Multi-tendency negative samples for multi-modal knowledge graph completion based on diffusion*

Xiaoceng Zeng
College of Computer Science and Technology
Xinjiang University
Urumqi, China
2754741263@qq.com

Hui Zhao*
* *College of Computer Science and Technology*
Xinjiang University
Urumqi, China
zhaohui@xju.edu.cn*

***Abstract*—Multimodal Knowledge Graph Completion (MMKGC) seeks to automatically uncover unseen true knowledge from multimodal knowledge graphs by jointly modeling entity triplets and multimodal data. However, real-world MMKGs face two key challenges: modality diversityand modality missingness.To tackle these challenges, this paper proposes a Multimodal Knowledge Graph Completion framework based on Relation-Guided Adaptive Fusion and Different Modality Tendency Negative Sampling (MTNS), featuring two core innovations: 1) adaptive fusion of any modalities via dynamic weight allocation; 2) a diffusion model-driven negative sampling strategy that generates diverse negative samples), effectively mitigating modality missingness issues.Empirical results against SOTASOTA baselines confirm MTNS's superiority.**

***Index Terms*—Multimodal Knowledge Graph, Knowledge Graph Completion, Negative Sampling.**

I. Introduction

Multimodal Knowledge Graphs [1], [2] (MMKGs), as an extended form of traditional knowledge graphs, construct a multi-dimensional feature space among entities by semantically associating multimodal features such as text and images with structured triplets (head entity, relation, tail entity). This fused representation paradigm provides more reliable knowledge support for downstream tasks such as recommendation systems, visual question answering, and large language models. However, due to the difficulty in collecting multimodal corpora, existing multimodal knowledge graphs often face the issue of incompleteness. Therefore, Knowledge Graph Completion (KGC) technology, which discovers potential facts by mining latent associations in the graph, has become an important research direction in the field of knowledge graphs.

Traditional knowledge graph completion methods typically learn structural embeddings to model triplets and evaluate their validity. However, for multimodal knowledge graphs, relying solely on structural embeddings is insufficient to capture the complex relationships and interactions between modalities. It is necessary to effectively integrate information from different modalities to achieve more accurate predictions of missing entities. In existing research on multimodal knowledge graph completion, multimodal information of entities is usually used as supplementary data for encoding entity representations. Pre-trained models are employed to extract features from different modalities. Then, the encoded entity representations are concatenated with relation representations and fed into a triplet encoder for prediction. Nevertheless, these methods still face the problems of entity modality diversity and modality missingness.

Modality Diversity: In the real world, Multimodal Knowledge Graphs (MMKGs) encompass a variety of heterogeneous modalities, such as structured data, text, images, and audio. The semantic representations and structural features of different modalities exhibit significant differences, and they contribute differently during the completion process. Existing methods overly rely on visual and textual modalities and adopt simple feature concatenation or dot-product fusion strategies. This leads to insufficient model generalization ability, making it difficult to adapt to complex and diverse modalities. Moreover, the semantic gap between heterogeneous modalities further exacerbates the difficulty of feature fusion, limiting the improvement of completion performance.Modality Missingness: During the actual completion process of MMKGs, there are significant differences in the acquisition difficulty and sparsity of different modalities, resulting in the frequent absence of key modality information. Existing research has not paid sufficient attention to this issue and merely relies on complete modality data for completion. Additionally, current negative sampling strategies, such as random sampling and adversarial-based sampling, mainly generate negative triplets based on structural features. They neglect multimodal semantic information, which easily leads to the generation of simple or ineffective negative samples. Furthermore, these strategies are overly dependent on pre-sampled triplets.

To address the aforementioned issues, we propose a Multimodal Knowledge Graph Completion model based on Diffusion Model for generating Multi-Tendency Negative Samples (MTNS). It can fuse diverse entity modalities through an adaptive fusion module and generate negative samples with various tendencies using a diffusion model to combat modality missingness. MTNS consists of two key modules: the adaptive fusion module and the negative sampling module. The fusion module facilitates diverse multi-modal fusion with any input

This article is funded by the Key Research and Development Program of Xinjiang Uygur Autonomous Region (Project Number: 2023B01032).
*(Corresponding author: Hui Zhao)

modality and adjusts the weight of each modality through relation guidance. The negative sampling module designs a conditional diffusion model for generating negative samples. By replacing specific modalities of positive samples and dynamically adjusting the time steps, it generates negative samples with different tendencies and difficulties, thereby enhancing the missing modality information. Our specific contributions are mainly as follows:

1) We propose a novel multi - modal knowledge graph completion model, MTNS. This model obtains enhanced modal semantics through relation - guided modal fusion, addressing the ambiguity of single - modal entities. Moreover, by generating multi - tendency negative samples and simulating missing scenarios, it improves the model's robustness.
2) We present an effective negative sample generation module. By dynamically controlling the difficulty of negative samples and generating negative samples with different tendencies, it enhances the model's stability in the face of modality missingness.
3) We conduct extensive experiments on two public datasets against multiple models, demonstrating the effectiveness of our proposed model.

II. Related Work

Multimodal knowledge graphs significantly enhance the richness of knowledge representation and reasoning capabilities by integrating data from multiple different modalities, such as structural information, textual information, image information, and audio information. This provides strong support for a more comprehensive and accurate understanding and processing of knowledge. Existing research methods mainly make improvements from the following three key aspects: multimodal fusion, comprehensive decision - making, and negative sampling:(1) Multimodal fusion [6]–[8]. The key to multimodal knowledge graph completion lies in effectively fusing multisource information from text, vision, and structure. To this end, researchers have proposed various fusion strategies to achieve semantic complementarity among modalities. (2) Comprehensive decisionmaking [11], [12]. This category of methods focuses on the integration of multimodal information at the decision-making stage, aiming to enhance the accuracy and stability of the model through joint prediction.(3) Negative sampling optimization [14]–[16]. Negative sampling is a crucial step in the training of Multimodal Knowledge Graph Completion (MMKGC) models, and its design directly affects the model's discriminative ability and generalization performance. In response to the complex characteristics of multimodal data, some studies have introduced improved sampling mechanisms.

III. Method

In this section, we offer an in - depth introduction to our MTNS model. As depicted in Figure 1, it comprises two main modules designed for cross - modal entity fusion and accuracy improvement under modal missingness. First, entity modal information, combined with relations, is used to calculate weights, resulting in entity embeddings that incorporate multi - modal data. In the negative sampling module, conditional diffusion generates realistic negative samples. By swapping modalities of positive samples, we create negative samples with varied tendencies to mimic real - world modal missing situations. Moreover, we dynamically adjust negative sample difficulty to enhance model stability.

A. Relation-Guided Modal Fusion

To utilize modal information, we first perform modal encoding to capture modal features, which is a very common step in different MMKGC methods. We use BERT as the text encoder and VGG16 as the image encoder. Digital information can also be extended into sequences, and BERT is adopted to capture its modal information. The pre-trained encoders are frozen to capture modal features and will not be fine-tuned during the training process. The modal encoding process produces different modal embeddings e_m for each entity e.In the context of Multi-Modal Knowledge Graph (MMKG), embeddings from other modalities serve as auxiliary information to enhance the structure of entities. To uniformly represent multi-modal information, we designate the structural modality as S and the structural embedding e as e_S, which are learnable parameters.

After obtaining the modal embeddings of entities, in order to extract the rich semantic information contained within the entities, it is common to fuse these multi-modal embeddings. Existing methods typically employ fusion mechanisms such as concatenation, dot-product, or gating to achieve modal fusion. However, these methods are mainly designed for specific modalities (e.g., text and images), and may not fully serve scenarios where a richer combination of modalities is intertwined. Moreover, given the diversity of entity modalities, different modalities need to play different roles and provide different evidence for robust predictions in different contexts.

To address the above issues, we propose a relation-guided modal fusion module within our framework. We believe that the importance of modal information varies across different relations. For example, the "color" relation focuses more on the image modality, while the "creator" relation pays more attention to the text modality. Therefore, we calculate the weight of each modality through relations, which can be dynamically adjusted when the importance of modal information varies, and add relation context to entities. More specifically, given a triplet (h, r, t), with the head entity h and relation context r, the adaptive weight of each modality m of h is represented as:

$$a_m = \frac{\exp\left(\frac{\sum(h_m \odot r)}{\sqrt{d}}\right)}{\sum_{j \in K} \exp\left(\frac{\sum(h_j \odot r)}{\sqrt{d}}\right)} \tag{1}$$

By using adaptive weights, the joint embedding of the head entity h is aggregated as:

$$h_{joint} = \sum_{m \in \mathcal{M} \cup \{S\}} a_m h_m \tag{2}$$

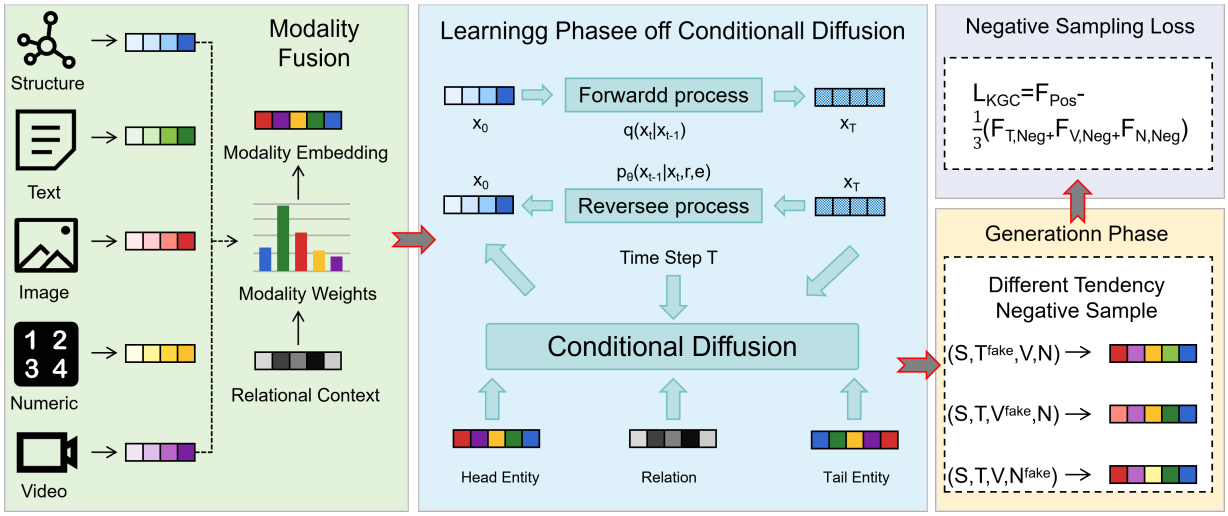

Fig. 1. The MTNS model is mainly divided into three stages.Relationship guided fusion, diffusion negative sampling, and multi-directional modal replacement

Traditional MMKGC methods typically adopt entity-centric modal fusion approaches, regardless of the triplet context. The advantage of our design lies in its ability to dynamically adjust the modal weights of different entities in different triplets. It allows these weights to be dynamically adjusted according to their relation contexts, while taking into account both the modal information of entities and the relation contexts. Moreover, our method can handle an unlimited number of input modalities, rather than considering specific modalities.

Once the joint embeddings of entities are obtained, we use the RotatE scoring function to distinguish the plausibility of triplets. The scoring function is expressed as:

$$\mathcal{F}(h,r,t) = -\left\| h_{joint} \circ r - t_{joint} \right\| \tag{3}$$

where $\odot$ is the rotation operator in the complex space. The higher the score, the more plausible the triplet is. We choose RotatE as the scoring function because RotatE can model the most common relation patterns. During the training process, we adopt a negative sampling based loss function to optimize the parameters, which can be expressed as:

$$\begin{aligned}\mathcal{L}_{kgc} = & \sum_{(h,r,t)\in\mathcal{T}} -\log\sigma(\gamma + \mathcal{F}(h,r,t)) \\ & - \sum_{i=1}^{K} p(h'_i, r'_i, t'_i) \log\sigma(-\mathcal{F}(h'_i, r'_i, t'_i) - \gamma)\end{aligned} \tag{4}$$

Here, σ is the sigmoid function; γ is a fixed margin; $(h'_i, r'_i, t'_i) \in \mathcal{T}'$, $(i = 1, 2, \ldots, K)$ are K negative samples of the triplet (h, r, t). Additionally, $p(h'_i, r'_i, t'_i)$ is the self-adversarial weight proposed in RotatE.

B. Negative Sample Generation via Conditional Diffusion

The previous sampling strategy is straightforward, but it tends to select simple negative samples that contribute little to model training. In fact, multimodal features offer rich semantic and contextual information, which can be leveraged to guide the selection of more realistic negative samples and help the model understand triplets better. Therefore, we aim to generate negatively-biased samples. To achieve this goal, we utilize a conditional diffusion model, taking relations and entities as additional information for negative sample generation.

1) Training Phase: The forward diffusion process can be represented by a Markov chain:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\alpha_t}\mathbf{x}_{t-1}, \alpha_t\mathbf{I}) \tag{5}$$

And the closed form of noisy entity embedding x_t is:

$$\mathbf{x}_t = \sqrt{\bar{\beta}_t}\mathbf{x}_0 + \sqrt{1-\bar{\beta}_t}\boldsymbol{\epsilon}_t \tag{6}$$

where $\beta_t = 1 - \alpha_t$, $\bar{\beta}_t = \prod_{i=1}^{t} \beta_i$, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In the reverse diffusion process, it iteratively denoises the noisy entity embedding $\mathbf{x}_t$ to obtain a synthetic entity embedding. Given the noisy entity embedding $\mathbf{x}_t$ at time - step t, the reverse process is conditioned on the embeddings of an input entity $\mathbf{x}_h$ and a relation $\mathbf{x}_r$, and can be expressed as:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, PE(t), C(\mathbf{x}_e, \mathbf{x}_r)), \alpha_t\mathbf{I}) \tag{7}$$

where θ represents the parameterization of the diffusion model. $\mathrm{PE}(t)$ denotes the positional embedding of time - step t. $\mathbf{x}_e, \mathbf{x}_r$ are conditions from the observed entity E and relation r in the triplet, guiding the generation of a negative triplet entity embedding.

The mean $\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathrm{PE}(t), \mathbf{x}_e, \mathbf{x}_r)$ of the Gaussian distribution in reverse diffusion is parameterized as:

$$\begin{aligned}\mu_\theta(\mathbf{x}_t, PE(t), C(\mathbf{x}_e, \mathbf{x}_r)) = & \frac{1}{\sqrt{\beta_t}}\mathbf{x}_t \\ & - \frac{\beta_t}{\sqrt{\beta_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t, PE(t), C(\mathbf{x}_e, \mathbf{x}_r))\end{aligned} \tag{8}$$

The entity embeddings generated in reverse diffusion at different steps are:

$$\hat{\mathbf{x}}_{t-1} = \frac{1}{\sqrt{\beta_t}}\hat{\mathbf{x}}_t - \frac{\beta_t}{\sqrt{\beta_t}\sqrt{1-\beta_t}}\boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t, PE(t), C(\mathbf{x}_e, \mathbf{x}_r)) + \alpha_t\boldsymbol{\epsilon}_t \tag{9}$$

where $\hat{\mathbf{x}}_T$ is pure noise at step T, and $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{x}}_{t-1}$ ($t \in [1, T-1]$) are embeddings at starting and intermediate steps..

2) Generation Phase: The negative sample generation module integrates multiple modalities to produce semantically rich negative triplets. Specifically, by corrupting the head entity of a positive triplet (e_h, r, e_t), we input the structural features $x^{e_h}_{\text{struc}}$, visual features $x^{e_h}_{\text{vis}}$, and textual features $x^{e_h}_{\text{text}}$ of the tail entity, along with the relation embedding x_r, into the diffusion process to compute multimodal conditions. These conditions guide the reverse processes, generating modality-specific embeddings $\hat{x}^{e_t,t}_{\text{struc}}$, $\hat{x}^{e_t,t}_{\text{vis}}$, and $\hat{x}^{e_t,t}_{\text{text}}$ to construct the generated negative triplets. The diffusion model loss is:

$$\mathcal{L}_{\text{diff}} = \sum_{m=1}^{M} \left\| \epsilon_\theta(x_{t,m}, t, c) - \epsilon_t \right\|^2 \tag{10}$$

To prevent negative sample difficulty from being too low or high, we propose a dynamic adjustment mechanism. It generates negative samples based on the model's current performance, enhancing robustness.

Specifically, at each training stage, we create negative samples of varying difficulty. Entity embeddings are generated from complete noise, iteratively removing predicted noise at each step (a diffusion model process). We adjust diffusion time steps to control negative triplet difficulty: smaller steps yield harder-to-distinguish samples closer to positives, while larger steps produce easier ones. After obtaining embeddings at different steps, we generate increasingly difficult negative samples using a fixed-step function:

$$\text{stage\_idx} = int\left\lfloor \frac{\text{epoch}}{S} \right\rfloor \tag{11}$$

Where S is a fixed value.We find negative samples at the $\frac{T}{\text{stage\_idx}}$ time step from the time steps.This strategy ensures model stability by avoiding overly hard negatives early on and overly easy ones later.

3) Negative Samples with Different Tendency: To address the issue of modal missingness, we utilize diffusion models to generate negative samples. To further enhance model performance and simulate real-world modal missingness scenarios, we propose constructing negative samples with different inclinations. Specifically, we define negative samples with different inclinations as follows.

Visual tendency negative samples: Formed by replacing the original entity text modality and numerical modality embeddings:(h_s, h_v, h_t^*, h_n^*)

Text tendency negative samples: Formed by replacing the original entity visual modality and numerical modality embeddings:(h_s, h_v^*, h_t, h_n^*)

Numeric tendency negative samples: Formed by replacing the original entity visual modality and text modality embeddings:(h_s, h_v^*, h_t^*, h_n)

TABLE I
STATISTICAL INFORMATION OF THE DATASET USED

Dataset	$\|E\|$	$\|R\|$	Train	Valid	Test	$\|T\|$	$\|V\|$
DB15K	12842	279	79222	9902	9904	12818	12842
MKG-W	15000	169	34196	4276	4274	14463	14123

Unlike traditional methods that sample negatives at the entity level, our strategy operates on individual modality embeddings. We generate synthetic triplets, calculate their scores using a fusion module and scoring function F to evaluate plausibility, and denote these as synthetic set S(h,r,t).Therefore, in the overall design, the contrastive loss between negative samples and positive samples can be expressed as:

$$\mathcal{L}_{kgc} = \sum_{(h,r,t)\in\mathcal{T}} \left(-\mathcal{F}(h,r,t) + \frac{1}{|S|} \sum_{\substack{(h^*,r,t^*) \\ \in\mathcal{S}(h,r,t)}} \mathcal{F}(h^*,r,t^*) \right) \tag{12}$$

IV. EXPERIMENT

A. Experimental Setup

1) Datasets: In this paper, we employ two public MMKGC benchmarks, DB15K and MKG-W, to evaluate the model performance. Detailed information about the datasets is presented in Table 1. The raw data for each modality is obtained from their official release sources. During the training process, each dataset is divided into a training set, a validation set, and a test set with a split ratio of 8:1:1.

2) Evaluation Indicators: Building upon prior research, we employ rank-based metrics including Mean Rank Reversal (MRR) and Hits@K (K=1,3,10) to evaluate prediction outcomes. Furthermore, we apply filter settings to exclude candidate triplets in the training data, ensuring fair comparisons. The metrics MRR and Hits@K are defined as follows:

$$\text{MRR} = \frac{1}{|T_{\text{test}}|} \sum_{i=1}^{|T_{\text{test}}|} \left(\frac{1}{r_{h,i} + \frac{1}{r_{t,i}}} \right) \tag{13}$$

$$\text{Hits@K} = \frac{1}{|T_{\text{test}}|} \sum_{i=1}^{|T_{\text{test}}|} \left(1_{(r_{h,i}\le K)} + 1_{(r_{t,i}\le K)} \right) \tag{14}$$

where $r_{h,i}$ and $r_{t,i}$ represent the predicted head entity ranking and tail entity ranking, respectively, while T_{test} denotes the test triple.

3) Baseline: From the perspective of modalities, baselines can be divided into two categories: for single-modal methods, the model is trained using only the structural information of triples; for multi-modal methods, the model is trained with both textual and visual information.

(1) Single-modal KGC methods: TransE [1], ComplEx [2], RotatE [3], PairRE [4], and Tucker [5].

(2) MMKGC methods: IKRL [6], TBKGC [7], TransAE [8], MMKRL [9], VBKGC [10], OTKGE [11], IMF [12], VISTA [13], AdaMF [14], MANS [15], MMRNS [16].

4) Implementation Details: We implemented MTNS using PyTorch, setting the number of training epochs to 1000 and the batch size to 1024. We employed the tokenizers of VGG16 and BERT as the visual/text encoders, with the embedding dimension set to 250. The model was optimized using the Adam optimizer with a learning rate of 5e-4. All experiments were conducted on a Linux server with the Ubuntu 20.04.1 operating system and an NVIDIA RTX 3090 TI GPU.

B. Main Results

The main results of MTNS and baseline methods on link prediction are presented in Table II. Several observations can be drawn from it.

Firstly, MTNS outperforms all baselines on four metrics across both datasets, showing early modality interaction enhances entity representation quality. It improves more significantly in Hit@1 and MRR than in Hit@10, indicating relation-guided fusion boosts prediction accuracy.

Secondly, MTNS, with simpler modules, outperforms other complex multi-modal knowledge graph completion models, demonstrating its effectiveness and efficiency. It uses relation-computed weights to guide fusion, effectively utilizing multi-modal information.

Thirdly, MTNS performs better on DB15K than on MKG-Y because DB15K entities have richer image and text information, aiding multi-modal info utilization.

C. Ablation Experiments

To show MTNS module effectiveness, we did ablation experiments. We removed some modules under different settings for MMKGC tests (results in Table III). Three experiment sets verified each module's contribution. Modal info removal shows all modalities help, with image less so than text. Removing the modal fusion layer hurt performance, proving its role. Negative sampling module tests showed our method notably improves accuracy.

D. Parameter Sensitivity

To assess MTNS's parameter sensitivity and robustness, we ran extensive tests on diffusion model hyperparameters, loss function robustness, and diffusion noise variance impact (results in Figure 2). For diffusion time steps, varying T from 5 to 100 showed that performance slightly drops after T=25, suggesting excessive steps may add redundant computations and harm generalization. Regarding diffusion loss function robustness, L2 loss generally performed best among L1, L2, and Huber losses.

E. Negative Sampling Analysis

To further assess our negative sampling (NS) module, we tested KGE models (TransE, DistMult, RotatE) in MMKGC tasks with a replaced scoring function. When paired with RotatE, MTNS significantly outperforms other NS strategies across all datasets and metrics. For TransE and DistMult, MTNS achieves top or near-top MRR and H@10 scores on all datasets. Notably, MTNS consistently surpasses both traditional and advanced NS methods, like MMRNS, by leveraging multi-modal info across models and datasets. This highlights the advantage of generating hierarchical embeddings for diverse, high-quality negative triplets, rather than simple sampling.

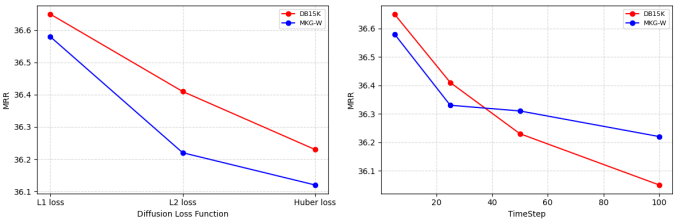

Fig. 2. Parameter Sensitivity.

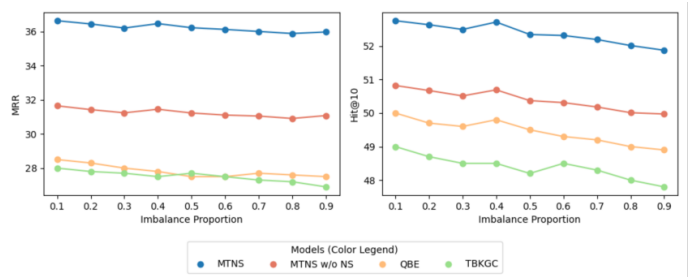

Fig. 3. Imbalance MMKGC Experiments.

V. Conclusion

This paper presents a novel multi-tendency negative sampling framework (MTNS) for multimodal knowledge graph completion. The core innovations include a relation-guided adaptive multimodal fusion module that dynamically assigns modality weights based on relation semantics, and a diffusion-based negative sampling strategy that generates high-quality hard negatives to simulate modality missingness and imbalance. Extensive experiments on two benchmark datasets show that our model significantly outperforms state-of-the-art methods. The relation-guided fusion improves the rationality of multimodal integration, and the diffusion-generated negative samples enhance model discrimination and generalization, jointly leading to consistent performance gains. In the future, we will further explore fine-grained relational-semantic interaction and lightweight optimization of the negative sampling pipeline.

References

[1] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, Translating embeddings for modeling multi-relational data," Advances in neural information processing systems, vol. 26, 2013.

[2] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, Complex embeddings for simple link prediction," in International conference on machine learning, PMLR, 2016, pp. 2071–2080.

[3] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space," arXiv preprint arXiv:1902.10197, 2019.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS

Model Type	Model	DB15K				MKG-W			
		MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Uni-modal	TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	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	26.80	36.68	46.73
	PairRE	31.13	21.62	35.91	49.30	34.40	28.24	36.71	46.04
	Tucker	33.86	25.33	37.91	50.38	30.39	24.44	32.91	41.25
Multi-modal	IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07
	TBKGC	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24
	TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72
	MMKRI	26.81	13.85	35.07	49.39	30.10	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	49.44	30.61	24.91	33.01	40.88
	OTKGE	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88
	IMF	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44
	VISTA	30.42	22.49	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	49.26	30.88	24.89	33.63	41.78
	MMRNS	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47
	MTNS	**36.64**	**27.99**	**41.42**	**53.75**	**36.58**	**29.83**	**38.82**	**51.85**

TABLE III
RESULTS OF ABLATION STUDIES ON DB15K.

Setting	MRR	Hit@1	Hit@3	Hit@10
w/o MF	34.17	24.17	40.23	51.66
w/o Negative	31.65	20.44	38.90	50.82
w/o loss	35.61	26.54	40.91	51.97
w/o MT	35.60	26.50	41.04	51.98
Full MTNS	36.64	27.99	41.42	53.75

TABLE IV
MODEL PERFORMANCE WITH DIFFERENT NS STRATEGIES ON DB15K

Models	NS Strategies	DB15K		
		MRR	Hit@1	Hit@10
TransE	KBGAN	24.00	5.47	50.36
	MMRNS	26.61	13.40	50.60
	DHNS	32.63	20.95	52.86
	MTNS	34.77	24.68	52.93
DistMult	KBGAN	23.36	16.43	36.30
	MMRNS	25.92	14.06	40.42
	DHNS	27.20	19.07	43.01
	MTNS	29.33	22.64	44.23
RotatE	MMRNS	29.67	17.89	51.01
	DHNS	34.36	23.34	53.72
	MTNS	36.64	27.99	53.75

[4] L. Chao, J. He, T. Wang, and W. Chu, Pairre: Knowledge graph embeddings via paired relation vectors," in Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers), 2021, pp. 4360–4369.

[5] I. Balažević, C. Allen, and T. Hospedales, "Tucker: Tensor factorization for knowledge graph completion," in Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 5185–5194.

[6] R. Xie, Z. Liu, H. Luan, and M. Sun, Image-embodied knowledge representation learning," arXiv preprint arXiv:1609.07028, 2016.

[7] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, A multimodal translation-based approach for knowledge graph representation learning," in Proceedings of the seventh joint conference on lexical and computational semantics, 2018, pp. 225–234.

[8] Z. Wang, L. Li, Q. Li, and D. Zeng, Multimodal data enhanced representation learning for knowledge graphs," in 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[9] 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, 2022, Springer.

[10] Y. Zhang and W. Zhang, Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling," arXiv preprint arXiv:2209.07084, 2022.

[11] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang, Otkge: Multimodal knowledge graph embeddings via optimal transport," Advances in neural information processing systems, vol. 35, pp. 39090–39102, 2022.

[12] X. Li, X. Zhao, J. Xu, Y. Zhang, and C. Xing, IMF: interactive multimodal fusion model for link prediction," in Proceedings of the ACM web conference 2023, 2023, pp. 2572–2580.

[13] 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, 2023, pp. 7314–7328.

[14] Y. Zhang, Z. Chen, L. Liang, H. Chen, and W. Zhang, Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion," arXiv preprint arXiv:2402.15444, 2024.

[15] 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, 2023, pp. 1–8.

[16] 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, 2022, pp. 3857–3866.