Multi-Scale Hypergraph Convolutional Network for Low-Quality Text-Based Scene Retrieval

Fengjing Song[1,2,3], Shengrong Zhao[1,2,3], Hu Liang[1,2,3*]
1 Key Laboratory of Computing Power Network and Information Security,
Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
2 Shandong Engineering Research Center of Big Data Applied Technology,
Faculty of Computer Science and Technology,
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
3 Shandong Provincial Key Laboratory of Industrial Network and Information System Security,
Shandong Fundamental Research Center for Computer Science, Jinan, China

***Abstract*—Image-text retrieval is a core task in vision-language. However, current cross-modal methods often rely heavily on high-quality training data, which tends to cause issues such as performance degradation when the model is applied to scenarios with semantic ambiguity or incomplete data. To address this problem, we propose the Modal Multi-Scale Hyperedge Retrieval (MMHR) model. Specifically, its Multi-Scale Hyperedge Module (MHM) introduces a fine-grained hyperedge taxonomy to reinforce cross-modal/cross-scale node connections, while the Multi-Dimensional Matrix Hypergraph Convolution Module (MMHC) designs dual multidimensional matrices to quantify node contribution weights and enhance node-hyperedge interactions, helping the model prioritize nodes with high information density during the modeling process. These designs enable the model to adjust relationships between nodes of different modalities, thereby calibrating cross-modal retrieval. Experiments on Flickr30K and MSCOCO demonstrate that MMHR outperforms existing methods in cross-modal retrieval.**

***Index Terms*—Image-text retrieval, hypergraph convolution, multimodal alignment**

I. Introduction

Image-text retrieval matches images and text via content-driven semantic similarity, with relevance to tasks like image captioning [1], text-to-image synthesis [2], action understanding [3], multimodal machine translation [4], scene graph generation [5], and zero-shot learning [6]. Despite academic and industrial attention and prior progress, the semantic gap between image content and linguistic descriptions in complex scenarios remains a key barrier to practical systems.

A. Maintaining the Integrity of the Specifications

It has two core tasks: text-to-image retrieval (selecting relevant images for text) and image-to-text retrieval (finding descriptive sentences for images). Early works employed two-branch networks to map different modalities into a shared embedding space for similarity calculation. While these methods

This work was supported by Natural Science Foundation of Shandong Province (ZR2025MS1082), National Key Research and Development Program (No.2023YFB3308403), Natural Science Foundation of Shandong Province (No.ZR2022LZH008). (*Corresponding author: Hu Liang.)

Fig. 1. Schematic diagrams corresponding to different semantic methods: (a) Cross-attention retrieval method (b) Hypergraph Convolution retrieval method.

are efficient and scalable, they primarily capture global correspondences, lack fine-grained interactions, and consequently fail to bridge the semantic gap in complex scenarios.Recent methods (SCAN [7], IMRAM [8]) adopt cross-attention for dynamic bimodal alignment, outperforming coarse-grained retrieval, yet cross-attention is computationally heavy, reducing efficiency and scalability.

Despite extensive research, the cross-modal semantic gap persists due to differing modal complexities and information volumes. Most methods assume complete training data, yet real-world data incompleteness or misalignment exacerbates heterogeneity gaps; cross-attention further introduces redundant alignments and ambiguity, with irrelevant segments disrupting cross-modal interaction. In multi-object or ambiguous text scenarios, hypergraphs outperform cross-attention by avoiding redundancy: as shown in Figure 1, cross-attention computes weights for all visual region-text block pairs (inducing redundancy), while hypergraphs use hyperedges to link semantically similar visual regions.

Recent graph-based methods have made efforts to improve cross-modal alignment: HGAN [9] proposes a hierarchical graph alignment framework to capture multi-level correlations,

Multilateral Semantic Relations Modeling [10] focuses on modeling diverse semantic connections between modalities, and Heterogeneous Graph Fusion Network [11] integrates heterogeneous graph structures for cross-modal fusion. However, these methods have three limitations in handling low-quality text: (1) they lack multi-scale hyperedge design, failing to simultaneously capture spatial, semantic, and cross-modal correlations; (2) they do not model node contribution differences within hyperedges, leading to interference from irrelevant nodes; (3) they ignore correlations between hyperedges, limiting contextual information propagation. Our work complements these efforts by designing a multi-scale hypergraph framework with adaptive weight calibration, specifically tailored for low-quality text scenarios.

Based on the above analysis, this paper proposes the Modal Multi-Scale Hyperedge Retrieval (MMHR) model, which advances existing hypergraph-based cross-modal retrieval methods through three novel designs, addressing the core challenge of semantic ambiguity in low-quality text scenarios. Unlike prior works that either adopt single-type hyperedges [18,19] or ignore fine-grained node-hyperedge interaction modeling [20], MMHR integrates the Multi-Dimensional Matrix Hypergraph Convolution Module (MMHC) and Multi-Scale Hyperedge Module (MHM) to strengthen cross-scale vision-text semantic associations.

Specifically, MHM introduces a fine-grained multi-scale hyperedge taxonomy (three types) that simultaneously captures spatial, semantic, and cross-modal correlations—an innovation compared to existing hypergraph methods that only focus on partial associations. The three hyperedge types are: cross-modal hyperedges, intra-modal positional hyperedges, and intra-modal conventional hyperedges. For the visual modality, we use Euclidean distance to connect target regions with their neighboring regions (γ=5, optimized on Flickr30K validation set) to construct intra-modal positional hyperedges, which explicitly model spatial structure and resolve ambiguity in multi-object scenarios. For the cross-modal setting, we select text-visual nodes with cosine similarity $\geq$0.3 (Flickr30K validation set-optimized threshold) to build cross-modal hyperedges—this threshold-guided selection balances fine-grained alignment and redundancy reduction, avoiding "global-information-only shortcuts" that plague traditional methods.

Additionally, MMHC proposes a multi-dimensional matrix-driven hypergraph convolution mechanism, which complements the limitations of prior hypergraph convolution that either neglect node contribution differences or hyperedge correlations. MMHC optimizes hypergraph convolution via two core matrices: (1) a node contribution matrix that quantifies node weights using feature transformation and hyperedge-specific attention, eliminating irrelevant node interference; (2) a hyperedge correlation matrix that integrates feature similarity and node connectivity to capture semantic connections between hyperedges. These two matrices collaborate to strengthen node-hyperedge interactions, enabling MMHR to prioritize high-information-density nodes and align cross-modal semantics in semantically ambiguous or data-incomplete scenarios.

Compared with existing graph-based methods, MMHR's key innovations are threefold: (1) a tripartite hyperedge taxonomy integrating spatial, semantic, and cross-modal correlations; (2) threshold-guided cross-modal hyperedge construction (τ=0.3) that balances fine-grained alignment and redundancy suppression; (3) adaptive hypergraph convolution driven by dual-matrix collaboration. These designs synergistically address the core challenges of low-quality text, including semantic ambiguity and information incompleteness.

To summarize, this work makes three key contributions: (1) We propose a multi-scale hyperedge module that captures spatial, semantic, and cross-modal correlations simultaneously, addressing the limitations of single-type hyperedge designs. (2) We design a dual-matrix hypergraph convolution mechanism to quantify node contributions and hyperedge correlations, enhancing feature aggregation under semantic ambiguity. (3) We conduct comprehensive experiments on low-quality text scenarios (random word removal, synonym replacement, word reordering) to verify the robustness of MMHR, providing new insights for cross-modal retrieval under imperfect data.

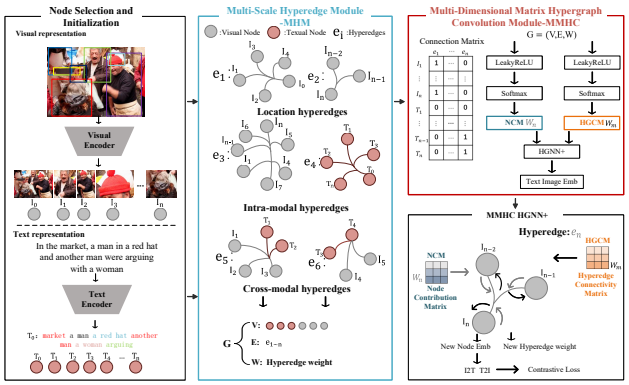

Fig. 2. The MMHR architecture comprises two single-modal encoders (image and text) and integrates two core modules—MMHC and MHM—which perform hypergraph convolution to support subsequent cross-modal retrieval.

II. METHOD

This section presents the proposed Modal Multi-Scale Hyperedge Retrieval Model (MMHR), which advances existing hypergraph-based cross-modal retrieval through a structured multi-scale hyperedge design and a novel multi-dimensional matrix convolution mechanism. As shown in Figure 2, MMHR comprises two single-modal encoders (image and text) and two core modules: the Multi-Scale Hyperedge Module (MHM) and the Multi-Dimensional Matrix Hypergraph Convolution Module (MMHC), which are detailed below.

A. The Multi-Scale Hyperedge Module

Multimodal hypergraphs contain visual and textual nodes, connected by hyperedges to establish their multivariate key correlations. To capture high-order semantic correlations between the two node types, we design three hyperedge types: intra-modality, intra-modality positional, and cross-modality.

For the visual modality, regions exhibit spatial correlations. Let b_i denote the bounding box of the i-th region from Faster R-CNN [12]. Using Euclidean distance to measure inter-region distances, we construct a hyperedge containing I_i and its nearest γ neighbors to ensure each represents a locally coherent spatial group. In experiments, $\gamma = 5$ (optimized on the Flickr30K validation set) achieves the highest performance, while other γ values yield lower performance: smaller γ misses semantically related regions, while larger γ introduces redundant background. Intra-modal Positional Hyperedges: The Euclidean distance-based neighbor calculation formula follows:

$$Dis(I_i, I_j) = \sqrt{(c_{i_x} - c_{j_x})^2 + (c_{i_y} - c_{j_y})^2}, \quad (1)$$

where (c_{i_x}, c_{i_y}) are the central coordinates of image region I_i's bounding box, (c_{j_x}, c_{j_y}) those of I_j's, and $Dis(I_i, I_j)$ the distance between I_i and I_j.

To capture cross-modality interactive relationships and high-order semantic correlations, we design cross-modal hyperedges: first calculate text-visual node similarity via cosine similarity to determine co-hyperedge connectivity, then set a similarity threshold $\tau = 0.3$ (Flickr30K validation set-optimized) and select node pairs with $Sim(w_i, I_j) \geq \tau$ for construction. Their definition is as follows:

$$\mathcal{E}_{\text{text-visual}} = (w_1, \ldots, w_n, I_1, I_2, \ldots, I_m), \quad (2)$$

$$\mathcal{E}_{\text{text}} = (w_1, w_2, \ldots, w_n), \quad (3)$$

$$\mathcal{E}_{\text{visual}} = (I_1, I_2, \ldots, I_m), \quad (4)$$

where $\mathcal{E}_{text}$ denotes intra-modality semantic hyperedges for textual nodes, $\mathcal{E}_{\text{visual}}$ for visual nodes, w_n the n-th text node, and I_m the m-th visual node.$\mathcal{E}_{\text{text-visual}}$ denotes the cross-modal hyperedge connecting text node w_1 and visual node I_m, with $i \in [1, m]$.

B. The Multi-Dimensional Matrix Hypergraph Convolution Module

Traditional hypergraph convolution methods either treat all nodes within a hyperedge as equally important or ignore the semantic correlations between hyperedges, leading to suboptimal feature aggregation under low-quality data. To address this, we theoretically derive the dual-matrix collaboration mechanism: the node contribution matrix C_e is motivated by the attention mechanism, which adaptively assigns weights based on node-hyperedge relevance, while the hyperedge correlation matrix A extends the graph Laplacian theory to hypergraphs, enabling contextual information propagation between semantically related hyperedges.

Hypergraph convolution relies on effective modeling of node-hyperedge interactions. During the feature update process, node features are first aggregated into hyperedges, and then hyperedge features propagate back to update nodes. To address the ambiguity of node contributions and the lack of correlation modeling between hyperedges in traditional hypergraph convolution, the MMHC module optimizes the process through two core multidimensional matrices (node contribution matrix and hyperedge correlation matrix) with clear functional divisions, ensuring interpretable and efficient cross-modal feature refinement.

1) Node Contribution Matrix: The node contribution matrix C_e quantifies the importance of each node within a specific hyperedge e , which is critical for filtering irrelevant nodes and highlighting high-information-density elements (e.g., key text words or core visual regions). Its construction integrates feature transformation and hyperedge-specific attention, with the specific definition as follows:

$$C_e = Softmax\left((X^{(l)} W_c) \cdot a^e\right), \quad (5)$$

where C_e(dimension N×1) is the node contribution matrix for hyperedge e, where $C_e(i)$ represents the weight of the i-th node to hyperedge e;$X^{(l)}$(dimension N×D) is the node feature matrix at the l-th iteration, where N denotes the total number of nodes and D denotes the feature dimension;W_c (dimension D×D') is the feature transformation matrix, which maps the original D-dimensional node features to a D'-dimensional space to enhance discriminability;a^e(dimension D'×1) is the learnable attention vector specific to hyperedge e, which measures the relevance between each transformed node feature and hyperedge e.

Notably, nodes that do not belong to hyperedge e are assigned a contribution value of 0 in C_e, which effectively avoids interference from irrelevant nodes and ensures that only semantically related nodes participate in hyperedge feature aggregation.

2) Hyperedge Correlation Matrix: To capture the semantic connections between different hyperedges and promote contextual information propagation, we construct a hyperedge correlation matrix A by fusing two key cues: feature similarity (reflecting semantic consistency) and node connectivity (reflecting structural association). The correlation between any two hyperedges e_i and e_j is defined as:

$$A_{i,j} = \alpha \cdot \frac{f_i \cdot f_j}{\|f_i\| \cdot \|f_j\|} + (1 - \alpha) \cdot \frac{n_{i,j}}{\max n_{i,k}}, \quad (6)$$

where $A_{i,j}$ is the correlation coefficient between hyperedge e_i and e_j (ranging from 0 to 1); f_i and f_j are the feature vectors of hyperedge e_i and e_j (obtained by aggregating node features within the hyperedge via weighted sum based on C_e); $n_{i,j}$ is the number of shared nodes between e_i and e_j (reflecting the structural connectivity of the two hyperedges); $\max_k n_{i,k}$ is the maximum number of shared nodes between hyperedge e_i and any other hyperedge e_k (normalizing the node connectivity term); $\alpha \in [0, 1]$ is a balancing coefficient (set to 0.5 after cross-validation on the Flickr30K validation set) that adjusts the weights of feature similarity and node connectivity.

Based on the hyperedge correlation matrix A, the hyperedge weight matrix is updated to dynamically adjust the importance of each hyperedge during feature propagation. The update formula is:

$$W_{\text{hyper}}^{(l+1)} = Softmax_{i,j}\left(W_{hyper}^{l} \cdot \Theta \cdot A\right), \quad (7)$$

where W_{hyper}^{l}(dimension $E \times E$) is the initial hyperedge weight matrix at the l-th iteration;Θ(dimension $E \times E$) is a learnable matrix that dynamically adjusts the relevance between hyperedges;The Softmax function normalizes the matrix

row-wise to ensure the sum of weights for each hyperedge is 1.

3) Hypergraph Convolution Layer: The hypergraph convolution layer is stacked with residual structures to avoid gradient vanishing. The complete feature update process integrates the node contribution matrix C and the updated hyperedge weight matrix $W_{\text{hyper}'}^{(l+1)}$, with the specific formula as follows:

$$X^{(l+1)} = \sigma\left(D_v^{-\frac{1}{2}} H W_{hyper}^{(l+1)} D_e^{-1} H^T C D_v^{-\frac{1}{2}} X^{(l)} \Theta^{(l)}\right), \quad (8)$$

where $X^{(l+1)}$ (dimension $N \times D'$) is the updated node feature matrix at the $(l+1)$-th iteration; $\sigma(\cdot)$ denotes the activation function (ReLU is adopted in this work); $D_v^{-\frac{1}{2}}$ (dimension $N \times N$) is the square root inverse of the node degree matrix, where $\tilde{D}_v(i,i) = \sum_{j=1}^{E} H(i,j)$; D_e^{-1} (normalizing node features to avoid scale imbalance); D_e^{-1} (dimension $E \times E$) is the inverse of the hyperedge degree matrix, where $D_e(j,j) = \sum_{i=1}^{N} H(i,j)$ (normalizing hyperedge features); H (dimension $N \times E$) is the hypergraph incidence matrix, where $H(i,j) = 1$ if the i-th node belongs to the j-th hyperedge, otherwise $H(i,j) = 0$; H^T is the transpose of the incidence matrix, enabling feature conversion between nodes and hyperedges; C (dimension $N \times N$) is a diagonal matrix, where the diagonal element $C(i,i)$ is the average of the contribution weights of node i in all hyperedges containing it (derived from Eq. (5), with non-participating nodes contributing 0); $\Theta^{(l)}$ (dimension $D' \times D'$) is the learnable linear transformation matrix at the l-th iteration, enhancing the expressive ability of the updated features.

In summary, the node contribution matrix C_e resolves intra-hyperedge node importance ambiguity, while the hyperedge correlation matrix A compensates for inter-hyperedge correlation neglect. Their synergy optimizes hypergraph convolution, empowering the model to adaptively prioritize high-value nodes/hyperedges under semantic ambiguity.

C. The Modal Multi-Scale Hyperedge Alignment Model

This section details the core components of the proposed model: visual/textual representation learning, hypergraph convolution mechanism and matching strategy.

Visual Representation: Pre-trained Faster R-CNN captures visual content of an image, generating n proposals with feature and bounding box vectors. The image representation is $V = \{v^0, v^1, \ldots, v^n\}$, where $v^i = [x^i, b^i]$ (concatenating visual feature x^i and bounding box b^i). v^0 is the global image feature (with b^0 for the entire image). Images and regions are converted into nodes; cross-modal alignment is established via hyperedge and hypergraph convolution modules.

Text Representation. BERT [13] and BiGRU [14] model text semantics: BiGRU processes GloVe-embedded words, deriving query features from averaged hidden states; BERT generates $n+1$ tokens for an n-word sentence, with text represented as $X = \{x^0, x^1, ..., x^n\}$ (global x^0, local word representations $x^1, ..., x^n$). Sentences and words become nodes; cross-modal alignment is achieved via hyperedge and hypergraph convolution modules, emphasizing regional and global text features.

Hypergraph Convolution. This module has two sub-modules: Multi-Dimensional Matrix Hypergraph Convolution (MMHC) and Multi-Scale Hyperedge (MHM). MMHC uses a node contribution matrix to evaluate intra-hyperedge node contributions and inter-node importance, plus a hyperedge connectivity matrix (fusing node connectivity and hyperedge similarity) to capture inter-hyperedge correlations and strengthen node-hyperedge connections.MHM classifies nodes via multi-scale hyperedges, with each hyperedge type serving a unique role.

Matching Method. Fine-grained retrieval is used: local retrieval computes similarity by aligning multimodal local elements via their features (measured on hypergraph-convolved nodes). Let $X = \{x_1, ..., x_{n_x}\}$ and $Y = \{y_1, ..., y_{n_y}\}$ denote post-convolution image and text node features, respectively. Samples are aligned via cosine similarity between each pair (X: local image features; Y: local text features), with local similarity defined as:

$$S_l(X,Y) = \frac{1}{n_y} \sum_{j \in [1,n_y]} \max_{i \in [1,n_x]} \frac{x_i^T \times y_j}{||x_i|| \times ||y_j||}, \quad (9)$$

where $\frac{x_i^T y_j}{\|x_i\| \|y_j\|}$ denotes the similarity between image node x_i and text node y_j; cross-modal elements are aligned via maximum similarity. Let n_y be the number of words in Y: for each element in Y, its best match in X is identified, with local sample similarity averaging these best similarities.

III. Experiment

A. Datasets and Experimental Setup

To ensure fair comparisons with existing methods, we evaluate our model on two public datasets: Flickr30K and MSCOCO.

Flickr30K contains 31,000 images (5 captions each) with standard splits: 29K training, 1K validation, 1K testing.MS-COCO includes 123,287 images (616,435 captions) with annotations. Following standard splits (113,287 training, 5K validation, 5K testing), results are averaged over 5-fold 1K tests with full 5K results reported.Evaluation uses top-K recall (R@K) with R@1, R@5, R@10 as metrics.

Implementation: Built with PyTorch 2 and trained on an RTX 3090. Text uses Hugging Face's bert-base-uncased (768D); images use Faster-R-CNN (4096D). Optimized with AdamW (weight decay=0.01, batch size=16, lr=4e-5) for 30 epochs (2 hypergraph iterations), evaluated from epoch 10, selecting the best validation model. All baselines use identical splits.

B. Comparison Results

We compare MMHR with recent methods on two benchmarks (Flickr30K and MS COCO), testing all methods on both high-quality datasets and low-quality ones. The low-quality text datasets are constructed with three types of semantic degradation to comprehensively evaluate robustness: (1) random word removal (5%, 10%, 20%, 30%); (2) synonym replacement (5%, replacing words with WordNet synonyms to simulate paraphrasing ambiguity); (3) word reordering (10%,

reordering consecutive word pairs to simulate grammatical distortion). The following are the results of high-quality datasets, and the low-quality dataset results are presented in Table 4.

For cross-modal retrieval (e.g., image-text), we use RSUM to quantify performance: RSUM=R@1+R@5+R@10 (R@K= Recall@K, proportion of correct top-K matches). Covering image→text and text→image directions, RSUM sums six values (R@1/R@5/R@10 for each direction) to evaluate overall recall. Text is processed via BiGRU and pre-trained BERT.

Table 1 shows the quantitative results of our model in the Flickr30k test set. The enhanced retrieval performance demonstrates that the MMHR framework effectively captures the cross modal shared semantics through the hypergraph model.

TABLE I
RESULT OBTAINED ON FLICKR 30K.

METHOD	Flickr30K Dataset						
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD+BIGRU							
HREM [15]	79.5	94.3	97.4	59.3	85.1	91.2	506.8
CHAN [16]	79.7	94.5	97.3	60.2	85.3	90.7	507.8
BOOM [17]	82.1	96.4	98.3	61.5	86.0	91.4	515.7
CHAN+ESSE [18]	80.2	94.6	97.2	60.9	85.6	90.8	509.3
RANet* [19]	81.1	96.1	97.6	62.5	87.1	92.2	516.5
MMHR(ours)	**82.5↑**	**96.9↑**	**98.2↑**	62.5	**88.2↑**	**92.6↑**	**520.9↑**
BUTD+BERT							
HREM [15]	83.3	96.0	98.1	63.5	87.1	92.4	520.4
CHAN [16]	80.6	96.1	97.8	63.9	87.5	92.6	518.5
BOOM [17]	84.3	96.5	98.5	65.2	88.6	93.1	526.2
RANet* [19]	83.3	97.2	98.6	67.4	90.1	94.3	531.0
MMHR(ours)	**85.1↑**	**97.7↑**	**99.1↑**	67.3	**90.6↑**	**94.9↑**	**534.7↑**

Table 2 shows the results of our model on MS COCO(5K and 5-fold 1K test set). The proposed MMHR framework performs well on both low-quality and high-quality text datasets.

TABLE II
RESULT OBTAINED ON MS COCO(5K AND 5-FOLD 1K TEST)

METHOD	COCO5-fold 1K Test						
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD+BIGRU							
NAAF [20]	78.1	96.1	98.6	63.5	89.6	95.3	521.2
CHAN [16]	79.7	96.7	98.7	63.8	90.4	95.8	525.0
BOOM [17]	80.5	96.7	99.0	64.5	91.1	96.3	528.1
MMHR(ours)	80.3	**96.9↑**	**99.2**	**65.0↑**	**91.4↑**	**96.7↑**	**529.5↑**
BUTD+BERT							
CHAN [16]	81.4	96.9	98.9	66.5	92.1	96.7	532.6
BOOM [17]	82.3	97.2	99.1	67.7	92.6	96.8	535.7
RANet* [19]	82.8	97.5	98.9	69.6	93.1	97.2	539.2
MMHR(ours)	**83.4↑**	**97.9↑**	**99.6↑**	**69.9↑**	**93.8↑**	**97.6↑**	**542.2↑**

C. Ablation Studies

In the ablation study, the MMHR framework has two key modules: the Multi-Dimensional Matrix Hypergraph Convolution Module (MMHC) and Multi-Scale Hyperedge Module (MHM). We assessed each component's impact on retrieval performance via separate ablation. The baseline here is the CHAN model—a two-branch model using Faster R-CNN (visual encoder) and BERT/BiGRU (text encoder)—which employs attention for alignment and computes cross-modal similarity directly via cosine similarity.

METHOD	COCO 5K Test						
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD+BIGRU							
NAAF [20]	58.9	85.2	92.0	42.5	70.9	81.4	430.9
CHAN [16]	60.2	85.9	92.4	41.7	71.5	81.7	433.4
BOOM [17]	60.5	86.2	92.6	43.4	72.1	82.5	437.3
MMHR(ours)	**60.8↑**	86.1	**93.1↑**	**44.0↑**	71.6	**83.3↑**	**438.9↑**
BUTD+BERT							
CHAN [16]	59.8	87.2	93.3	44.9	74.5	84.2	443.9
BOOM [17]	63.1	87.7	93.4	45.2	75.5	85.7	450.6
RANet* [19]	60.6	86.3	93.0	45.3	75.6	85.5	446.3
MMHR(ours)	**62.8↑**	**88.2↑**	**93.6↑**	**45.0↑**	**75.8↑**	**86.0↑**	**451.4↑**

Table 3 shows the ablation study results of our model on the Flickr30K test set.

TABLE III
RESULT OBTAINED ON FLICKR30K.

METHOD	Flickr30K Dataset						
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
BUTD+BIGRU							
Baseline	79.7	94.5	97.3	60.2	85.3	90.7	507.8
MMHR w/o MMHC	80.4	95.1	97.6	60.1	85.5	90.5	509.2
MMHR w/o MHM	81.2	96.1	97.3	61.0	84.9	91.3	511.8
MMHR(ours)	**82.5**	**96.9**	**98.2**	**62.5**	**88.2**	**92.6**	**520.9**
BUTD+BERT							
Baseline	80.6	96.1	97.8	63.9	87.5	92.6	518.5
MMHR w/o MMHC	82.2	96.4	98.2	64.7	88.2	93.1	522.8
MMHR w/o MHM	83.9	96.7	98.0	65.3	88.7	93.5	526.1
MMHR(ours)	**85.1**	**97.7**	**99.1**	**67.3**	**90.6**	**94.9**	**534.7**

D. Case Studies and Visualization

In this analysis, we compared our model's retrieval performance with state-of-the-art models under low-quality text scenarios. Experiments were conducted on complete text datasets and low-quality ones (with 5%, 10%, 20%, 30% of words randomly removed).

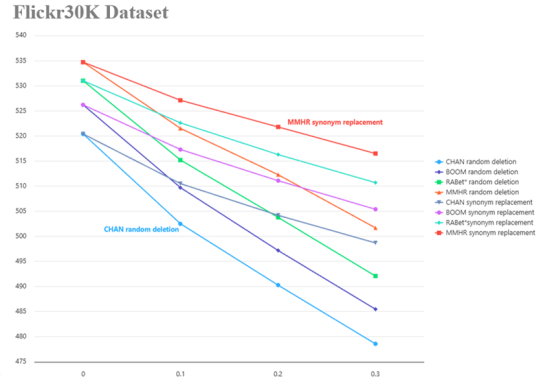

Fig. 3. On the Flickr30K dataset, comparative method experiments were conducted on the proposed MMHR framework under varying degrees of text semantic loss and synonym substitution.

IV. CONCLUSION

To improve the retrieval accuracy in scenarios with semantic ambiguity, this paper designs a novel cross-modal retrieval model named MMHR from the perspective of hypergraph convolution. During the hyperedge construction process, the MMHC is used to quantify the node contribution, which

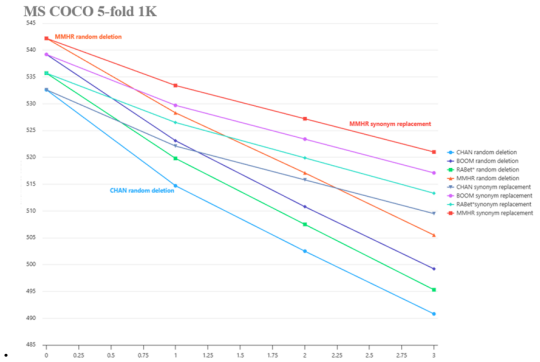

Fig. 4. On the MS C0C0 5-fold 1K dataset, comparative method experiments were conducted on the proposed MMHR framework under varying degrees of text semantic loss and synonym substitution.

helps the model adjust the relationships between nodes across different modalities and weaken the interference of redundant information. In addition, the MHM is utilized to enhance the connectivity between nodes and improve the model's ability to capture the associations between objects. Experimental results demonstrate that these improvements not only exhibit excellent performance in the presence of semantic ambiguity but also achieve good retrieval effectiveness in high-quality datasets for the cross-modal retrieval task.

V. Reference

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[2] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[3] L. Zhang, X. Chang, J. Liu, M. Luo, Z. Li, L. Yao, and A. Hauptmann, "Tn-zstad: Transferable network for zero-shot temporal activity detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3848–3861, 2022.

[4] X. Fu, J. Xiao, Y. Zhu, A. Liu, F. Wu, and Z.-J. Zha, "Continual image deraining with hypergraph convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9534–9551, 2023.

[5] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1–26, 2021.

[6] C. Yan, X. Chang, Z. Li, W. Guan, Z. Ge, L. Zhu, and Q. Zheng, "Zeronas: Differentiable generative adversarial networks search for zero-shot learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9733–9740, 2021.

[7] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.

[8] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 655–12 663.

[9] J. Guo, M. Wang, Y. Zhou, B. Song, Y. Chi, W. Fan, and J. Chang, "Hgan: Hierarchical graph alignment network for image-text retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 9189–9202, 2023.

[10] Z. Wang, Z. Gao, K. Guo, Y. Yang, X. Wang, and H. T. Shen, "Multilateral semantic relations modeling for image text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2830–2839.

[11] X. Qin, L. Li, G. Pang, and F. Hao, "Heterogeneous graph fusion network for cross-modal image-text retrieval," *Expert Systems with Applications*, vol. 249, p. 123842, 2024.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[14] S. Zubair, F. Yan, and W. Wang, "Dictionary learning based sparse coefficients for audio classification with max and average pooling," *Digital Signal Processing*, vol. 23, no. 3, pp. 960–970, 2013.

[15] Z. Fu, Z. Mao, Y. Song, and Y. Zhang, "Learning semantic relationship among instances for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 159–15 168.

[16] Z. Pan, F. Wu, and B. Zhang, "Fine-grained image-text matching by cross-modal hard aligning network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 275–19 284.

[17] Z. Li, L. Zhang, K. Zhang, Y. Zhang, and Z. Mao, "Improving image-text matching with bidirectional consistency of cross-modal alignment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6590–6607, 2024.

[18] Z. Wang, Z. Gao, M. Han, Y. Yang, and H. T. Shen, "Estimating the semantics via sector embedding for image-text retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 10 342–10 353, 2024.

[19] G. Xiong, M. Meng, T. Zhang, D. Zhang, and Y. Zhang, "Reference-aware adaptive network for image-text matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9678–9691, 2024.

[20] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 661–15 670.