# Not All Modalities at Once: Dynamic Dropout and Bidirectional Fusion for Robust Multi-modal Knowledge Graph Completion

**Jiashun Peng, Fu Zhang*, Hongzhi Chen, Jingwei Cheng, Yingsong Ning, Xiaoke Wang**
School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
2472039@stu.neu.edu.cn, {zhangfu, chengjingwei}@neu.edu.cn

## Abstract

Multi-modal Knowledge Graph Completion (MKGC) aims to infer missing links in multi-modal knowledge graphs by leveraging structured triples together with auxiliary modalities such as text and images. Existing MKGC methods typically train with all modalities available, implicitly assuming consistent complementarity; however, this practice often induces modality dependence and modality competition under heterogeneous noise, which can hinder robust multi-modal fusion and limit overall performance. To address these issues, we propose **MDBGF**, a **M**odality **D**ropout and **B**idirectional **G**ated **F**usion framework for MKGC. MDBGF introduces a *dynamic, probability-based modality dropout* schedule. When the dropout is activated, MDBGF drops either the textual or visual modality during training while always preserving the structural information, encouraging the model to reduce over-reliance on any single auxiliary modality and to learn complementary cues under missing-modality conditions. When the dropout is not activated (i.e., all modalities are present), we further design a *bidirectional gated fusion* mechanism that enables mutual modulation between textual and visual modalities, enhancing cross-modal interaction and flexible fusion. In addition, we propose an *adaptive proportional hybrid negative sampling* strategy to strengthen MDBGF's discriminative ability on hard negatives. Experiments on three benchmarks show that MDBGF consistently outperforms existing baselines and achieves new state-of-the-art results. Our code is available at https://github.com/ferryman-ship/MDBGF.

## 1 Introduction

Multi-modal Knowledge Graphs (MKGs) (Liu et al., 2019) are advanced versions of traditional knowledge graphs, representing knowledge as structured triples consisting of head entities, relations, and tail entities, while also incorporating multi-modal information associated with entities (e.g., text descriptions and images). Current MKGs still experience serious incompleteness. This highlights the significance of multi-modal knowledge graph completion (MKGC) (Chen et al., 2024), which aims to leverage multi-modal information alongside structured triples to learn richer knowledge representations and infer missing links.

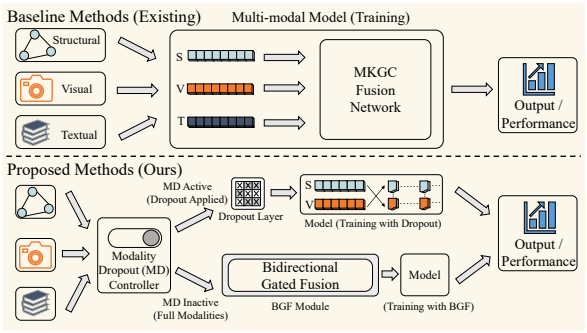


Figure 1: Comparison between existing MKGC methods and our MDBGF. Existing methods typically train on all available modalities and fuse modality-specific representations. In contrast, MDBGF introduces a dynamic, probability-based modality dropout schedule and a bidirectional gated fusion mechanism to promote robust multi-modal learning and fusion.

Existing MKGC approaches (Cao et al., 2022; Lee et al., 2023; Li et al., 2023) typically fuse modality-specific embeddings via dedicated fusion modules to assess the plausibility of triples. To better exploit multi-modal information[1], recent MKGC studies are beginning to focus on the imbalance problem of modalities and the design of more flexible fusion mechanisms (Zhang et al., 2024, 2025a), while negative sampling is also increasingly adopted to help models learn from incorrect samples (Xu et al., 2022; Zhang et al., 2025b).

* Corresponding author.

[1] For simplicity, we also refer to structured triple information as the *structural modality*.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 17928–17940

Despite these advances, most existing MKGC methods train models by indiscriminately using all available modalities as shown in Fig. 1, implicitly assuming that modalities always provide complementary and useful information. As a result, models can be prone to modality dependence, over-relying on one or a few modalities while under-utilizing latent complementary signals from others. Moreover, given the heterogeneous noise levels across modalities, such training can suffer from modality competition (Huang et al., 2022), potentially leading to suboptimal solutions and limiting the robustness of MKGC. However, existing MKGC studies have not sufficiently explored or evaluated how individual modalities affect training dynamics at different stages. This motivates us to explicitly regulate modality usage during training and to design more robust fusion strategies.

To this end, we propose **MDBGF**, a modality dropout and bidirectional gated fusion framework for MKGC, together with an adaptive proportional hybrid negative sampling strategy. Instead of permanently discarding any modality, we introduce a dynamic, probability-based **modality dropout (MD)** mechanism. Specifically, based on our probability formulation and a gradually increasing dropout schedule, we decide for each training epoch whether to perform dropout; *when dropout is activated*, we randomly drop either the textual or the visual modality (while always keeping the structural modality). This decouples the model from dependence on any specific auxiliary modality (textual or visual): it is encouraged to learn to infer missing entities from structure+text even without images, and from structure+vision even without text. Moreover, *when dropout is not activated* (i.e., under complete-modality cases that occur with high probability in the early stages of training), we additionally design a **bidirectional gated fusion (BGF)** mechanism, in which the textual and visual modalities influence each other's fusion weights through learnable gating networks. Overall, the dynamic MD and BGF effectively regulate modality usage during training and improves the robustness of multi-modal fusion in MKGC.

Furthermore, to help the proposed framework learn patterns from incorrect samples, we design a novel negative sampling strategy. In contrast to previous MKGC methods that rely on contrastive-learning-based negative sampling (Zhang et al., 2025b) or generate new samples by injecting noise within a modality (Zhang et al., 2024), we propose an **adaptive proportional hybrid negative sampling (AHNS)** strategy. AHNS employs a hybrid scheme that combines in-batch random negatives with high-score negatives, and uses an adaptively adjusted proportional mechanism to control their respective numbers, thereby further enhancing MDBGF's discriminative ability on hard negatives.

In summary, our contributions are as follows:

- We propose an MKGC framework that incorporates a dynamic, probability-based modality dropout mechanism, mitigating modality dependence and improving robustness in multi-modal learning.
- We design a bidirectional gated fusion mechanism for complete-modality cases when dropout is not activated, where textual and visual modalities modulate each other's fusion weights through learnable gating networks. This enhances cross-modal interaction and yields more effective multi-modal fusion.
- We introduce a new adaptive proportional hybrid negative sampling strategy to further strengthen the model's discriminative ability, especially on hard negatives.
- Experiments on three datasets demonstrate that MDBGF outperforms previous methods, achieving new state-of-the-art results. Further analysis and experiments show that dynamic dropout mechanism consistently improves performance across multiple backbone models, verifying its generalizability for MKGC.

## 2 Related Work

### 2.1 MKGC Methods

MKGC has attracted increasing attention. Early methods like IKRL (Xie et al., 2016), TBKGC (Mousselly-Sergieh et al., 2018), TransAE (Wang et al., 2019) employ multiple translation-based scoring functions for embeddings from diverse modalities. Later methods take modal fusion into account; for example, OTKGE (Cao et al., 2022), VISTA (Lee et al., 2023), and AdaMF (Zhang et al., 2024) introduce sophisticated feature fusion techniques such as optimal transfer, transformer, and adversarial training into multi-modal fusion.

Recently, finer-grained multi-modal information is being explored. MyGO (Zhang et al., 2025b) optimizes visual and textual modal information at the token level. MoMoK (Zhang et al., 2025a)

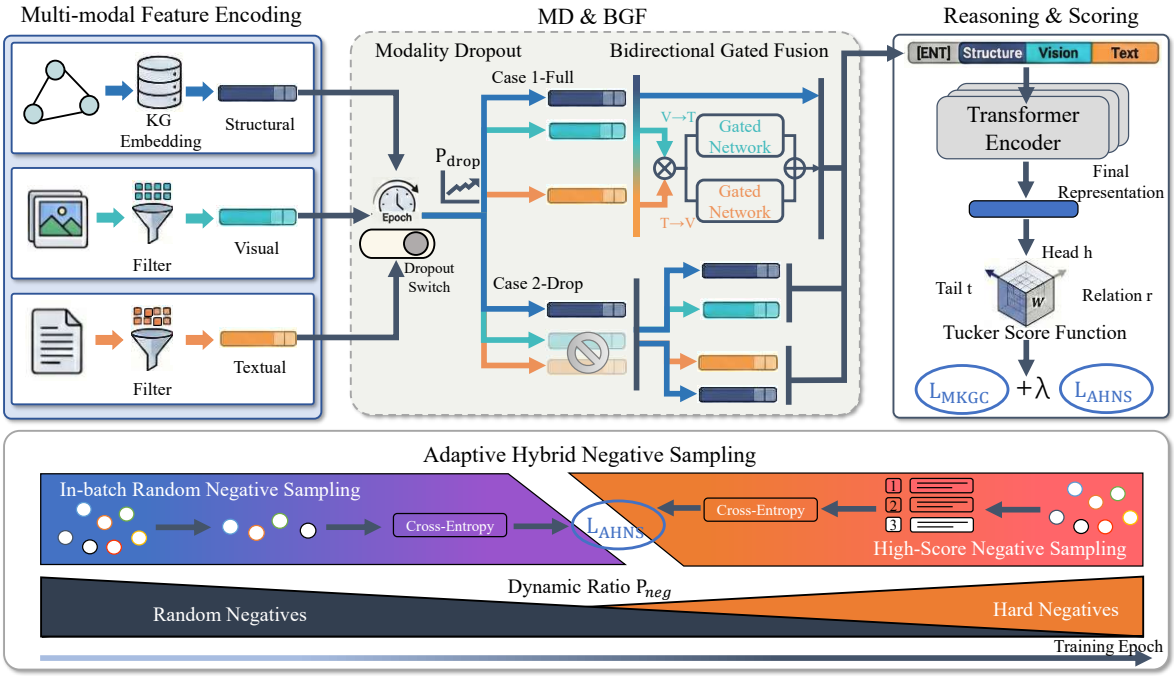


Figure 2: Overview of our framework MDBGF. Firstly, **Multi-modal Feature Encoding** encodes structural, textual, and visual inputs into modality-specific embeddings. Then, when dropout is activated based on our probability formulation and a gradually increasing dropout schedule, **Modality Dropout** (MD) drops either the textual or the visual modality; otherwise, when dropout is not activated (i.e., under complete-modality cases), **Bidirectional Gated Fusion** (BGF) enables mutual weight adjustment between textual and visual modalities for effective cross-modal fusion. Finally, MDBGF performs scoring-based prediction and is jointly trained with our **Adaptive Hybrid Negative Sampling** (AHNS), which dynamically balances random and high-score negatives.

leverages a mixture of experts to decouple mutual information. SRE (Wang et al., 2025) captures entities' semantic proximity by segmenting semantic similarity. APKGC (Jian et al., 2025) employs a noise-enhanced penalty mechanism to alleviate excessive modal trust. MCKGC (Gao et al., 2025) embeds multi-modal features into hyperbolic, spherical, and Euclidean spaces simultaneously, and achieves adaptive information fusion.

Despite these advances, existing MKGC methods typically train models by indiscriminately using all available modalities, without systematically exploring how individual modalities affect training dynamics and model quality. To address this gap, we propose a probability-based modality dropout and bidirectional gated fusion framework that finegrainedly regulates modality usage during training and enables more effective fusion for MKGC.

### 2.2 Negative Sampling Strategies for MKGC

To enhance MKGC by improving negative sample quality, recent studies increasingly incorporate negative sampling optimization into MKGC. MMRNS (Xu et al., 2022) introduces relation embeddings in the sampling stage to guide visual and textual features in generating more challenging and diverse negative examples. AdaMF (Zhang et al., 2024) adds noise to visual and textual modalities to generate negative samples. MyGO (Zhang et al., 2025b) adopts a contrastive negative sampling strategy.

Unlike existing strategies, we propose an adaptive proportional hybrid method that dynamically combines in-batch random negatives with high-score negatives, thereby further enhancing the model's discriminative ability.

## 3 Problem Formulation

The MKG $= \{\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M}\}$ represents a structured knowledge graph composed of an entity set $\mathcal{E}$, a relation set $\mathcal{R}$, a triplet set $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ , and a modality set $\mathcal{M} = \{\mathcal{V}, \mathcal{D}\}$, where $\mathcal{V}$ and $\mathcal{D}$ denote the sets of vision and text. The goal of MKGC is to predict missing entities for a given query $(h, r, ?)$ or $(?, r, t)$.

## 4 Methodology

In this section, we propose our framework **MDBGF**. Fig. 2 provides an overview of the design.

### 4.1 Multi-modal Feature Encoding

In MKGC, each entity is associated with heterogeneous multi-modal information, including structural, visual, and textual modalities. We first encode raw inputs from different modalities into a unified embedding space.

For the structural modality, entities and relations are represented as triples $(h, r, t)$. We use randomly initialized learnable embeddings to obtain structural representation $\mathbf{e}_{str}$, which serves as structural priors for multi-modal modeling.

For visual and textual modalities, entity images and descriptions are first tokenized using a pre-trained visual encoder and a textual encoder, respectively. To reduce noise and redundancy, a frequency-based filtering strategy (Zhang et al., 2025b) is applied to select representative visual and textual tokens. The retained tokens are then projected into continuous embeddings, resulting in visual and textual representations $\mathbf{e}_{vis}$ and $\mathbf{e}_{txt}$.

### 4.2 Modality Dropout and Bidirectional Gated Fusion

After initial multi-modal feature encoding, the model further integrates cross-modal semantics to learn improved joint representations.

#### 4.2.1 Modality Dropout

To mitigate modality dependence and better exploit cross-modal complementarity, we introduce a dynamic, probability-based modality dropout mechanism. Instead of permanently removing any modality, we compute dropout probabilities and adopt a gradually increasing schedule during training.

Specifically, in each training epoch, this module randomly masks textual or visual modal inputs with a dynamically increasing probability, where the dropout probability is defined as follows:

$$P_{drop} = min(\alpha, \frac{epoch}{N_{epoch}} * \alpha) \tag{1}$$

where $P_{drop}$ denotes the modality dropout probability, $\alpha$ is a hyperparameter that controls the maximum dropout probability, $epoch$ represents the current training round, and $N_{epoch}$ is the total number of epochs. In the early training stage, after the $\min$ operation, the value of $P_{drop}$ is relatively small, allowing the model to fully learn the complete information of each modality. As training progresses, $P_{drop}$ will gradually increase to a value close to $\alpha$, enabling the model to gradually develop the ability to reason in modality-missing scenarios without relying on any single modality with large information volume (either visual or textual).

After processing by the MD module, we define the following two types of data to indicate whether the modality dropout operation has been performed in the current training epoch:

$$\begin{cases} \text{Case 1 (no dropout)}: & \mathbf{e} = \{\mathbf{e}_{str}, \mathbf{e}_{vis}, \mathbf{e}_{txt}\} \\ \text{Case 2 (dropout)}: & \begin{cases} \mathbf{e} = \{\mathbf{e}_{str}, \mathbf{e}_{vis}\} \\ \mathbf{e} = \{\mathbf{e}_{str}, \mathbf{e}_{txt}\} \end{cases} \end{cases} \tag{2}$$

where $\mathbf{e}$ denotes the set of multi-modal representations. Case 1 corresponds to the complete-modality setting without dropout, while Case 2 activates dropout, where either the visual or textual modality is randomly discarded with equal probability.

Subsequently, we will conduct subsequent processing separately for these two cases to highlight their differences in the overall workflow.

#### 4.2.2 Bidirectional Gated Fusion

This module focuses on bidirectional interaction between textual and visual modalities and is therefore applied only in the complete-modality setting (Case 1). When the modality dropout is activated (Case 2), the model directly proceeds to the subsequent process in Eq. (11). Unlike the direct attention fusion (Zhang et al., 2024) or one-way expert decision-making (Zhang et al., 2025a), BGF employs learnable gating networks to enable mutual modulation between textual and visual modalities, leading to more effective cross-modal interaction.

In BGF, we first perform average pooling on the feature representations of the textual and visual modalities respectively to obtain the global feature representation of each modality. we extract the visual modality $\mathbf{e}_{vis}$ and text modality $\mathbf{e}_{txt}$ from the set $\mathbf{e}$ in Eq. (2), and obtain their global representations by averaging over the feature dimension:

$$\mathbf{e}_{vis\_avg} = \mathcal{M}ean(\mathbf{e}_{vis}, \ keepdim = true) \tag{3}$$

$$\mathbf{e}_{txt\_avg} = \mathcal{M}ean(\mathbf{e}_{txt}, \ keepdim = true) \tag{4}$$

where $keepdim$ keeps the dimensions unchanged, and $\mathcal{M}ean$ averages multiple visual or textual features of the same entity to obtain a global representation. The global features of the two modalities are then concatenated and fed into gated networks to learn cross-modal adjustment weights. The gated weight *from vision to text* is defined as:

$$\mathcal{X}_{v \to t} = \{\mathbf{e}_{vis\_avg}, \ \mathbf{e}_{txt\_avg}\} \tag{5}$$

$$\mathcal{G}_{v \to t} = \mathcal{S}igmoid(\mathcal{W}_{v \to t} \cdot \mathcal{X}_{v \to t} + b_{v \to t}) \tag{6}$$

where $\mathcal{G}_{v\to t}$ denotes the gated weight from the visual to the textual modality; $\mathcal{W}_{v\to t}$ is a learnable weight matrix, and $b_{v\to t}$ is a bias term. And the gated weight *from text to vision* is defined as:

$$\mathcal{X}_{t\to v} = \{\mathbf{e}_{txt\_avg},\ \mathbf{e}_{vis\_avg}\} \tag{7}$$

$$\mathcal{G}_{t\to v} = \mathcal{S}igmoid(\mathcal{W}_{t\to v}\cdot\mathcal{X}_{t\to v} + b_{t\to v}) \tag{8}$$

After obtaining the bidirectional gated weights between the visual and textual modalities, we use its weights to perform element-wise weighting on the original modal features, so as to achieve cross-modal interaction:

$$\mathbf{e}'_{vis} = \mathbf{e}_{vis}\cdot\mathcal{G}_{v\to t} \tag{9}$$

$$\mathbf{e}'_{txt} = \mathbf{e}_{txt}\cdot\mathcal{G}_{t\to v} \tag{10}$$

where $\mathbf{e}'$ denotes the embedding of each modality obtained after bidirectional gating adjustment. After performing the operations of MD and BGF, we redefine the results of Case 1 and Case 2:

$$\begin{cases}\text{Case 1 (no dropout)}: & \mathbf{e}_{input} = (\text{[ENT]}, \mathbf{e}_{str}, \mathbf{e}'_{vis}, \mathbf{e}'_{txt})\\ \text{Case 2 (dropout)}: & \begin{cases}\mathbf{e}_{input} = (\text{[ENT]}, \mathbf{e}_{str}, \mathbf{e}_{vis}, \mathbf{0})\\ \mathbf{e}_{input} = (\text{[ENT]}, \mathbf{e}_{str}, \mathbf{0}, \mathbf{e}_{txt})\end{cases}\end{cases} \tag{11}$$

where $\mathbf{e}_{input}$ denotes the concatenated embedding sequence fed into the encoder, [ENT] is the special token prepended to the sequence, and $\mathbf{0}$ indicates that a modality is masked by MD. After constructing $\mathbf{e}_{input}$ for Case 1 and Case 2, we feed the sequence into a Transformer encoder (Vaswani et al., 2017) to obtain the final representation of the head (or tail) entity, denoted as $\mathbf{h}$ (or $\mathbf{t}$).

### 4.3 Triple Scoring and Prediction

For a given query $(h, r, ?)$, we obtain the head embedding $\mathbf{h}$ for entity $h$, combine it with a randomly initialized learnable relation embedding $\mathbf{r}$, and feed the result into a relation decoder for entity prediction. During inference, the decoder ranks candidate entities via a scoring function. We employ TuckER (Balažević et al., 2019) as the scoring function:

$$\mathcal{S}(h, r, t) = \mathcal{W}\times_1\mathbf{h}\times_2\mathbf{r}\times_3\mathbf{t} \tag{12}$$

where $\mathcal{W}$ is the core tensor learned during training, $\times_i$ represents the tensor product along the $i$-th mode. Our model is optimized using a cross-entropy loss applied to each triple. In this process, $t$ is regarded as the true label when compared with the entire entity set $\mathcal{E}$, and the same way is adopted for the head prediction $(?, r, t)$. Consequently, the overall training objective is formulated as a cross-entropy loss:

$$\mathcal{L}_{MKGC} = -\sum_{(h,r,t)\in\mathcal{T}} log\frac{exp(\mathcal{S}(h,r,t))}{\sum_{t'\in\mathcal{E}} exp(\mathcal{S}(h,r,t'))} \tag{13}$$

where $\mathcal{T}$ denotes the triple set.

### 4.4 Adaptive Hybrid Negative Sampling

To enhance model's ability to learn from negative samples, we propose AHNS, which adaptively balances in-batch random and high-score negatives via a dynamic proportional strategy.

Specifically, AHNS mainly consists of an adaptively adjusted proportional mechanism, where hybrid negative sampling combines in-batch random negatives and high-score negatives (hard negatives), and dynamically adjusts the number of the two types according to the adaptive proportion. Suppose that negative samples need to be sampled in each batch, and this proportion is defined as:

$$P_{neg} = min(\beta, \frac{epoch}{N_{epoch}} * \beta) \tag{14}$$

where $P_{neg}$ denotes the adaptive proportion that changes with the training epochs, the parameter $\beta$ is a manually set maximum proportion. Subsequently, the total number of negative samples $N_{neg}$ is multiplied by the proportional $P_{neg}$ to obtain the number of high-score negative samples, and the remaining part $(1 - P_{neg})$ corresponds to the number of random negative samples:

$$N_{hard} = N_{neg}\times P_{neg} \tag{15}$$

$$N_{random} = N_{neg} - N_{hard} \tag{16}$$

where $N_{hard}$ denotes the number of high-score samples selected as hard negatives, i.e., the top-ranked samples according to the model's scores. These samples typically lie near the decision boundary between positive and negative instances and help improve discriminative accuracy. $N_{random}$ denotes the number of randomly sampled negatives. Notably, $P_{neg}$ increases progressively with training epochs; accordingly, the model emphasizes random negatives in the early stage to preserve sample diversity and gradually shifts its focus to hard negatives in later stages to enhance discrimination.

AHNS also adopts a cross-entropy loss function:

$$\mathcal{L}_{AHNS} = -\sum_{(h,r,t)\in\mathcal{T}'} log\frac{exp(\mathcal{S}(h,r,t))}{\sum_{t'\in\mathcal{E}'} exp(\mathcal{S}(h,r,t'))} \tag{17}$$

| Model | MKG-W | | | | MKG-Y | | | | DB15K | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | H@1 | H@3 | H@10 | MRR | H@1 | H@3 | H@10 | MRR | H@1 | H@3 | H@10 |
| **Uni-modal KGC Methods** | | | | | | | | | | | | |
| TransE (Bordes et al., 2013) | 29.19 | 21.06 | 33.20 | 44.23 | 30.73 | 23.45 | 35.18 | 43.37 | 24.86 | 12.78 | 31.48 | 47.07 |
| DistMult (Yang et al., 2015) | 20.99 | 15.93 | 22.28 | 30.86 | 25.04 | 19.33 | 27.80 | 35.95 | 23.03 | 14.78 | 26.28 | 39.59 |
| RotatE (Sun et al., 2019) | 33.67 | 26.80 | 36.68 | 46.73 | 34.95 | 29.10 | 38.35 | 45.30 | 29.28 | 17.87 | 36.12 | 49.66 |
| TuckER (Balažević et al., 2019) | 30.39 | 24.44 | 32.91 | 41.25 | 37.05 | 34.59 | 38.43 | 41.45 | 33.86 | 25.33 | 37.91 | 50.38 |
| **Multi-modal KGC (MKGC) Methods** | | | | | | | | | | | | |
| IKRL (Xie et al., 2016) | 32.36 | 26.11 | 34.75 | 44.07 | 33.22 | 30.37 | 34.28 | 38.26 | 26.82 | 14.09 | 34.93 | 49.09 |
| TBKGC (Mousselly-Sergieh et al., 2018) | 31.48 | 25.31 | 33.98 | 43.24 | 33.99 | 30.47 | 35.27 | 40.07 | 28.40 | 15.61 | 37.03 | 49.86 |
| TransAE (Wang et al., 2019) | 30.00 | 21.23 | 34.91 | 44.72 | 28.10 | 25.31 | 29.10 | 33.03 | 28.09 | 21.25 | 31.17 | 41.17 |
| OTKGE (Cao et al., 2022) | 34.36 | 28.85 | 36.25 | 44.88 | 35.51 | 31.97 | 37.18 | 41.38 | 23.86 | 18.45 | 25.89 | 34.23 |
| IMF (Li et al., 2023) | 34.50 | 28.77 | 36.62 | 45.44 | 35.80 | 33.00 | 37.10 | 40.60 | 32.25 | 24.20 | 36.00 | 48.19 |
| VISTA (Lee et al., 2023) | 32.91 | 26.12 | 35.38 | 45.61 | 30.45 | 24.87 | 32.39 | 41.53 | 30.42 | 22.49 | 33.56 | 45.94 |
| MoMoK (Zhang et al., 2025a) | 35.89 | 30.38 | 37.08* | 45.48* | 37.91 | 35.09 | 39.15* | 42.38* | 39.57 | 32.38 | 43.45 | 54.14 |
| SRE (Wang et al., 2025) | 37.20 | 30.80 | 39.90 | 48.90 | 37.30 | 34.50 | 39.20 | 42.90 | 37.60 | 30.30 | 41.00 | 51.50 |
| APKGC (Jian et al., 2025) | 37.40 | 30.60 | 40.40 | 50.10 | - | - | - | - | 36.40 | 28.20 | 41.30 | 52.70 |
| MCKGC (Gao et al., 2025) | 36.88 | 31.32 | 38.92 | 47.43 | 38.92 | **35.49** | 40.57 | 45.21 | 39.79 | 31.92 | 43.80 | 54.66 |
| **MKGC with Negative Sampling Methods** | | | | | | | | | | | | |
| MMRNS (Xu et al., 2022) | 35.03 | 28.59 | 37.49 | 47.47 | 35.93 | 30.53 | 39.07 | 45.47 | 32.68 | 23.01 | 37.86 | 51.01 |
| AdaMF (Zhang et al., 2024) | 34.27 | 27.21 | 37.86 | 47.21 | 38.06 | 33.49 | 40.44 | 45.48 | 32.51 | 21.31 | 39.67 | 51.68 |
| MyGO (Zhang et al., 2025b) | 36.10 | 29.78 | 38.54 | 47.75 | 36.57 | 33.48 | 38.40 | 41.76 | 37.72 | 30.08 | 41.26 | 52.21 |
| **MDBGF (ours)** | **40.10** | **33.42** | **42.83** | **52.35** | **39.13** | 35.15 | **40.78** | **46.06** | **40.58** | **32.77** | **44.23** | **55.53** |
| *Improvements* | 7.22% | 6.70% | 6.01% | 4.49% | 0.54% | - | 0.52% | 1.28% | 1.99% | 2.66% | 0.98% | 1.59% |

Table 1: The main results on three widely used MKGC datasets. Best results are in bold and second best results are underlined. Results marked with * are from our reproduction, while the others are taken from the original papers.

where $\mathcal{T}'$ denotes the in-batch triple set, and $\mathcal{E}'$ denotes the in-batch tail entity set. Therefore, the total training loss $\mathcal{L}$ of the model is expressed as:

$$\mathcal{L} = \mathcal{L}_{MKGC} + \lambda \mathcal{L}_{AHNS} \tag{18}$$

where $\lambda$ is a hyper-parameter to control the weight of the $\mathcal{L}_{AHNS}$.

## 5 Experiments and Results

### 5.1 Experimental Setup

**Datasets.** We evaluate our method on three widely used MKGC datasets: DB15K (Liu et al., 2019), MKG-W (Xu et al., 2022), and MKG-Y (Xu et al., 2022), which are detailed in **Appendix A.1**.

**Baselines.** We compare our method with prior SOTA models, including: **(1)** ***Uni-modal KGC***: TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), RotatE (Sun et al., 2019), TuckER (Balažević et al., 2019). (2) ***Multi-modal KGC (MKGC)***: IKRL (Xie et al., 2016), TBKGC (Mousselly-Sergieh et al., 2018), TransAE (Wang et al., 2019), OTKGE (Cao et al., 2022), IMF (Li et al., 2023), VISTA (Lee et al., 2023), MoMoK (Zhang et al., 2025a), SRE (Wang et al., 2025), APKGC (Jian et al., 2025), MCKGC (Gao et al., 2025). (3) ***MKGC with negative sampling***: MMRNS (Xu et al., 2022), AdaMF (Zhang et al., 2024), MyGO (Zhang et al., 2025b).

**Evaluation Metrics.** We use standard evaluation metrics: Mean Reciprocal Rank (MRR) and Hit Rate (H@1, H@3, and H@10). **Appendix A.2** provides detailed descriptions of these metrics.

**Implementation Details.** We conduct our experiments on a single A800 GPU. For modality tokenization, we employ the tokenizer of BEIT (Peng et al., 2022) and BERT (Devlin et al., 2019) as our visual/textual tokenizers. More implementation details can be found in **Appendix A.3**.

### 5.2 Main Results

The main results are reported in Table 1. Overall, our MDBGF consistently outperforms both uni-modal KGC methods and recent MKGC baselines on all three datasets, achieving new state-of-the-art performance. On MKG-W, MDBGF yields the largest gains, improving over the best prior result by **7.22**% in MRR and **6.70**% in H@1. On DB15K and MKG-Y, MDBGF also outperforms strong baselines on most metrics, further validating the effectiveness of our method.

Compared with recent methods MoMoK (Zhang et al., 2025a), APKGC (Jian et al., 2025), SRE (Wang et al., 2025), and MCKGC (Gao et al., 2025), our model still achieves consistent gains. The advantage of MDBGF mainly stems from several designs. First, the proposed MD module constructs

| Model | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| MDBGF | **40.10** | **33.42** | **42.83** | **52.35** |
| w/o MD | 38.63 | 32.34 | 40.93 | 50.63 |
| w/o BGF | 38.70 | 32.35 | 40.91 | 51.20 |
| w/o AHNS | 38.63 | 31.91 | 41.62 | 50.80 |

Table 2: Ablation study on the MKG-W dataset.

diverse modality-availability scenarios during training, alleviating reliance on a specific modality and encouraging the model to exploit complementary cues across modalities. Second, the BGF module introduces bidirectional gating between textual and visual modalities, allowing them to modulate each other's fusion weights dynamically, which strengthens cross-modal interaction and yields more effective fusion. These benefits are further supported by the results that MDBGF also surpasses MKGC methods equipped with negative sampling strategies (e.g., MMRNS, AdaMF, and MyGO).

### 5.3 Ablation Study

To investigate contributions brought by each module, we conduct ablation experiments in Table 2:

$w/o$ MD. Training the model with all three modalities consistently, instead of employing the diverse modality combinations enabled by MD, leads to a significant performance drop. This suggests that MD mitigates the model's reliance on individual modalities and facilitates the exploitation of cross-modal complementary information.

$w/o$ BGF. When BGF is removed and replaced with a simple attention-based fusion mechanism, performance consistently degrades across all metrics. This observation validates the effectiveness of the proposed BGF approach, indicating that bidirectional gated weighting of the information-rich visual and textual modalities is necessary.

$w/o$ AHNS. The removal of AHNS causes significant performance drops. By gradually shifting from random to hard negative sampling during training, AHNS facilitates fine-grained discrimination, especially for Top-1 ranking, which explains the larger decreases observed in MRR and Hits@1.

### 5.4 Effect of Modality Dropout Strategies

To investigate the impact of different modality dropout strategies on performance, we conduct the experiments presented in Table 3.

The "dp X" denotes discarding X modality when performing MD. Discarding a single modality alone fails to achieve optimal performance. In contrast, randomly dropping either the textual or visual modality during training exposes the model to diverse reasoning scenarios, leading to the best overall performance (base).

| Modality | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| dp visual | 38.84 | 32.41 | 41.09 | 51.49 |
| dp textual | 38.77 | 32.24 | 41.39 | 50.88 |
| dp vis + txt | 39.27 | 32.76 | 41.62 | 51.23 |
| base (dp vis or txt) | **40.10** | **33.42** | **42.83** | **52.35** |

Table 3: Results of different modality dropout strategies.

| Training Stage | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| Early | 34.61 | 28.54 | 36.38 | 46.44 |
| Late | 37.06 | 30.42 | 39.58 | 49.70 |
| $P_{drop}$ (ours) | **40.10** | **33.42** | **42.83** | **52.35** |

Table 4: Impact of modality dropout at different training stages.

Moreover, we investigate the impact of MD at different training stages in Table 4. With a boundary at 500 epochs, early and late apply modality dropout only before or after the boundary, respectively. Our probability-based MD strategy $P_{drop}$ performs best by dynamically applying dropout throughout training, enabling more balanced multimodal fusion across stages. This further confirm the effectiveness of our dynamic, probability-based modality dropout mechanism for MKGC.

### 5.5 Generalization of Modality Dropout

To verify the generalization of MD, we integrate it into different backbone models in Fig. 3.

We adopt recent MyGO and AdaMF as backbone models. The upper parts of the bars indicate the performance gains brought by MD, which demonstrate the generality of MD and its strong performance within the MDBGF framework. The results further confirm the generalization of our dynamic, probability-based modality dropout mechanism for MKGC, and may also offer a useful perspective for future MKGC research.

### 5.6 Analysis of Bidirectional Gated Fusion

To investigate whether our designed BGF is particularly beneficial for the text–vision interaction, we evaluate BGF with different modality pairings.

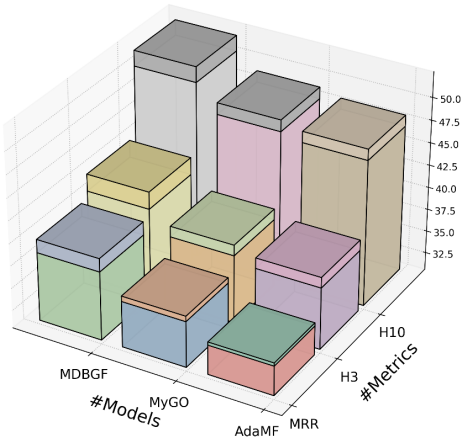


Figure 3: Generalization of our proposed dynamic modality dropout (MD) across different backbones on MKG-W dataset. The upper parts of the bars indicate the performance gains brought by MD.

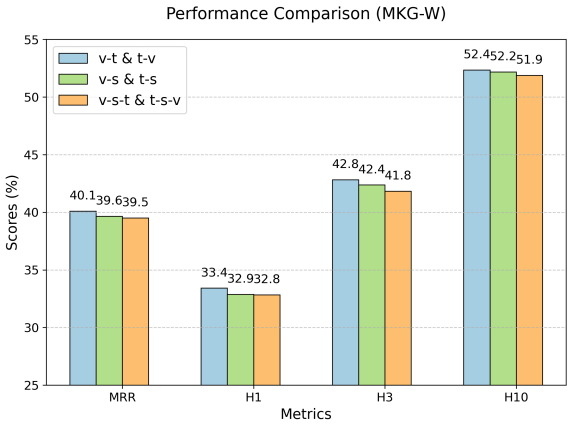


Figure 4: The results of bidirectional gated fusion after combining different modalities on MKG-W dataset.

The results are shown in Fig. 4, where "v-t & t-v" indicates that the input $\mathcal{X}$ combinations used in Eq. (5) and the rest can be done in the same manner. Comparing different modality combinations in BGF, we observe that bidirectional gating between visual and textual modalities yields the most significant gains. This is because structural information is relatively stable and dense, whereas textual and visual modalities exhibit stronger semantic complementarity. Gating-based fusion of these two modalities thus produces more discriminative cross-modal representations, making BGF particularly effective for visual–textual interaction.

### 5.7 Generalization of AHNS

To balance performance and sampling overhead, we propose the AHNS and compare it with other negative sampling methods in Table 5.

| Method | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| MDBGF (w/o) | 38.63 | 31.91 | 41.62 | 50.80 |
| + FGCL | 39.98 | 33.36 | 42.76 | 52.07 |
| + Noise | 39.32 | 33.06 | 41.91 | 51.11 |
| + AHNS | **40.10** | **33.42** | **42.83** | **52.35** |
| MyGO (w/o) | - | - | - | - |
| + FGCL | 36.10 | 29.78 | 38.54 | 47.75 |
| + AHNS | 37.50 | 31.64 | 39.77 | 49.30 |

Table 5: Performance comparison of AHNS with other negative sampling methods on the MKG-W dataset. (w/o) denotes the results without negative sampling.

| Model | GPU Memory(MB) |
|---|---|
| Base | 20622 |
| +MD | 20622 |
| +BGF | 21284 |
| +AHNS | 20624 |
| +MD& BGF | 21284 |
| +MD& BGF& AHNS | **21286** |

Table 6: Fine-Grained analysis of resource consumption.

We compare AHNS with two representative alternatives in terms of performance and overhead: (i) FGCL, a contrastive negative sampling in MyGO (Zhang et al., 2025b), with a memory overhead of 39551M, and (ii) Noise, the adversarial noise-based strategy in AdaMF (Zhang et al., 2024) that jointly trains a noise generator, incurring a higher overhead of 53350M. In contrast, AHNS combines random negatives with high-score negatives and adaptively adjusts their proportions, improving the hardness and diversity of negatives with minimal overhead (21286M). As shown in the results, AHNS consistently delivers better performance and further boosts the backbone when used as a plug-in sampler, demonstrating superior effectiveness and generalization with minimal overhead.

### 5.8 Fine-Grained Analysis of Resource Consumption

To further investigate the efficiency of our method, we incrementally add modules layer by layer based on the base model to identify the main sources of overhead in our approach. The results are shown in Table 6.

As can be observed, the memory overhead of the MD module is negligible since it adopts a modality dropout strategy during training. The BGF contains a bidirectional gating network between visual and

| Baseline Model | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| structural + visual + textual | 37.3 | 30.5 | 40.0 | 49.4 |
| structural + textual | **38.9** | **31.4** | **40.9** | **50.2** |
| structural + visual | 36.7 | 30.4 | 39.2 | 48.6 |
| visual + textual | 36.4 | 30.2 | 38.6 | 48.2 |

Table 7: Performance of different modality combinations on MKG-W dataset.

| Test Scenario | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| Full | **40.10** | **33.42** | **42.83** | **52.35** |
| Deficient | 38.19 | 31.76 | 40.62 | 50.35 |

Table 8: Comparison of different test scenarios.

text modalities, making it the main source of overhead in our method. The overhead of the AHNS module is also low, as it only samples sample scores at the end of each training epoch and then weights their loss into the final model training loss.

### 5.9 Analysis of Modality Competition and Interference

To further investigate the role of each modality in MKGC, we conducted experiments with different modality combinations, the results are shown in Table 7.

We perform ablation studies by removing the MD and BGF modules, retaining only the baseline model. Inference results under different modality combinations show that adding visual modality to structural and textual modalities degrades some metrics, and performance drops further when key modalities (e.g., structural) are removed. As discussed in Appendix C, redundant visual features can interfere with other modalities. This motivates our MD module, which probabilistically discards text or visual modality to mitigate noise from excessive useless features.

### 5.10 Scenario-Deficient Testing: A Study

To better justify the core motivation for robustness, we conduct additional simulations and evaluations on missing scenarios over the test set, with results illustrated in Table 8.

Full denotes the complete test scenario, while Deficient refers to the scenario with partially missing test data. As no standard missing-scenario test set exists for MKGC, we apply the same MD operation to test data as in training. Specifically, we mask partial modality information during inference, consistent with the MD module in training, to evaluate its robustness. Although no official benchmarks are available, our model still performs well under the constructed missing inference setting, with only a 1–2 point performance drop compared to the full data. This validates the effectiveness of MD under various inference conditions.

### 5.11 Hyperparameter Analysis and Case Study

The analysis of hyperparameters (including $\alpha$ in dropout probability of Eq. (1), the sample proportion $\beta$ in Eq. (14), the loss weight $\lambda$ in Eq. (18), and modality token scales) is reported in **Appendix B**. Case studies are also provided in **Appendix C**.

## 6 Conclusion

We propose MDBGF, a framework for MKGC that addresses modality dependence and interaction through three collaborative components. The dynamic MD mechanism improves robustness by introducing diverse modality scenarios during training, the BGF enables adaptive visual–textual interaction for selective integration of high-quality semantics, and AHNS dynamically balances random and hard negatives to enhance discriminative capability. Experiments demonstrate that MDBGF effectively learns and fuses multi-modal information, and the dynamic modality dropout mechanism may offer a useful perspective for MKGC research.

## Limitations

Although our MDBGF framework demonstrates significant advantages in MKGC, it still has several limitations. First, while the MD module integrates diverse modality training strategies to alleviate modality dependence, identifying the most appropriate modality combinations in diverse real-world scenarios remains an open problem and warrants further investigation. Second, although AHNS achieves a favorable trade-off between performance and computational cost, it is still difficult to derive clear principles for selecting the optimal loss function combination, and determining the optimal number of negative samples remains challenging. We plan to investigate this further in future work.

## Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly improved this paper. This work is supported by the National Natural Science Foundation of China (62276057).

## References

Ivana Balažević, Carl Allen, and Timothy Hospedales. 2019. 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)*, pages 5185–5194.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems (NeurIPS)*, 26.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in neural information processing systems (NeurIPS)*, 35:39090–39102.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, and 1 others. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *arXiv preprint arXiv:2402.05391*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Yuxiao Gao, Fuwei Zhang, Zhao Zhang, Xiaoshuang Min, and Fuzhen Zhuang. 2025. Mixed-curvature multi-modal knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 11699–11707.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning (ICML)*, pages 9226–9259.

Yue Jian, Xiangyu Luo, Zhifei Li, Miao Zhang, Yan Zhang, Kui Xiao, and Xiaoju Hou. 2025. Apkgc: Noise-enhanced multi-modal knowledge graph completion with attention penalty. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 15005–15013.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Whang. 2023. Vista: Visual-textual knowledge graph representation learning. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 7314–7328.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and 1 others. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. Imf: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM web conference (WWW)*, pages 2572–2580.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *European Semantic Web Conference (ESWC)*, pages 459–474. Springer.

Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the seventh joint conference on lexical and computational semantics (SEM)*, pages 225–234.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yunpeng Wang, Bo Ning, Xin Wang, Chengfei Liu, and Guanyu Li. 2025. Segmentation similarity enhanced semantic related entity fusion for multi-modal knowledge graph completion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1176–1185.

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Image-embodied knowledge representation learning. *arXiv preprint arXiv:1609.07028*.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM international conference on multimedia (MM)*, pages 3857–3866.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025a. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2025b. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 13322–13330.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17120–17130.

# A Detailed Experimental Setup

## A.1 Dataset Statistics

In this paper, we employ three public MKGC benchmarks: DB15K (Liu et al., 2019), MKG-W (Xu et al., 2022), and MKG-Y (Xu et al., 2022) to evaluate the model performance. DB15K is derived from DBPedia (Lehmann et al., 2015) and enhanced with images obtained through a search engine. MKG-W comprises subsets of Wikidata (Vrandečić and Krötzsch, 2014). MKG-Y comprises subsets of YAGO (Suchanek et al., 2007). All feature rich image and text data from official releases. Detailed information about the datasets is presented in Table 9.

| Dataset | Entity | Rel | Train | Valid | Test | Visual | Textual |
|---|---|---|---|---|---|---|---|
| DB15K | 12842 | 279 | 79222 | 9902 | 9904 | 12818 | 12842 |
| MKG-W | 15000 | 169 | 34196 | 4276 | 4274 | 14463 | 14123 |
| MKG-Y | 15000 | 28 | 21310 | 2665 | 2663 | 14244 | 12305 |

Table 9: Dataset statistics.

## A.2 Evaluation Metrics

We use two widely adopted metrics to evaluate our model: Mean Reciprocal Rank (MRR) and Hit Rate (Hits@k, where k takes the values of 1, 3, and 10). MRR is used to calculate the reciprocal of the correct rank for a single query and then takes the average value across all queries. Hits@k is used to calculate the proportion of cases where the correct result appears among the top k candidate answers. It is worth noting that MRR and Hits@1 are more sensitive to the candidate rankings, these two metrics are better at measuring the accurate prediction ability of the model, while Hits@3 and Hits@10 have a more lenient assessment, which better reflects the tolerance rate of the model evaluation.

## A.3 Hyperparameter Settings

In our experiment, we use the PyTorch 2.0.0 framework and complete the code execution on a single A800 GPU. For modality tokenization, we employ the tokenizer of BEIT (Peng et al., 2022) and BERT (Devlin et al., 2019) as our visual/textual tokenizers. The codebook size of BEIT is 8192 and the vocabulary size of BERT tokenizer is 32000. We keep 3 images for each entity with visual information. During training, we set the epoch to 5000, the batch size to 2048, and the embedding dimension to 256. The maximum number of selected visual tokens and textual tokens is set to 16 and 24 respectively.

# B Hyperparameter Analysis

## B.1 Hyperparameter Analysis of Modality Dropout Proportion

To analyze the impact of the dynamically increasing proportion $P_{drop}$ on model performance, we conduct experiments on the parameter $\alpha$ under the condition of fixing the random seed and compared different modal dropout strategies. The experimental results are shown in Table 10.

It can be seen from Table 10 that if the parameter $\alpha$ in $P_{drop}$ is set too high or too low, the model performance will decrease to varying degrees. This indicates that during the training process, the number of modality missing scenarios generated by the

MD module should be neither excessive nor insufficient: an excessive number will weaken the model's learning ability under the condition of complete modalities, while an insufficient number will make it difficult to provide sufficient diversity of missing scenarios.

| Parameter $\alpha$ | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| 0.15 | 39.13 | 32.64 | 41.85 | 50.82 |
| 0.30 | **40.10** | **33.42** | **42.83** | **52.35** |
| 0.45 | 39.56 | 32.94 | 42.20 | 51.74 |

Table 10: The results of different parameters $\alpha$ on the MKG-W dataset.

### B.2 Hyperparameter Analysis of Sampling Proportion and Adjusting Loss

The primary motivation of designing AHNS is to achieve good performance while minimizing sampling overhead. To this end, we conduct an experimental analysis on the key parameters of AHNS and compare it with other negative sampling methods. The results are shown in Table 11.

When the total number of negative samples $N_{neg} = 20$, parameter $\beta = 0.8$, and $\lambda = 0.01$, the overall performance of the model reaches its optimal level.

### B.3 Exploration on Modality Tokenization

To investigate the impact of MD on model performance at different modality information scales, we vary the maximum token numbers of the textual and visual modalities, as shown in Fig. 5.

We conduct a set of experiments and visualize the results with a layered bar chart, where the bottom and top layers correspond to settings without and with MD, respectively. MD brings larger performance gains when the token count is high, as modalities with more tokens contain richer yet noisier information, from which MD can benefit by introducing diverse incomplete-information scenarios. When the token count is small, the information is already compact, limiting the effectiveness of MD and resulting in smaller improvements.

In contrast, when the token count is small, the modality already contains compact and dense information, leaving limited room for MD to generate diverse missing-information patterns. As a result, the performance improvement becomes smaller.

| Setting | MKG-W | | | |
|---|---|---|---|---|
| | MRR | Hits@1 | Hits@3 | Hits@10 |
| *Number of samples* | | | | |
| $N_{neg}$ = 5 | 39.71 | 33.08 | 42.44 | 52.12 |
| $N_{neg}$ = 10 | 39.82 | 33.26 | 42.29 | 52.16 |
| $N_{neg}$ = 20 | **40.10** | **33.42** | **42.83** | 52.35 |
| $N_{neg}$ = 30 | 39.71 | 33.28 | 42.14 | 51.60 |
| $N_{neg}$ = 40 | 39.61 | 32.86 | 42.37 | **52.37** |
| *Parameter $\beta$ in $P_{neg}$* | | | | |
| $\beta$ = 0.4 | 39.53 | 33.00 | 42.21 | 51.32 |
| $\beta$ = 0.6 | 39.95 | 33.37 | 42.63 | 52.30 |
| $\beta$ = 0.8 | **40.10** | **33.42** | **42.83** | **52.35** |
| *Parameter $\lambda$ in Loss L* | | | | |
| $\lambda$ = 0.1 | 39.16 | 32.74 | 41.41 | 51.32 |
| $\lambda$ = 0.01 | **40.10** | **33.42** | **42.83** | **52.35** |
| $\lambda$ = 0.001 | 39.49 | 32.99 | 42.06 | 51.16 |

Table 11: Hyperparameter analysis of sampling proportion and adjusting loss on MKG-W dataset.

## C Case Study

In this case study, we consider the query (Taylor Swift, Educated_At, ?) as shown in Table 12.

For visual modality: the visual content related to Taylor Swift is dominated by images of her luxury residences in Los Angeles, Hollywood red carpet appearances, and California-style street photography. Conventional visual encoders (e.g., BEIT) tend to extract highly frequent yet irrelevant cues such as California, sunshine, and Los Angeles–style architecture.

For textual modality: textual descriptions primarily focus on her musical achievements, Grammy awards, and the fact that she mainly resides in California (Beverly Hills) and Nashville. These descriptions further reinforce strong geographic associations unrelated to her educational background.

For baseline: due to the high frequency and salience of visual and textual signals, conventional fusion-based baseline models overemphasize these modalities and mistakenly associate geographic information with educational relations. As a result, they incorrectly predict institutions such as the University of Southern California (USC) or the University of California, Los Angeles (UCLA).

For structural evidence and MD mechanism: in contrast, the knowledge graph contains a sparse but accurate structural reasoning path: (Taylor Swift, Received_Honorary_Degree, Doctor of Fine

| Query | Multi-modal Inputs (Features) | Reasoning Process | Method | Prediction |
|---|---|---|---|---|
| **Head Entity:** Taylor Swift<br><br>**Relation:** Educated_At<br><br>**Target:** **?** | **Visual Modality ($e_{vis}$):** Images of her residence in ***Beverly Hills, California***; Hollywood red carpet photos; Concert stages; ..... → *Noise: Strongly suggests "California" context.* | **Baseline Logic (50% visual, 35% textual, and 15% structural):** The model fuses features directly. The high frequency of "California" visual tokens and "Los Angeles" text tokens overwhelms the embedding space. The model infers education location based on residence location. High modality dominance caused by abundant textual and visual information. | **Baseline** | ✗USC |
| | **Textual Modality ($e_{txt}$):** "Taylor Swift is an American singer-songwriter and a dominant figure in pop culture. She maintains a high-profile lifestyle with primary residences in Nashville and Beverly Hills, California, frequently attending Hollywood industry events. As a global superstar, her influence spans music, cinema, and fashion....." → *Noise: Focuses on career & location, lacks academic details.* | **MDBGF Logic (55% structural, 45% visual or textual):** 1. **Modality Dropout:** During training, the model learned to predict without visual cues, reducing reliance on background scenery. 2. **Structure Priority:** The structural embedding captures the triple **(Taylor Swift, Received_Honorary_Degree, NYU)**. The **BGF** module gates out the conflicting visual noise and highlights this structural link. | **MDBGF** | ✓NYU |

Table 12: Case study comparison: resolving modality dependence with MDBGF, which demonstrates how our proposed method corrects false predictions caused by misleading visual and textual noise (e.g., location bias) by leveraging structural priors.

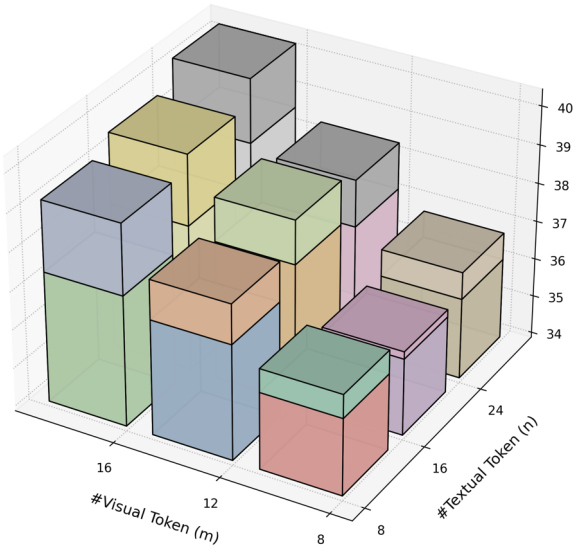


Figure 5: Performance improvement of the MD module under different numbers of modality tokens.

Arts)→(Doctor of Fine Arts, awarded_by, New York University). During training, the MD mechanism randomly drops high-noise visual and textual modalities, forcing the model in certain epochs to rely solely on structural embeddings. This process encourages the model to avoid over-trusting visually dominant but semantically irrelevant cues.

For BGF effect at inference: at inference time, although visual features still strongly point to California, the BGF mechanism detects a semantic mismatch between visual content (e.g., concert and lifestyle images) and the queried educational relation. Consequently, it suppresses the visual modality, amplifies the structural signal, and ultimately enables MDBGF to correctly predict New York University (NYU).