu Du:** Writing - review & editing, Visualization, Data curation; **Hao Li:** Writing - review & editing, Formal analysis; **Zhuang Hu:** Writing - review & editing, Software, Data curation.

Data availability

Data will be made available on request.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Nos. 62407009, 62507035) and Youth Project of Chongqing Municipal Education Commission (No. KJQN202400642).

References

Bai, S., Zhang, F., & Torr, P. H. S. (2021). Hypergraph convolution and hypergraph attention. *Pattern Recognition*, *110*, 107637.

Bi, S., Liu, J., Miao, Z., & Min, Q. (2024). Difficulty-controllable question generation over knowledge graphs: A counterfactual reasoning approach. *Information Processing & Management*, *61* (4), 103721.

Bloch, I., & Bretto, A. (2024). Functional analysis on hypergraphs: Density and zeta functions - applications to molecular graphs and image analysis. *Information Sciences*, *676*, 120850.

Dixit, C., & Satapathy, S. M. (2024). *Expert Systems with Applications*, *240*, 122579.
Dou, Z.-Y., Xu, Y., Gan, Z., Wang, J., Wang, S., & Wang (2022). An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18166–18176).
Duan, J., Xiong, J., Li, Y., & Ding, W. (2024). Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, *112*, 102536.
Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020). Early vs late fusion in multimodal convolutional neural networks. In *IEEE 23rd international conference on information fusion, FUSION* (pp. 1–6). IEEE.
Gao, X., Wang, X., Chen, Z., Zhou, W., & Hoi, S. C. H. (2024). Knowledge enhanced vision and language model for multi-modal fake news detection. *IEEE Transactions on Multimedia*, *26*, 8312–8322.
Gaudet, C. J., & Maida, A. S. (2018). Deep quaternion networks. In *2018 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.
Guizzo, E., Weyde, T., Scardapane, S., & Comminiello, D. (2023). Learning speech emotion representations in the quaternion domain. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1200–1212.
Guo, J., Zhao, M., Yu, J., Yu, R., Song, J., Wang, Q., Xu, L., & Yu, M. (2025). EHPR: Learning evolutionary hierarchy perception representation based on quaternion for temporal knowledge graph completion. *Information Sciences*, *688*, 121409.
Guo, Z., Zhao, J., Jiao, L., Liu, X., & Liu, F. (2021). A universal quaternion hypergraph network for multimodal video question answering. *IEEE Transactions on Multimedia*, *25*, 38–49.
Hamilton, W. R. (1844). II. On quaternions; or on a new system of imaginaries in algebra. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *25* (163), 10–13.
Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L.-p., & Poria, S. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction* (pp. 6–15). Association for Computing Machinery.
Huang, J., Pu, Y., Zhou, D., Cao, J., Gu, J., Zhao, Z., & Xu, D. (2024). Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing*, *565*, 126992.
Huang, X., & Gong, H. (2024). A dual-attention learning network with word and sentence embedding for medical visual question answering. *IEEE Transactions on Medical Imaging*, *43*, 832–845.
Ji, L., Tu, R., Lin, K., Wang, L., & Duan, N. (2022). Multimodal graph neural network for video procedural captioning. *Neurocomputing*, *488*, 88–96.
Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., & Chen, X. (2020). In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.
Jiang, J., Wei, Y., Feng, Y., Cao, J., & Gao, Y. (2019). Dynamic hypergraph neural networks. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19* (pp. 2635–2641).
Jin, Y., Yin, W., Wang, H., & He, F. (2024). Capturing word positions does help: A multi-element hypergraph gated attention network for document classification. *Expert Systems with Applications*, *251*, 124002.
Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th international conference on machine learning* (pp. 5583–5594). PMLR (*vol. 139*). Proceedings of Machine Learning Research.
Lameesa, A., Silpasuwanchai, C., & Alam, M. S. B. (2025). VG-CALF: A vision-guided cross-attention and late-fusion network for radiology images in medical visual question answering. *Neurocomputing*, *613*, 128730.
Li, D., Gao, Y., Wang, Z., Qiu, H., Liu, P., Xiong, Z., & Zhang, Z. (2024a). Homogeneous graph neural networks for third-party library recommendation. *Information Processing & Management*, *61* (6), 103831.
Li, D., Tang, J., Wang, P., Li, S., & Wang, T. (2025a). Maximizing discrimination masking for faithful question answering with machine reading. *Information Processing & Management*, *62* (1), 103915.
Li, D., Xia, T., Wang, J., Shi, F., Zhang, Q., Li, B., & Xiong, Y. (2024b). SDFormer: A shallow-to-deep feature interaction for knowledge graph embedding. *Knowledge-Based Systems*, *284*, 111253.
Li, H., Wang, J., Du, X., Hu, Z., & Yang, S. (2023a). KBHN: A knowledge-aware bi-hypergraph network based on visual-knowledge features fusion for teaching image annotation. *Information Processing & Management*, *60* (1), 103106.
Li, M., Zhang, Y., Li, X., Cai, L., & Yin, B. (2022). Multi-view hypergraph neural networks for student academic performance prediction. *Engineering Applications of Artificial Intelligence*, *114*, 105174.
Li, M., Zhou, S., Chen, Y., Huang, C., & Jiang, Y. (2024c). Educross: Dual adversarial bipartite hypergraph learning for cross-modal retrieval in multimodal educational slides. *Information Fusion*, *109*, 102428.
Li, S., Gong, C., Zhu, Y., Luo, C., Hong, Y., & Lv, X. (2024d). Context-aware multi-level question embedding fusion for visual question answering. *Information Fusion*, *102*, 102000.
Li, X., Guo, S., Wu, J., & Zheng, C. (2025b). An interpretable polytomous cognitive diagnosis framework for predicting examinee performance. *Information Processing & Management*, *62*, 103913.
Li, Z., Zhang, Q., & Zhu, F. (2023b). Knowledge graph representation learning with simplifying hierarchical feature propagation. *Information Processing & Management*, *60* (4), 103348.
Li, Z., Zhang, Q., Zhu, F., Li, D., Zheng, C., & Zhang, Y. (2023c). Knowledge graph representation learning with simplifying hierarchical feature propagation. *Information Processing & Management*, (*60*, 103348.
Liao, W., Liu, Z., Dai, H., Wu, Z., Zhang, Y., Huang, X., Chen, Y., Jiang, X., Liu, D., Zhu, D., Li, S., Liu, W., Liu, T., Li, Q., Cai, H., & Li, X. (2024). Mask-guided BERT for few-shot text classification. *Neurocomputing*, *610*, 128576.
Liu, H., Jiao, P., Guo, X., Wu, H., Gao, M., & Zhang, J. (2024a). HGN2T: A simple but plug-and-play framework extending hgnns on heterogeneous temporal graphs. *IEEE Transactions on Big Data* (pp. 620–632). (*vol. 10*).
Liu, M., Zhang, J., Nyagoga, L. M., & Liu, L. (2024b). Student-AI question cocreation for enhancing reading comprehension. *IEEE Transactions on Learning Technologies*, *17*, 815–826.
Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurIPS 2019, december 8-14, 2019, vancouver, BC, canada* (pp. 13–23).
Lu, Q., Sun, X., Long, Y., Zhao, X., Zou, W., Feng, J., & Wang, X. (2025). Multimodal dual perception fusion framework for multimodal affective analysis. *Information Fusion*, *115*, 102747.
Luo, Y., Chen, H., Yin, T., Horng, S.-J., & Li, T. (2024). Dual hypergraphs with feature weighted and latent space learning for the diagnosis of Alzheimer's disease. *Information Fusion*, *112*, 102546.
Minemoto, T., Isokawa, T., Nishimura, H., & Matsui, N. (2017). Feed forward neural network with random quaternionic neurons. *Signal Processing*, *136*, 59–68.
Morgado, P., Vasconcelos, N., & Misra, I. (2021). Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12475–12486).
Nguyen, D. Q., Nguyen, T. D., & Phung, D. (2021a). Quaternion graph neural networks. In *Proceedings of the 13th asian conference on machine learning* (pp. 236–251). PMLR (*vol. 157*).
Nguyen, T., Bui, T., Fujita, H., Hong, T.-P., Loc, H. D., Snasel, V., & Vo, B. (2021b). Multiple-objective optimization applied in extracting multiple-choice tests. *Engineering Applications of Artificial Intelligence*, *105*, 104439.
Nie, W., Liang, Q., Wang, Y., Wei, X., & Su, Y. (2020). MMFN: Multimodal information fusion networks for 3D model classification and retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16* (4), 1–22.
Nizarudeen, S., & Shanmughavel, G. R. (2024). Comparative analysis of resnet, resnet-SE, and attention-based ranet for hemorrhage classification in CT images using deep learning. *Biomedical Signal Processing and Control*, *88*, 105672.

Ogri, O. E., EL-Mekkaoui, J., Benslimane, M., & Hjouji, A. (2024). Automatic lip-reading classification using deep learning approaches and optimized quaternion meixner moments by GWO algorithm. *Knowledge-Based Systems*, *304*, 112430.
Pan, H., & Huang, J. (2022). Multimodal high-order relational network for vision-and-language tasks. *Neurocomputing*, *492*, 62–75.
Parcollet, T., Morchid, M., & Linarès, G. (2020). A survey of quaternion neural networks. *Artificial Intelligence Review*, *53* (4), 2957–2982.
Peng, D., & Li, Z. (2025). Unbiased VQA via modal information interaction and question transformation. *Pattern Recognition*, *162*, 111394.
Peng, L., Jian, S., Kan, Z., Qiao, L., & Li, D. (2024). Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. *Information Processing & Management*, *61* (1), 103564.
Pham, N.-Q., Niehues, J., Ha, T.-L., & Waibel, A. (2019). Improving zero-shot translation with language-independent constraints. In *Proceedings of the fourth conference on machine translation (volume 1: Research papers)* (pp. 13–23). Association for Computational Linguistics.
Qiu, C., Xie, Z., Liu, M., & Hu, H. (2024). Explainable knowledge reasoning via thought chains for knowledge-based visual question answering. *Information Processing & Management*, *61* (4), 103726.
Schwung, A., Pöppelbaum, J., & Nutakki, P. C. (2021). Rigid body movement prediction using dual quaternion recurrent neural networks. In *2021 22nd IEEE international conference on industrial technology (ICIT)* (pp. 756–761). IEEE (*vol. 1*).
Shao, R., Bi, X.-J., & Chen, Z. (2024). Hybrid ViT-CNN network for fine-grained image classification. *IEEE Signal Processing Letters*, *vol. 31*, 1109–1113.
Shao, W., Peng, Y., Zu, C., Wang, M., & Zhang, D. (2020). Hypergraph based multi-task feature selection for multimodal classification of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *80*, 101663.
Silva, V. A., Bittencourt, I. I., & Maldonado, J. C. (2018). Automatic question classifiers: A systematic review. *IEEE Transactions on Learning Technologies*, *12* (4), 485–502.
Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., & Rohrbach, M. (2019). Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.
Song, J., Wang, Y., Zhang, C., & Xie, K. (2024). Self-attention and forgetting fusion knowledge tracking algorithm. *Information Sciences*, *680*, 121149.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15* (1), 1929–1958.
Sun, B., Zhu, Y., Xiao, Y., Xiao, R., & Wei, Y. (2019). Automatic question tagging with deep neural networks. *IEEE Transactions on Learning Technologies*, *12* (1), 29–43.
Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6558–6569). Association for Computational Linguistics.
Udandarao, V., Agarwal, A., Gupta, A., & Chakraborty, T. (2021). InPHYNet: Leveraging attention-based multitask recurrent networks for multi-label physics text classification. *Knowledge-Based Systems*, *211*, 106487.
Wang, C., Huang, K., & Shi, W. (2022). An accurate and efficient quaternion-based visualization approach to 2D/3D vector data for the mobile augmented reality map. *ISPRS International Journal of Geo-Information*, *11* (7), 383.
Wang, J., Li, H., Du, X., Hung, J.-L., & Yang, S. (2023a). S-KMN: Integrating semantic features learning and knowledge mapping network for automatic quiz question annotation. *Journal of King Saud University - Computer and Information Sciences*, *35* (7), 101594.
Wang, J., Zhang, Q., Shi, F., & Li, D. (2023b). Knowledge graph embedding model with attention-based high-low level features interaction convolutional network. *Information Processing & Management*, *60* (4), 103350.
Wang, Q., Ji, R., Peng, T., Wu, W., Li, Z., & Liu, J. (2024). Soft knowledge prompt: Help external knowledge become a better teacher to instruct LLM in knowledge-based VQA. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2024* (pp. 6132–6143).
Wang, X., Han, X., Huang, W., Dong, D., & Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.
Wu, F., Jing, X.-Y., Wei, P., Lan, C., Ji, Y., Jiang, G.-P., & Huang, Q. (2022). Semi-supervised multi-view graph convolutional networks with application to webpage classification. *Information Sciences*, *591*, 142–154.
Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.
Yadav, A., & Vishwakarma, D. K. (2023). A deep multi-level attentive network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing Communications and Applications*, *19*, (1),19.
Yao, K., Liang, J., Liang, J., Li, M., & Cao, F. (2022). Multi-view graph convolutional networks with attention mechanism. *Artificial Intelligence*, *307*, 103708.
Yu, Z., Yu, J., Xiang, C., Fan, J., & Tao, D. (2018). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, *29* (12), 5947–5959.
Zhang, Q., Li, Y., Zou, J., Zhu, J., Liu, D., & Jiao, J. (2024). Unifying multimodal interactions for rumor diffusion prediction with global hypergraph modeling. *Knowledge-Based Systems*, (*301*, 112246.
Zhao, C., Xiong, C., Boyd-Graber, J., & Daumé III, H., (2021). Distantly-supervised dense retrieval enables open-domain question answering without evidence annotation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9612–9622).
Zhao, Q., Xia, Y., Long, Y., Xu, G., & Wang, J. (2025). Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. *Information Processing & Management*, *62* (1), 103876.
Zhou, D., Huang, J., & Schölkopf, B. (2006). Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems , 19*.
Zhu, X., Xu, Y., Xu, H., & Chen, C. (2018). Quaternion convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 631–647).