SAMEF-DHNN: Sample-Aware Multimodal Expert Fusion and Dynamic Hard Negatives Network for Multimodal Knowledge Graph Completion

1st Yi Zhou
School of Computer Science and Technology
Southwest University of Science and Technology
Mianyang, China
553580053@qq.com

2nd Qingming Zhang*
School of Computer Science and Technology
Southwest University of Science and Technology
Mianyang, China
qmzhang@swust.edu.cn

***Abstract*—Multimodal Knowledge Graph Completion (MMKGC) leverages multimodal data such as images and texts to predict missing triples. Two major challenges are static multimodal fusion strategies overlook differences in modality reliability across samples, and contrastive learning balancing negative sample quality with computational efficiency. To address these issues, we propose a Sample-Aware Multimodal Expert Fusion and Dynamic Hard Negatives Network (SAMEF-DHNN). First, to address the limitation of static fusion, we propose the Sample-Aware Multimodal Expert Fusion Block. Inspired by Mixture of Experts (MoE), it treats each modality as an expert and employs a novel gating network to dynamically assign adaptive fusion weights for each input entity sample, thereby enhancing fusion robustness against variable modality quality. Second, we introduce the Dynamic Hard Negative Sampling Contrastive Learning (DHNSCL). Instead of static or pre-generated negatives, Its core is a real-time mining strategy that, for each anchor, selects only the top-K most semantically similar samples in the current batch as hard negatives. This improves discriminative ability while maintaining efficiency. Experiments on DB15K and MKG-W demonstrate that SAMEF-DHNN outperforms 16 existing baselines across multiple metrics, validating its effectiveness.**

***Index Terms*—knowledge graph, knowledge graph completion, multi-modal knowledge graph completion, contrastive learning, link prediction**

I. Introduction

Multimodal Knowledge Graphs (MMKGs) [1], [2] represent knowledge as structured triples (head, relation, tail) augmented with multimodal data such as images and texts. They serve as a critical knowledge base for applications like recommendation systems [3] and multimodal understanding [4]. However, MMKGs are often incomplete, necessitating the task of Multimodal Knowledge Graph Completion (MMKGC) to predict missing triples [1], [2].

In recent years, significant progress has been made in MMKGC research. Early works such as IKRL [5] first integrated visual information into the TransE [6] framework, enriching entity representations. Subsequently, TBKGC [7] established a joint modeling mechanism for text and images, while RSME [8] further advanced multimodal fusion by screening key visual information through relevance scoring. In recent years, VBKGC [9] leveraged VisualBERT to enhance modal alignment and fusion capabilities, MMKGR [10] introduced gated attention mechanisms to enable multi-hop reasoning, and MoSE [11] improved representation quality through splitting and fusion processing of multimodal data. These methods have progressively refined the technical roadmap for multimodal fusion, laying the foundation for more accurate knowledge reasoning.

In the domain of contrastive learning and negative sampling, traditional approaches typically employ random sampling or head/tail entity replacement to construct negative samples. To enhance sample quality and discriminative capability, generative methods such as IGAN [12], KBGAN [13], and MMKRL [14] introduce adversarial training strategies, synthesizing hard negative samples through generators. DHNS [15] further leverages diffusion models to generate multimodal negative sample embeddings. On the other hand, non-generative methods like NSCaching [16] improve hard sample mining efficiency by caching high-scoring samples, while VBKGC [9] and MANS [17] propose fine-grained negative sampling mechanisms based on visual-structural alignment. These works improve the quality and training efficiency of negative samples. Despite the considerable advancements in modal fusion and negative sampling achieved by existing methods, we identify two limitations: current fusion strategies often use fixed weighting, lacking adaptability to sample-specific modality quality, and negative sampling methods either incur high cost or lack dynamic hardness adjustment. To address these challenges, we propose the Sample-Aware Multimodal Expert Fusion and Dynamic Hard Negatives Network (SAMEF-DHNN). First, we introduce the Sample-Aware Multimodal Expert Fusion Block (SAMEF Block) as a key innovation for adaptive fusion. It moves beyond fixed weighting schemes through a trainable gating network. This network evaluates the reliability of each modality expert per sample and computes sample-specific fusion weights. This enables the model to

*Corresponding author. This work was supported by the Sichuan Province Science and Technology Achievement Transfer and Transformation Demonstration Project, China (Grant No. 2024ZHCG0002)

flexibly emphasize more trustworthy modalities for each entity. Second, we introduce a DHNSCL module. Its core is a real-time mining strategy that, for each anchor, selects only the top-K most semantically similar samples in the current batch as hard negatives. This dynamic and efficient approach ensures the model continuously focuses on the most challenging discriminative boundaries.

1) We propose the SAMEF Block that achieves input-conditional adaptive multimodal fusion, enhancing the model's robustness to noisy modalities and its efficiency in leveraging informative modal features.
2) We introduce the DHNSCL that efficiently mines a small number of easily confusable hard samples through semantic similarity computation, improving the discrimination of fine-grained semantic differences in contrastive learning.
3) Experiments on multiple public MMKGC datasets demonstrate that the proposed method outperforms 16 existing baseline models across multiple metrics.

II. Methodology

A. Task Definition

MMKG integrates visual and textual knowledge, defined as $G = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{V}, \mathcal{D})$. Here $\mathcal{E}$ and $\mathcal{R}$ denote entity and relation sets, while $\mathcal{T} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ represents triples. $\mathcal{V}$ and $\mathcal{D}$ correspond to visual images and textual descriptions of entities. MMKGC utilizes multimodal information $(e, V(e), D(e))$ to establish a scoring function $\mathcal{S}(h, r, t)$ that evaluates triple plausibility, maximizing confidence for positive triples [6].

B. Base Framework

As shown in Fig. 1, our model has two main parts: (1) the Hierarchical Triple Modeling Module (HTM Module), containing a Cross-Modal Encoding Fusion Module (CMEF) and a Contextual Triple Encoder and Relation Decoder (CTE & RD); (2) the DHNSCL Module.

C. Hierarchical Triple Modeling Module

1) Cross-Modal Encoding Fusion Module: To extract features from text and image modalities, we use a pre-trained BERT model [18] for text subword embeddings and BEIT [19] for image patch encoding, obtaining entity, visual, and text representations: $x_{\text{ent}}, x_{\text{vis}}, x_{\text{txt}} \in \mathbb{R}^{N_{\text{ent}} \times L_m \times D}$. These are concatenated along the sequence dimension to form a unified multimodal feature matrix:

$$X_{\text{ent}} = [x_{\text{ent}} \oplus x_{\text{vis}} \oplus x_{\text{txt}}] \in \mathbb{R}^{N \times (L_{\text{ent}} + L_{\text{vis}} + L_{\text{txt}}) \times D} \quad (1)$$

Where $\oplus$ denotes the concatenation operation. A standard Transformer Encoder ($TE(\cdot)$) then models cross-modal interactions:

$$H = \text{TE}(X_{\text{ent}}) \in \mathbb{R}^{N \times L \times D} \quad (2)$$

We apply pooling to H to obtain global semantics $h_{\text{pool}} \in \mathbb{R}^{N \times D}$. Meanwhile, the SAMEF Block produces cross-modal enhanced representations $f_{\text{sum}} \in \mathbb{R}^{N \times D}$. Both are adaptively fused via a gating unit:

$$\gamma = \sigma(W_g[h_{\text{pool}} \oplus f_{\text{sum}}] + b_g) \in [0, 1]^{N \times D} \quad (3)$$

$$e_{\text{ent}} = \text{LN}\big(\gamma \odot h_{\text{pool}} + (1 - \gamma) \odot f_{\text{sum}}\big) \in \mathbb{R}^{N \times D} \quad (4)$$

Here, γ is a gating weight computed using $W_g \in \mathbb{R}^{2D \times D}$ and $b_g \in \mathbb{R}^D$,LN$(\cdot)$ denotes Layer Normalization,and $\sigma(\cdot)$ is the sigmoid activation function. The final entity embedding e_{ent} balances both information sources.

2) Context Triple Encoder & Relation Decoder: To enhance relational context interactions, we employ a Contextual Triple Encoder(CTE). Given a set of triples $X = \{(h, r, t)_i\}_{i=1}^N$, the model first maps each element to embeddings: head entity embedding $\mathbf{e}_h$, relation embedding $\mathbf{e}_r$, and tail entity embedding $\mathbf{e}_t$, looked up from embedding matrices. These are combined with positional biases $\mathbf{p}_{\text{bias}}$ to form sequential representations $\mathbf{h}_{\text{seq}}, \mathbf{r}_{\text{seq}}, \mathbf{t}_{\text{seq}} \in \mathbb{R}^D$.

A Relation Decoder (RD$(\cdot)$), based on the Transformer Decoder, processes the concatenated sequence $[\mathbf{h}_{\text{seq}} \oplus \mathbf{r}_{\text{seq}} \oplus \mathbf{t}_{\text{seq}}]$ to produce contextualized representations:

$$\mathbf{H}_{\text{dec}} = \text{RD}([\mathbf{h}_{\text{seq}} \oplus \mathbf{r}_{\text{seq}} \oplus \mathbf{t}_{\text{seq}}]) \in \mathbb{R}^{3 \times D} \quad (5)$$

We take the first, second, and third vector from $\mathbf{H}_{\text{dec}}$ as $\widetilde{\mathbf{h}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{t}}$ respectively. The triplet plausibility is scored using the Tucker [20] Score (TS) function:

$$\mathcal{S}(h, r, t) = \mathcal{W} \times_1 \widetilde{\mathbf{h}} \times_2 \widetilde{\mathbf{r}} \times_3 \widetilde{\mathbf{t}} \quad (6)$$

where $\mathcal{W}$ is a learnable core tensor, and $\times_i$ denotes the i-mode product between a tensor and a vector. The model is trained with cross-entropy loss:

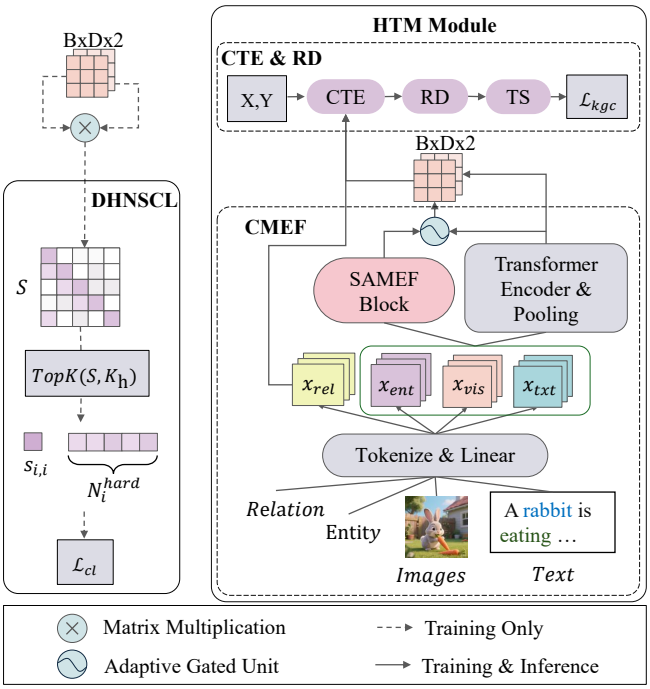

Fig. 1. The overview of SAMEF-DHNN base framework.

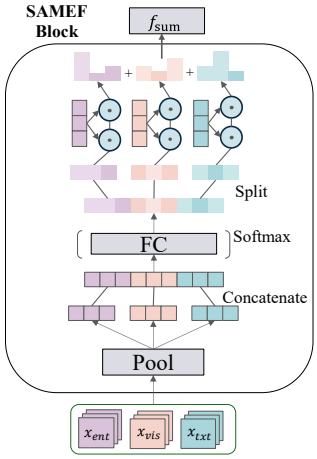

Fig. 2. The details of SAMEF Block.

$$L_{\text{kgc}} = \sum_{(h,r,t)\in\mathcal{T}} \log \frac{\exp(\mathcal{S}(h,r,t))}{\sum_{t'\in\mathcal{E}} \exp(\mathcal{S}(h,r,t'))} \tag{7}$$

D. Sample-Aware Multimodal Expert Fusion Block

The SAMEF Block proposed in this paper is inspired by MoE. It treats each modality $(V(e), D(e))$ as an independent expert, regards entity e as an input sample, and employs a gating network to assign adaptive fusion weights to each modality expert, thereby dynamically computing and integrating more reliable modal signals for each input sample.

As shown in Fig.2, SAMEF Block first obtains the most representative modality-level summary e_m for each modality by applying average pooling to entity features x_{ent}, visual features x_{vis}, and textual features x_{txt} respectively:

$$y_m = \frac{1}{T}\sum_{t=1}^{T} e_m^{(t)}, m \in \{ent, vis, txt\} \tag{8}$$

To establish inter-modal correlations, we concatenate these expert opinions along the feature dimension to obtain $e_{com} \in \mathbb{R}^{N\times 3D}$, which subsequently serves as input to the gating network:

$$e_{com} = [y_{ent} \oplus y_{vis} \oplus y_{txt}] \tag{9}$$

The core innovation enabling sample-aware capability lies in a learnable gating network that takes the joint opinions of different experts as input, evaluates the authority of each expert based on the current input sample, and generates learnable dynamic weights. First, to obtain the raw logits $\mathbf{g} = [g_{ent} \oplus g_{vis} \oplus g_{txt}] \in \mathbb{R}^{N\times 3}$ for the gating network, e_{com} is fed into a Fully Connected (FC) layer to learn a mapping from the joint feature space to the weight simplex:

$$\mathbf{g} = e_{com}W_c + b_c, \mathbf{g} \in \mathbb{R}^{N\times 3} \tag{10}$$

where W_c and b_c are trainable parameters. Then, to obtain sample-aware fusion weights $\alpha = [\alpha_{ent}, \alpha_{vis}, \alpha_{txt}], \alpha \in \mathbb{R}^{N\times 3}$, we apply Softmax normalization:

TABLE I
STATISTICAL INFORMATION OF THE BENCHMARKS.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#Vis	#Text	Train	#Valid	#Test
DB15K	12842	279	12818	12842	79222	9902	9904
MKG-W	15000	169	14463	14123	34196	4276	4274

$$\alpha = \text{softmax}(\mathbf{g}) = \left[\frac{e^{g_{\text{ent}}}}{\sum_m e^{g_m}}, \frac{e^{g_{\text{vis}}}}{\sum_m e^{g_m}}, \frac{e^{g_{\text{txt}}}}{\sum_m e^{g_m}}\right], \quad m \in \{\text{ent}, \text{vis}, \text{txt}\} \tag{11}$$

The Softmax ensures weight normalization ($\alpha_{ent} + \alpha_{vis} + \alpha_{txt} = 1$) and non-negativity, making α a valid convex combination coefficient. The weights α can be interpreted as the probability that each expert is the most reliable "answer" in the current sample context.

Finally, the fused feature f_{sum} is obtained through a convex combination of the three expert opinions. This paper adopts a quadratic interaction weighted fusion strategy using element-wise multiplication:

$$f_{\text{sum}} = \sum_{m\in\{\text{ent},\text{vis},\text{txt}\}} (\alpha_m \odot y_m) \odot y_m \tag{12}$$

E. Dynamic Hard Negative Sampling Contrastive Learning Module

To enhance the model's capability for fine-grained semantic discrimination of entity information, this paper introduces DHNSCL. By dynamically selecting the most challenging negative samples for each anchor sample, the model concentrates on a small subset of hard negative samples, thereby learning discriminative representations more efficiently and improving its discrimination ability.

First, given embedding matrices $E_1, E_2 \in \mathbb{R}^{B\times D}$ with batch size B and embedding dimension D, where E_1 is derived from e_{ent} in (4) and E_2 is obtained from (12), we construct a similarity matrix S:

$$S = \frac{E_1 \cdot E_2^{\top}}{\tau} \in \mathbb{R}^{B\times B} \tag{13}$$

where τ is a temperature hyperparameter that modulates the similarity distribution. For each anchor sample i, we retain the positive similarity score $s_{i,i} \in S$. Based on this, we perform hard negative sample mining for each anchor i by selecting the top-K most similar samples as hard negatives:

$$N_i^{\text{hard}} = \text{topK}\left(\{s_{i,j} \mid j \neq i\}, K_{\text{h}}\right) \tag{14}$$

Here, the Hard Negative Sample Count K_{h} is a hyperparameter significantly smaller than the batch size B, controlling the number of mined hard negatives. Finally, we employ the standard cross-entropy loss as the objective function:

$$\mathcal{L}_{\text{cl}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s_{i,i})}{\exp(s_{i,i}) + \sum_{n\in N_i^{\text{hard}}} \exp(s_{i,n})} \tag{15}$$

TABLE II
THE MAIN MMKGC RESULTS. THE BEST RESULTS ARE MARKED AS **BOLD** AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Model	DB15K				MKG-W			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.2	44.23
ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73
RotatE	29.28	17.87	36.12	49.66	33.67	26.8	36.68	46.73
TuckER	33.86	25.33	37.91	50.38	30.39	24.44	32.91	41.25
KBGAN(TransE)	25.73	9.91	36.95	51.93	29.47	22.21	34.87	40.64
KBGAN(TransD)	23.74	9.34	33.51	47.94	29.67	22.38	35.24	40.8
IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07
TBKGC	28.4	15.61	37.03	49.86	31.48	25.31	33.98	43.24
MMKRL	26.81	13.85	35.07	39.39	30.1	22.16	34.09	44.69
RSME	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43
VBKGC	30.61	19.75	37.18	39.44	30.61	24.91	33.01	40.88
VISTA	30.42	22.39	33.56	45.94	32.91	26.12	35.38	45.61
AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	**47.21**
MANS	28.82	16.87	36.58	39.26	30.88	24.89	33.63	41.78
MyGO	37.57	30.15	41.04	51.95	35.17	29.81	36.94	45.51
DHNS	34.36	23.34	41.56	53.72	35.68	28.98	38.68	48.12
Ours	**39.33**	**31.75**	**42.85**	**53.69**	**36.06**	**30.53**	**38.04**	46.79

The overall training objective of the framework is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{kgc}} + \lambda \cdot \mathcal{L}_{\text{cl}}$, where λ is the Contrastive Loss Weight.

1) Complexity Analysis: We analyzes the complexity of the loss function in (15) and compares it with the traditional InfoNCE loss (16). Let the batch size be N, and the number of mined hard negative samples be K ($K \ll N$). In (16), it is necessary to compute over all negative samples in the batch:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s_i \cdot s_i\prime/\tau)}{\sum_{j=1}^{N}\exp(s_i \cdot s_j\prime/\tau)} \tag{16}$$

The complexity of (16) is $O(N^2)$, which limits training efficiency and scalability. The proposed loss function in (15) significantly reduces the computational time complexity from $O(N^2)$ to $O(N \times K)$. Since $K \ll N$, it greatly saves computational and memory overhead in large-batch training scenarios, making model training more efficient.

III. EXPERIMENTS

A. Experiment Settings

1) Datasets: We evaluated our model on two public MMKGC benchmarks: DB15K [21] and MKG-W [22]. Dataset statistics are shown in Table I.

2) Task and Protocols: To systematically evaluate the capability of the model, we conducted link prediction tasks [6] on both datasets. We evaluate the model through link prediction tasks [6] on both datasets. To ensure fair comparison, we report MRR and Hit@K (K=1,10) [15], [23] as the most discriminative metrics. Hit@1 reflects precision and Hit@10 evaluates recall, while Hit@3 is omitted due to its correlation with MRR. The calculations are as follows:

$$MRR = \frac{1}{|\mathcal{T}_{test}|}\sum_{i=1}^{|\mathcal{T}_{test}|}\left(\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}}\right) \tag{17}$$

$$Hit@K = \frac{1}{|\mathcal{T}_{test}|}\sum_{i=1}^{|\mathcal{T}_{test}|}\left(1\left(r_{h,i} \le K\right) + 1\left(r_{t,i} \le K\right)\right) \tag{18}$$

where $r_{h,i}, r_{t,i}$ represent the prediction ranks for head entities and tail entities respectively.

3) Baselines: This paper conducts a comprehensive performance evaluation, comparing 16 different baseline models in the experiments. (TransE [6],RotatE [23],TuckER [20],KBGAN [13], IKRL [5], TBKGC [7], MMKRL [14], RSME [8], VBKGC [9], VISTA [24], AdaMF [25], MANS [17], MyGO [26],DHNS [15])

4) Implementation Details: We implemented our proposed model in PyTorch and utilized an RTX 4090 GPU. For multimodal features, images adopt BEIT [19] (codebook size=8192) and text adopts BERT [18] (vocabulary size=32,000). Training utilizes the Adam optimizer [27] for a total of 2000 epochs, with a batch size of 2048 and an embedding dimension of 256. Visual token length is 8, text token length is 12, Loss Weight λ is tuned in $\{0.05, 0.1, 0.5, 1.0, 5.0\}$, and The Number of Hard Negative Samples K_{h} is tuned in $\{2, 4, 5, 6, 8, 16, 32, 64\}$.

B. Main Results

The main experimental results in Table II show that our model achieves the highest or second-highest scores across

TABLE III
THE ABLATION STUDY RESULTS ON DB15K AND MKG-W.

Settings	DB15K			MKG-W		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
w/o SAMEF	38.81	31.09	53.40	35.85	30.21	46.70
w/o $\mathcal{L}_{cl}$	38.98	31.42	53.28	35.39	29.83	45.77
w/o SAMEF+$\mathcal{L}_{cl}$	38.69	31.13	53.18	35.44	30.11	45.69
Full Model	39.33	31.75	53.69	36.17	30.39	47.16

all metrics on both datasets, significantly outperforming 16 existing baselines.

Specifically, on DB15K, SAMEF-DHNN improves MRR by 4.68% and Hit@1 by 5.31% over the second-best model, DHNS. On MKG-W, it also achieves gains of 1.37% in MRR and 1.95% in Hit@1. These improvements, especially in Hit@1, confirm the model's strength in precise matching.

Further analysis reveals that the performance advantage stems mainly from the proposed SAMEF Block and DHNSCL. Unlike methods that simply concatenate multimodal features, our model dynamically computes fusion weights to reduce noise interference and capture discriminative features more accurately. Moreover, the negative sampling strategy enhances fine-grained discrimination more effectively than random or generated sample approaches.

C. Ablation Study

1) Model Design: We conducted systematic ablation studies on DB15K and MKG-W to evaluate the contribution of each module. As shown in Table III, all ablated variants underperform the full model, confirming that both SAMEF Block and DHNSCL play important roles.

Removing the SAMEF Block resulted in an average decrease of 1.10% in MRR and 1.34% in Hit@1, demonstrating that its multimodal fusion conditioned on input entity samples is crucial for accurate ranking. Removing the DHNSCL module led to an average reduction of 1.53% in MRR and 1.44% in Hit@1, highlighting its significant effect on enhancing the discriminative capability of the system. Moreover, removing DHNSCL caused a greater decline in Hit@10 (1.86%) compared to SAMEF Block removal (0.76%), demonstrating its stronger impact on discrimination and coverage in longer recommendation lists.

Joint removal of both modules resulted in performance degradation of 0.95%–3.12% across all metrics, exceeding the loss from individual removal. This synergistic effect confirms that SAMEF Block provides reliable multimodal representations to facilitate dynamic hard negative sampling in DHNSCL, while the discriminative features from DHNSCL in turn refine the multimodal fusion in SAMEF Block. Their collaboration collectively improves prediction accuracy in MMKGC.

D. Hyperparameter Analysis

We systematically analyzed the effects of the Contrastive Loss Weight λ and the Hard Negative Sample Count K_{h} on model performance.

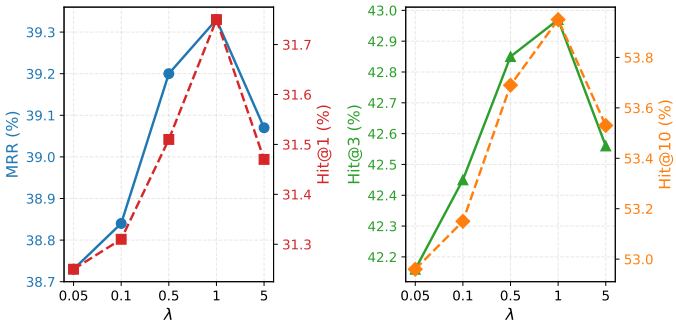

Fig. 3. Trends of MRR and Hit@K (k=1,3,10) across different λ values.

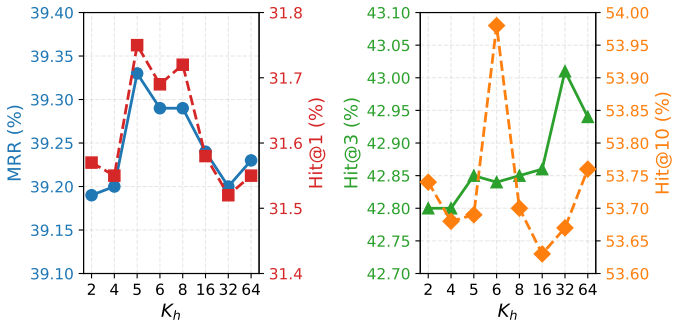

Fig. 4. Trends of MRR and Hit@K (k=1,3,10) across different K_{h} values.

1) Contrastive Loss Weight λ: The experimental results are shown in Fig. 3 (with $K_{\text{h}} = 5$ by default). As λ varies within the tested range, the model achieves optimal MRR and Hit@1 performance when $\lambda = 1.0$. This indicates that a higher λ enhances the model's ability to discriminate fine-grained semantic differences, thereby significantly improving Hit@1 accuracy. Simultaneously, a moderate weight balances the contributions of the main task and contrastive learning, further improving the global ranking quality of candidate entities and the overall robustness of the model.

2) Hard Negative Sample Count K_h: Fig. 4 shows the impact of different K_{h} values with $\lambda = 1.0$. When $K_{\text{h}} = 5$, MRR and Hit@1 reach their highest values, indicating that an appropriate number of high-difficulty negative samples helps the model focus on discriminative semantic boundaries, thereby improving precise matching capability. When $K_{\text{h}} = 6$, Hit@10 peaks, suggesting that a moderate negative sample difficulty expands the model's semantic discrimination boundary for sub-optimal candidates while maintaining core discriminative power. When $K_{\text{h}} = 32$, Hit@3 performs best, indicating that an expanded negative sample space forces the model to learn more global ranking robustness, enhancing the screening stability for mid-ranked candidates (Hit@3).

IV. CONCLUSION

In this paper, we propose SAMEF-DHNN, a MMKGC model integrating Sample-Aware Multimodal Expert Fusion

and Hard Negative Sampling. To address the static weighting issue in multimodal information fusion for MMKGC, we design the SAMEF Block, which introduces a dynamic weighting mechanism enabling adaptive evaluation and fusion of multimodal information based on input entity samples. Furthermore, we propose a dynamic hard negative sample mining strategy that enhances contrastive learning efficiency and effectiveness by real-time sampling of a limited number of hard negatives with high semantic similarity to target entities. Experimental results on two datasets demonstrate that SAMEF-DHNN significantly outperforms 16 existing baseline models, particularly showing notable improvements in precise prediction metrics such as Hit@1 and MRR, validating the effectiveness of the proposed modules. SAMEF-DHNN provides an efficient and robust solution for MMKGC, offering new perspectives and a reference framework for future research in this field.

REFERENCES

[1] Z. Chen, Y. Zhang, Y. Fang, Y. Geng, L. Guo, X. Chen, Q. Li, W. Zhang, J. Chen, Y. Zhu, J. Li, X. Liu, J. Z. Pan, N. Zhang, and H. Chen, "Knowledge graphs meet multi-modal learning: A comprehensive survey," *CoRR*, vol. abs/2402.05391, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.05391

[2] K. Liang, L. Meng, M. Liu, Y. Liu, W. Tu, S. Wang, S. Zhou, X. Liu, F. Sun, and K. He, "A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9456–9478, 2024. [Online]. Available: https://doi.org/10.1109/TPAMI.2024.3417451

[3] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:224281034

[4] Y. Zhu, H. Zhao, W. Zhang, G. Ye, H. Chen, N. Zhang, and H. Chen, "Knowledge perceived multi-modal pretraining in e-commerce," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. ACM, Oct. 2021, p. 2744–2752. [Online]. Available: http://dx.doi.org/10.1145/3474085.3475648

[5] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *International Joint Conference on Artificial Intelligence*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:9909815

[6] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Neural Information Processing Systems*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:14941970

[7] H. M. Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *International Workshop on Semantic Evaluation*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:44145776

[8] M. Wang, S. Wang, H. Yang, Z. Zhang, X. Chen, and G. Qi, "Is visual context really helpful for knowledge graph? a representation learning perspective," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. ACM, Oct. 2021, p. 2735–2743. [Online]. Available: http://dx.doi.org/10.1145/3474085.3475470

[9] Y. Zhang and W. Zhang, "Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling," *CoRR*, vol. abs/2209.07084, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.07084

[10] S. Zheng, W. Wang, J. Qu, H. Yin, W. Chen, and L. Zhao, "Mmkgr: Multi-hop multi-modal knowledge graph reasoning," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, Apr. 2023, pp. 96–109. [Online]. Available: http://dx.doi.org/10.1109/icde55515.2023.00015

[11] Y. Zhao, X. Cai, Y. Wu, H. Zhang, Y. Zhang, G. Zhao, and N. Jiang, "Mose: Modality split and ensemble for multimodal knowledge graph completion," in *Conference on Empirical Methods in Natural Language Processing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252918783

[12] P. Wang, S. Li, and R. Pan, "Incorporating gan for negative sampling in knowledge representation learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: http://dx.doi.org/10.1609/aaai.v32i1.11536

[13] L. Cai and W. Y. Wang, "Kbgan: Adversarial learning for knowledge graph embeddings," *ArXiv*, vol. abs/1711.04071, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3401524

[14] X. Lu, L. Wang, Z. Jiang, S. He, and S. Liu, "Mmkrl: A robust embedding approach for multi-modal knowledge graph representation learning," *Applied Intelligence*, vol. 52, no. 7, pp. 7480–7497, Sep. 2021. [Online]. Available: http://dx.doi.org/10.1007/s10489-021-02693-9

[15] G. Niu and X. Zhang, "Diffusion-based hierarchical negative sampling for multimodal knowledge graph completion," 2025. [Online]. Available: https://arxiv.org/abs/2501.15393

[16] Y. Zhang, Q. Yao, Y. Shao, and L. Chen, "Nscaching: Simple and efficient negative sampling for knowledge graph embedding," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, Apr. 2019, pp. 614–625. [Online]. Available: http://dx.doi.org/10.1109/icde.2019.00061

[17] Y. Zhang, M. Chen, and W. Zhang, "Modality-aware negative sampling for multi-modal knowledge graph embedding," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jun. 2023, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ijcnn54540.2023.10191314

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[19] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," 2022. [Online]. Available: https://arxiv.org/abs/2208.06366

[20] I. Balazevic, C. Allen, and T. M. Hospedales, "Tucker: Tensor factorization for knowledge graph completion," *ArXiv*, vol. abs/1901.09590, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59316623

[21] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, and D. S. Rosenblum, "Mmkg: Multi-modal knowledge graphs," *ArXiv*, vol. abs/1903.05485, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:76663467

[22] D. Xu, T. Xu, S. Wu, J. Zhou, and E. Chen, "Relation-enhanced negative sampling for multimodal knowledge graph completion," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. ACM, Oct. 2022, p. 3857–3866. [Online]. Available: http://dx.doi.org/10.1145/3503161.3548388

[23] X. Zhou, Y. Yi, and G. Jia, "Path-rotate: Knowledge graph embedding by relational rotation of path in complex space," in *2021 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, Jul. 2021, pp. 905–910. [Online]. Available: http://dx.doi.org/10.1109/iccc52777.2021.9580273

[24] J. Lee, C. Chung, H. Lee, S. Jo, and J. Whang, "Vista: Visual-textual knowledge graph representation learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023. [Online]. Available: http://dx.doi.org/10.18653/v1/2023.findings-emnlp.488

[25] Y. Zhang, Z. Chen, L. Liang, H. zeng Chen, and W. Zhang, "Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion," *ArXiv*, vol. abs/2402.15444, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267898006

[26] Y. Zhang, Z. Chen, L. Guo, Y. Xu, B. Hu, Z. Liu, W. Zhang, and H. Chen, "Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, pp. 13 322–13 330, Apr. 2025. [Online]. Available: http://dx.doi.org/10.1609/aaai.v39i12.33454

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980