Noise-powered Multi-modal Knowledge Graph Representation Framework

Anonymous ACL submission

Abstract

The rise of Multi-modal Pre-training highlights the necessity for a unified Multi-Modal Knowledge Graph (MMKG) representation learning framework. Such a framework is essential for embedding structured knowledge into multi-modal Large Language Models effectively, alleviating issues like knowledge misconceptions and multi-modal hallucinations. In this work, we explore the efficacy of models in accurately embedding entities within MMKGs through two pivotal tasks: Multi-modal Knowledge Graph Completion (MKGC) and Multi-modal Entity Alignment (MMEA). Building on this foundation, we propose a novel **SNAG** method that utilizes a Transformer-based architecture equipped with modality-level noise masking to robustly integrate multi-modal entity features in KGs. By incorporating specific training objectives for both MKGC and MMEA, our approach achieves SOTA performance across a total of ten datasets, demonstrating its versatility. Moreover, **SNAG** can not only function as a standalone model but also enhance other existing methods, providing stable performance improvements. Code and data are available at https://anonymous.4open.science/r/SNAG.

1 Introduction

Current efforts to integrate MMKG with pre-training are scarce. **Triple-level** methods (Pan et al., 2022) treat triples as standalone knowledge units, embedding the *(head entity, relationship, tail entity)* structure into Visual Language Model's space. On the other hand, **Graph-level** methods (Gong et al., 2023; Li et al., 2023b) capitalize on the structural connections among entities in a global MMKG. By selectively gathering multi-modal neighbor nodes around each entity featured in the training corpus, they apply techniques such as Graph Neural Networks (GNNs) or concatenation to effectively incorporate knowledge during the pre-training process.

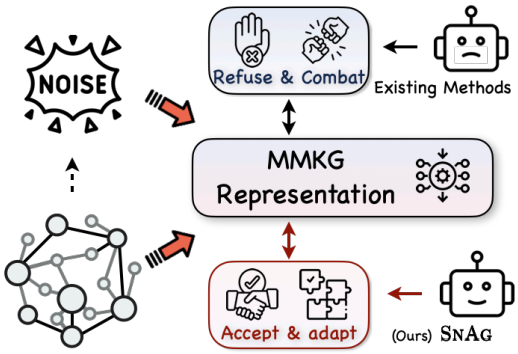

Figure 1: Unlike existing models that refuse and combat noise in MMKGs, our SNAG accepts and deliberately incorporates noise to mirror noisy real-world scenarios.

However, these approaches predominantly view MMKG from a traditional KG perspective, not fully separating the MMKG representation process from downstream or pre-training tasks. In this work, we revisit MMKG representation learning uniquely from the MMKG perspective itself, employing two tasks: Multi-modal Knowledge Graph Completion (MKGC) and Multi-modal Entity Alignment (MMEA) to validate our method.

Specifically, we introduce a unified Transformer-based framework (SNAG) that achieves SOTA results across an array of ten datasets by simply aligning it with task-specific Training targets. SNAG stands out for its **parameter-efficient** design and **adaptability**, incorporating components like Entity-Level Modality Interaction that can be seamlessly upgraded with advanced technologies.

A key aspect of our method is the Gauss Modality Noise Masking module, whose design sharply contrasts with previous MMKG-related efforts that primarily focus on designing methods to refuse and combat noise in MMKGs. In contrast, as shown in Fig. 1, our SNAG accepts and deliberately incorporates noise, adapting to the noisy real-world scenarios. Drawing inspiration from traditional mask-based multi-modal Pre-trained Language Models (PLMs) that enhance cross-modal alignment at the token level, our strategy innovates by **applying masking at the modality level**, signif-

icantly enhancing model's MMKG representation capabilities. Importantly, as the first MMKG effort to concurrently support both MKGC and MMEA tasks, this work demonstrates its adaptability of our strategy, highlighting its potential to interface with more training tasks in the future and paving the way for further research in MMKG Pre-training and Multi-modal Knowledge Injection.

2 Related Work

2.1 MMKG Representation

The current mainstream approaches to MMKG representation learning can broadly be classified into two distinct categories: ***(i)* Late Fusion** methods emphasize modality interactions and feature aggregation just prior to output generation. For example, MKGRL-MS (Wang et al., 2022) crafts unique single-modal embeddings, employing multi-head self-attention to determine each modality's contribution to semantic composition and **sum** the weighted multi-modal features for MMKG entity representation. MMKRL (Lu et al., 2022) learns cross-modal embeddings in a unified translational semantic space, merging them through **concatenation**. DuMF (Li et al., 2022) applies a bilinear layer for feature projection and an attention block for modality preference learning in each track, integrating features via a **gate network**. ***(ii)* Early Fusion** methods integrate multi-modal feature at an initial stage, enabling full modality interactions for complex reasoning. For example, Fang et al. (2023a) first normalizes entity modalities into a unified embedding using an MLP, then refines them by contrasting with perturbed negative samples. MMRotatH (Wei et al., 2023) utilizes a gated encoder to merge textual and structural data, filtering irrelevant information within a rotational dynamics-based KGE framework. Recent studies (Chen et al., 2022b; Lee et al., 2023) utilize (V)PLMs like BERT and ViT for multi-modal data integration. These methods convert graph structures, text, and images into sequences or dense embeddings suited for LMs, leveraging the LMs' reasoning abilities and embedded knowledge for tasks like Multi-modal Link Prediction. However, they rely heavily on pre-trained models, resulting in significant parameter sizes and training costs.

In this paper, we propose a Transformer-based method **SNAG** that introduces fine-grained, entity-level modality preference to enhance entity representation. This strategy combines the benefits of Early Fusion, with its effective modality interaction, while also aligning with the Late Fusion modality integration paradigm. Furthermore, our model is lightweight, with **only 13M parameters**, far fewer than traditional PLM-based methods, which often exceed **200M parameters**. This offers increased flexibility and wider applicability.

2.2 Multi-Modal Knowledge Graph Completion and Alignment

Multi-modal Knowledge Graph Completion (MKGC) is crucial for inferring missing triples in existing MMKGs (Lee et al., 2023; Zhao et al., 2022). Entity Alignment (EA) focuses on KG integration, aiming to identify identical entities across different KGs by leveraging relational, attributive, and literal (surface) features. Multi-Modal Entity Alignment (MMEA) enhances this by incorporating visual data, thereby improving alignment accuracy (Chen et al., 2020a; Li et al., 2023a). Further details are provided in Appendix A.1. Despite nearly five years of development, MMEA and MKGC have progressed independently within the MMKG field, lacking a unified framework that integrates these tasks. Given the advancements in multi-modal LLMs, it is timely to develop a comprehensive framework that addresses both MKGC and MMEA, offering enhanced multi-modal entity representations.

3 Method

3.1 Preliminaries

This paper focuses on A-MMKG (Zhu et al., 2022), where images are attached to entities as attributes.

Definition 1. ***Multi-modal Knowledge Graph.*** *A KG is defined as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{T}, \mathcal{V}\}$ where $\mathcal{T} = \{\mathcal{T}_{\mathcal{A}}, \mathcal{T}_{\mathcal{R}}\}$ with $\mathcal{T}_{\mathcal{R}} = \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ and $\mathcal{T}_{\mathcal{A}} = \mathcal{E} \times \mathcal{A} \times \mathcal{V}$. MMKG utilizes multi-modal data (e.g., images) as specific attribute values for entities or concepts, with $\mathcal{T}_{\mathcal{A}} = \mathcal{E} \times \mathcal{A} \times (\mathcal{V}_{KG} \cup \mathcal{V}_{MM})$, where $\mathcal{V}_{MM}$ are values of multi-modal data (e.g., images).*

Definition 2. ***MMKG Completion.*** *The objective is to augment the set of relational triples $\mathcal{T}_R$ within MMKGs by identifying and adding missing relational triples among existing entities and relations, potentially utilizing attribute triples $\mathcal{T}_{\mathcal{A}}$. Specifically, our focus is on Entity Prediction, which involves determining the missing head or tail entities in queries of the form $(head, r, ?)$ or $(?, r, tail)$.*

Definition 3. ***Multi-modal Entity Alignment.*** *Given two aligned MMKGs $\mathcal{G}_1$ and $\mathcal{G}_2$, the objec-*

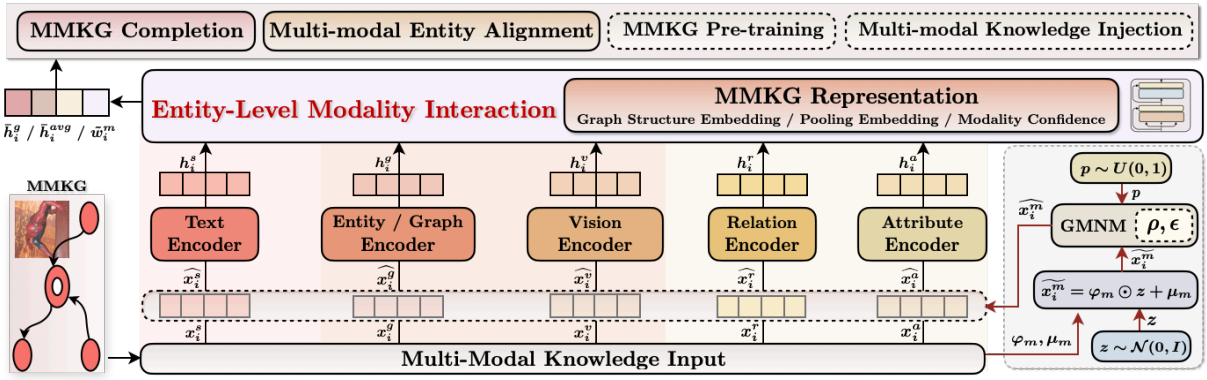

Figure 2: The overall framework of SNAG.

tive of MMEA is to identify entity pairs (e_i^1, e_i^2) *from* $\mathcal{E}_1$ *and* $\mathcal{E}_2$*, respectively, that correspond to the same real-world entity* e_i*. This process utilizes a set of pre-aligned entity pairs, divided into a training set (seed alignments* $\mathcal{S}$*) and a testing set* $\mathcal{S}_{te}$*, following a pre-defined seed alignment ratio* $R_{sa} = |\mathcal{S}|/|\mathcal{S} \cup \mathcal{S}_{te}|$*. The modalities associated with an entity are denoted by* $\mathcal{M} = \{g, r, a, v, s\}$*, signifying graph structure, relation, attribute, vision, and surface (i.e., entity names) modalities, respectively.*

3.2 Multi-Modal Knowledge Embedding

Graph Structure Embedding. Let $x_i^g \in \mathbb{R}^d$ represents the graph embedding of entity e_i, which is randomly initialized and learnable, with d representing the predetermined hidden dimension. In MKGC, we follow Zhang et al. (2024) to set $h_i^g = FC_g(W_g, x_i^g)$, where FC_g is a KG-specific fully connected layer applied to x_i^g with weights W_g. For MMEA, we follow Chen et al. (2023a) to utilize the Graph Attention Network (GAT) (Velickovic et al., 2018), configured with two attention heads and two layers, to capture the structural information of $\mathcal{G}$. This is facilitated by a diagonal weight matrix (Yang et al., 2015) $W_g \in \mathbb{R}^{d \times d}$ for linear transformation. The structure embedding is thus defined as $h_i^g = GAT(W_g, M_g; x_i^g)$, where M_g refers to the graph's adjacency matrix.

Relation and Attribute Embedding. Our MKGC study aligns with domain practices (Zhao et al., 2022; Li et al., 2023c) which focuses exclusively on relation triples. These are represented by learnable embeddings $x_j^r \in \mathbb{R}^{d/2}$, where j uniquely identifies each relation r_j, distinguishing it from entity indices. We exclude attribute triples to maintain consistency with methodological practices in the field. The choice of setting a dimensionality of $d/2$ is based on our application of the RotatE model (Sun et al., 2019), which assesses triple plausibility. RotatE interprets relations as rotations in a complex space, requiring the relation embedding's dimension to be half that of the entity embedding to account for the real and imaginary components of complex numbers. For MMEA, following Yang et al. (2019), we use bag-of-words features for relation (x^r) and attribute (x^a) representations of entities (detailed in § 4) . Separate FC layers, parameterized by $W_m \in \mathbb{R}^{d_m \times d}$, are employed for embedding space harmonization: $h_i^m = FC_m(W_m, x_i^m)$, where $m \in \{r, a\}$ and $x_i^m \in \mathbb{R}^{d_m}$ represents the input feature of entity e_i for modality m.

Visual and Surface Embedding. For visual embeddings, a pre-trained (and thereafter frozen) visual encoder, denoted as Enc_v, is used to extract visual features x_i^v for each entity e_i with associated image data. In cases where entities lack corresponding image data, we synthesize random image features adhering to a normal distribution, parameterized by the mean and standard deviation observed across other entities' images (Chen et al., 2023a,b; Zhang et al., 2024). Regarding surface embeddings for MKGC, we leverage Sentence-BERT (Reimers and Gurevych, 2019), a pre-trained textual encoder, to derive textual features from each entity's description. The `[CLS]` token serves to aggregate sentence-level textual features x_i^s. Consistent with the approach applied to other modalities, we utilize FC_m parameterized by $W_m \in \mathbb{R}^{d_m \times d}$ to integrate the extracted features x_i^v and x_i^s into the embedding space, yielding the embeddings h_i^v and h_i^s.

3.3 Gauss Modality Noise Masking

Recent research in MMKG (Chen et al., 2023b; Guo et al., 2023) suggests that models can tolerate certain noise levels without a noticeable decline in the expressive capability of multi-modal entity representations, a finding echoed across various machine learning domains (Jain et al., 2023; Chen et al., 2024). Additionally, Chen et al. (2023c)

demonstrate that cross-modal masking and reconstruction can improve a model's cross-modal alignment capabilities in Zero-shot Image Classification scenario. Inspired by evidence of model noise resilience, we hypothesize that introducing noise during MMKG modality fusion training could enhance both modal feature robustness and real-world performance. In light of these observations, we propose a new mechanism termed Gauss Modality Noise Masking (GMNM), aimed at enhancing modality feature representations through controlled noise injection at the training stage for MMKG. This stochastic strategy introduces a probabilistic transformation to each modality feature x_i^m at the beginning of every training epoch, described as :

$$\widehat{x_i^m} = \begin{cases} x_i^m, & \text{if } p > \rho\,, \\ (1-\epsilon)x_i^m + \epsilon\widetilde{x_i^m}\,, & \text{otherwise,} \end{cases} \quad (1)$$

where $p \sim U(0,1)$ denotes a uniformly distributed random variable that determines whether noise is applied, with ρ being the threshold probability for noise application to each x_i^m. Here, ϵ signifies the noise (mask) ratio. We define the generation of noise vector $\widetilde{x_i^m}$ as:

$$\widetilde{x_i^m} = \varphi_m \odot z + \mu_m\,, \quad z \sim \mathcal{N}(0, I), \quad (2)$$

where φ_m and μ_m represent the standard deviation and mean of the **modality-specific non-noisy data** for m, respectively, and z denotes a sample drawn from a Gaussian distribution $\mathcal{N}(0, I)$ with mean vector with mean 0 and identity covariance matrix I, ensuring that the introduced noise is statistically coherent with the intrinsic data variability of the respective modality. Additionally, the intensity of noise (ϵ) can be dynamically adjusted to simulate real-world data imperfections. This adaptive noise injection strategy is designed to foster a model resilient to data variability, capable of capturing and representing complex multi-modal interactions with enhanced fidelity in practical applications.

Note that after the transformation from x^m to $\widehat{x^m}$, these modified features are still subject to further processing through FC_m as detailed in § 3.2. This critical step secures the generation of the ultimate modal representation, symbolized as $\widehat{h^m}$. For clarity in subsequent sections, **we will treat h^m and h_i^m as representing their final states, $\widehat{h^m}$ and $\widehat{h_i^m}$**, unless specified otherwise.

3.4 Entity-Level Modality Interaction

This phase is designed for instance-level modality weighting and fusion, enabling dynamic adjustment of training weights based on modality information's signal strength and noise-induced uncertainty. We utilize a Transformer architecture (Vaswani et al., 2017) for this purpose, noted for its efficacy in modality fusion and its ability to derive confidence-based weighting for modalitieswhich improves interpretability and adaptability. The Transformer's self-attention mechanism is crucial for ensuring the model evaluates and prioritizes accurate and relevant modal inputs.

Specifically, we adapt the vanilla Transformer through integrating three key components: Multi-Head Cross-Modal Attention (MHCA), Fully Connected Feed-Forward Networks (FFN), and Instance-level Confidence (ILC).

(i) **MHCA** operates its attention function across N_h parallel heads. Each head, indexed by i, employs shared matrices $W_q^{(i)}, W_k^{(i)}, W_v^{(i)} \in \mathbb{R}^{d\times d_h}$ (where $d_h = d/N_h$), to transform input h^m into queries $Q_m^{(i)}$, keys $K_m^{(i)}$, and values $V_m^{(i)}$:

$$Q_m^{(i)}, K_m^{(i)}, V_m^{(i)} = h^m W_q^{(i)}, h^m W_k^{(i)}, h^m W_v^{(i)}\,.$$

The output for modality m's feature is then generated by combining the outputs from all heads and applying a linear transformation:

$$MHCA(h^m) = \bigoplus_{i=1}^{N_h} head_i^m \cdot W_0\,, \quad (3)$$

$$head_i^m = \sum\nolimits_{j\in\mathcal{M}} \beta_{mj}^{(i)} V_j^{(i)}\,, \quad (4)$$

where $W_0 \in \mathbb{R}^{d\times d}$. The attention weight β_{mj} calculates the relevance between modalities m, j:

$$\beta_{mj} = \frac{\exp(Q_m^\top K_j/\sqrt{d_h})}{\sum_{i\in\mathcal{M}}\exp(Q_m^\top K_i/\sqrt{d_h})}\,. \quad (5)$$

Besides, layer normalization (LN) and residual connection (RC) are incorporated to stabilize training:

$$\bar{h}^m = LayerNorm(MHCA(h^m) + h^m)\,. \quad (6)$$

(ii) **FFN**: This network, consisting of two linear transformations and a ReLU activation, further processes the MHCA output:

$$FFN(\bar{h}^m) = ReLU(\bar{h}^m W_1 + b_1)W_2 + b_2\,,$$

$$\bar{h}^m \leftarrow LayerNorm(FFN(\bar{h}^m) + \bar{h}^m)\,,$$

where $W_1 \in \mathbb{R}^{d\times d_{in}}$ and $W_2 \in \mathbb{R}^{d_{in}\times d}$.

(iii) **ILC**: To capture crucial inter-modal interactions and tailors the model's confidence for each entity's modality, we calculate the confidence $\tilde{w}^m$:

$$\tilde{w}^m = \frac{\exp(\sum_{j\in\mathcal{M}}\sum_{i=0}^{N_h}\beta_{mj}^{(i)}/\sqrt{|\mathcal{M}|\times N_h})}{\sum_{k\in\mathcal{M}}\exp(\sum_{j\in\mathcal{M}}\sum_{i=0}^{N_h}\beta_{kj}^{(i)}\sqrt{|\mathcal{M}|\times N_h})}\,. \quad (7)$$

3.5 Task-Specific Training

Building upon the foundational processes detailed in previous sections, we have derived multi-modal KG representations denoted as h^m (discussed in § 3.3) and $\bar{h}^m$ (elaborated in § 3.4, Eq. (6)), along with confidence scores $\tilde{w}^m$ for each modality m within the MMKG (introduced in § 3.4, Eq. (7)).

MMKG Completion. Within MKGC, we consider two methods for entity representation as candidates: ***(i)*** $\bar{h}^g$: Reflecting insights from previous research (Chen et al., 2023a; Zhang et al., 2024), graph structure embedding emerges as crucial for model performance. After being processed by the Transformer layer, $\bar{h}^g$ not only maintains its structural essence but also blends in other modal insights (refer to Eq. (3) and (4)), offering a comprehensive multi-modal entity representation. ***(ii)*** $\bar{h}^{avg}$: For an equitable multi-modal representation, we average all modality-specific representations via $\bar{h}^{avg} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \bar{h}^m$, where $\mathcal{M}$ is the set of all modalities. This averaging ensures equal modality contribution, leveraging the rich, diverse information within MMKGs. For consistency in the following descriptions, we will refer to both entity representations using the notation $\bar{h}$.

We apply the RotatE model (Sun et al., 2019) as our score function to assess the plausibility of triples. It is defined as:

$$\mathcal{F}(e^h, r, e^t) = ||\bar{h}^{head} \circ x^r - \bar{h}^{tail}||\,, \quad (8)$$

where $\circ$ represents the rotation operation in complex space, which transforms the head entity's embedding by the relation to approximate the tail entity's embedding.

To prioritize positive triples with higher scores, we optimize the embeddings using a sigmoid-based loss function (Sun et al., 2019). The loss function is given by:

$$\mathcal{L}_{kgc} = \frac{1}{|\mathcal{T}_\mathcal{R}|}\sum_{(e^h,r,e^t)\in\mathcal{T}_\mathcal{R}} \Big(-\log\sigma(\lambda - \mathcal{F}(e^h, r, e^t)) - \sum\nolimits_{i=1}^{K} \upsilon_i \log\sigma(\mathcal{F}(e^{h\prime}, r', e^{t\prime}) - \lambda)\Big)\,,$$

where σ denotes the sigmoid function, λ is the margin, K is the number of negative samples per positive triple, and υ_i represents the self-adversarial weight for each negatively sampled triple $(e^{h\prime}, r', e^{t\prime})$. Concretely, υ_i is calculated as:

$$\upsilon_i = \frac{\exp(\tau_{kgc}\mathcal{F}(e_i^{h\prime}, r_i', e_i^{t\prime}))}{\sum_{j=1}^{K}\exp(\tau_{kgc}\mathcal{F}(e_j^{h\prime}, r_j', e_j^{t\prime}))}\,, \quad (9)$$

with τ_{kgc} being the temperature parameter. Our primary objective is to minimize $\mathcal{L}_{kgc}$, thereby refining the embeddings to accurately capture MMKG's underlying relationships.

Multi-modal Entity Alignment. In MMEA, following (Chen et al., 2023b,a), we adopt the Global Modality Integration (GMI) derived multi-modal features as the representations for entities. GMI emphasizes global alignment by concatenating and aligning multi-modal embeddings with a learnable global weight, enabling adaptive learning of each modality's quality across two MMKGs. The GMI joint embedding h_i^{GMI} for entity e_i is defined as:

$$h_i^{GMI} = \bigoplus\nolimits_{m\in\mathcal{M}}[w_m h_i^m]\,, \quad (10)$$

where $\bigoplus$ signifies vector concatenation and w_m is the global weight for modality m, which is distinct from the entity-level dynamic modality weights $\tilde{w}^m$ in Eq. (7).

We note that the distinction between MMEA and MKGC lies in their focus: MMEA emphasizes aligning modal features between entities and distinguishing non-aligned entities, prioritizing original feature retention. In contrast, MKGC emphasizes the inferential benefits of modality fusion across different multi-modal entities. As demonstrated by Chen et al. (2023b), the modality feature is often smoothed by the Transformer Layer in MMEA, potentially reducing entity distinction. GMI addresses this by preserving essential information, aiding alignment stability.

Moreover, as a unified MMKG representation framework, modal features extracted earlier are optimized through MMEA-specific training objectives (Lin et al., 2022). Specifically, for each aligned entity pair (e_i^1,e_i^2) in training set (seed alignments $\mathcal{S}$), we define a negative entity set $\mathcal{N}_i^{ng} = \{e_j^1|\forall e_j^1 \in \mathcal{E}_1, j \neq i\} \cup \{e_j^2|\forall e_j^2 \in \mathcal{E}_2, j \neq i\}$ and utilize in-batch ($\mathcal{B}$) negative sampling (Chen et al., 2020b) to enhance efficiency. The alignment probability distribution is:

$$p_m(e_i^1, e_i^2) = \frac{\gamma_m(e_i^1, e_i^2)}{\gamma_m(e_i^1, e_i^2) + \sum_{e_j\in\mathcal{N}_i^{ng}}\gamma_m(e_i^1, e_j)}\,,$$

where $\gamma_m(e_i, e_j) = \exp({h_i^m}^\top h_j^m/\tau_{ea})$ and τ_{ea} is the temperature hyper-parameter. We establish a bi-directional alignment objective to account for MMEA directions:

$$\mathcal{L}_m = -\mathbb{E}_{i\in\mathcal{B}}\,\log[\,p_m(e_i^1, e_i^2) + p_m(e_i^2, e_i^1)\,]/2\,,$$

Table 1: MKGC performance on DB15K (Liu et al., 2019), MKG-W and MKG-Y (Xu et al., 2022) datasets. The best results are highlighted in **bold**, and the third-best results are underlined for each column.

Models	DB15K (Liu et al., 2019)				MKG-W (Xu et al., 2022)				MKG-Y (Xu et al., 2022)			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
IKRL (IJCAI '17) (Xie et al., 2017a)	.268	.141	.349	.491	.324	.261	.348	.441	.332	.304	.343	.383
TBKGC (NAACL '18) (Sergieh et al., 2018)	.284	.156	.370	.499	.315	.253	.340	.432	.340	.305	.353	.401
TransAE (IJCNN '19) (Wang et al., 2019)	.281	.213	.312	.412	.300	.212	.349	.447	.281	.253	.291	.330
RSME (ACM MM '21) (Wang et al., 2021)	.298	.242	.321	.403	.292	.234	.320	.404	.344	.318	.361	.391
VBKGC (KDD '22) (Zhang and Zhang, 2022)	.306	.198	.372	.494	.306	.249	.330	.409	.370	.338	.388	.423
OTKGE (NeurIPS '22) (Cao et al., 2022)	.239	.185	.259	.342	.344	.289	.363	.449	.355	.320	.372	.414
IMF (WWW '23) (Li et al., 2023c)	.323	.242	.360	.482	.345	.288	.366	.454	.358	.330	.371	.406
QEB (ACM MM '23) (Wang et al., 2023)	.282	.148	.367	.516	.324	.255	.351	.453	.344	.295	.370	.423
VISTA (EMNLP '23) (Lee et al., 2023)	.304	.225	.336	.459	.329	.261	.354	.456	.305	.249	.324	.415
MANS (IJCNN '23) (Zhang et al., 2023)	.288	.169	.366	.493	.309	.249	.336	.418	.290	.253	.314	.345
MMRNS (ACM MM '22) (Xu et al., 2022)	.297	.179	.367	.510	.341	.274	.375	.468	.359	.306	.391	.455
AdaMF (COLING '24) (Zhang et al., 2024)	.325	.213	.397	.517	.343	.272	.379	.472	.381	.335	.404	.455
SNAG (Ours)	**.363**	**.274**	**.411**	**.530**	**.373**	**.302**	**.405**	**.503**	**.395**	**.354**	**.411**	**.471**
- w/o GMNM	.357	.269	.406	.523	.365	.296	.398	.490	.387	.345	.407	.457

(i) The training objective denoted as $\mathcal{L}_{GMI}$ when using GMI joint embeddings, i.e., $\gamma_{GMI}(e_i, e_j)$ is set to $\exp(h_i^{GMI\top} h_j^{GMI}/\tau_{ea})$.

To integrate dynamic confidences into the training process and enhance multi-modal entity alignment, we adopt two specialized training objectives from Chen et al. (2023b): ***(ii)*** Explicit Confidence-augmented Intra-modal Alignment (ECIA): This objective modifies Eq. (3.5) to incorporate explicit confidence levels within the same modality, defined as: $\mathcal{L}_{ECIA} = \sum_{m\in\mathcal{M}} \widetilde{\mathcal{L}}_m$, where:

$$\widetilde{\mathcal{L}}_m = -\mathbb{E}_{i\in\mathcal{B}} \log[\,\phi_m(e_i^1, e_i^2) * (p_m(e_i^1, e_i^2) + p_m(e_i^2, e_i^1))\,]/2\,.$$

Here, $\phi_m(e_i^1, e_i^2)$ represents the minimum confidence value between entities e_i^1 and e_i^2 in modality m, i.e., $\phi_m(e_i, e_j) = Min(\tilde{w}_i^m, \tilde{w}_j^m)$, addressing the issue of aligning high-quality features with potentially lower-quality ones or noise. ***(iii)*** Implicit Inter-modal Refinement (IIR) refines entity-level modality alignment by leveraging the transformer layer outputs $\bar{h}^m$, aiming to align output hidden states directly and adjust attention scores adaptively. The corresponding loss function is: $\mathcal{L}_{IIR} = \sum_{m\in\mathcal{M}} \bar{\mathcal{L}}_m$, where $\bar{\mathcal{L}}_m$ is also a variant of $\mathcal{L}_m$ (Eq. (3.5)) with $\bar{\gamma}_m(e_i, e_j) = \exp(\bar{h}_i^{m\top} \bar{h}_j^m/\tau_{ea})$.

The comprehensive training objective is formulated as: $\mathcal{L}_{ea} = \mathcal{L}_{GMI} + \mathcal{L}_{ECIA} + \mathcal{L}_{IIR}$. Note that our **SNAG** framework can not only function as a standalone model but also enhance other existing methods, providing stable performance improvements in MMEA, as demonstrated in Table 2.

4 Experiments Setup

Iterative Training for MMEA. We employ a probation technique for iterative training, which acts as a buffering mechanism, temporarily storing a cache of mutual nearest entity pairs across KGs from the testing set (Lin et al., 2022). Specifically, at every K_e (where $K_e = 5$) epochs, models identify and add mutual nearest neighbor entity pairs from different KGs to a candidate list $\mathcal{N}^{cd}$. An entity pair in $\mathcal{N}^{cd}$ is then added to the training set if it continues to be mutual nearest neighbors for K_s ($= 10$) consecutive iterations. This iterative expansion of the training dataset serves as data augmentation in the EA domain, enabling further evaluation of the model's robustness across various scenarios.

Implementation Details. **MKGC:** ***(i)*** Following Zhang et al. (2024), vision encoders Enc_v are configured with VGG (Simonyan and Zisserman, 2015) for DBP15K, and BEiT (Bao et al., 2022) for MKG-W and MKG-Y. For entities associated with multiple images, the feature vectors of these images are averaged to obtain a singular representation. ***(ii)*** The head number N_h in MHCA is set to 2. For entity representation in DBP15K, graph structure embedding $\bar{h}^g$ is used, while for MKG-W and MKG-Y, mean pooling across modality-specific representations $\bar{h}^{avg}$ is employed. This distinction is made due to DBP15K's denser KG and greater absence of modality information compared to MKG-W and MKG-Y. ***(iii)*** We simply selected a set of candidate parameters in AdaMF (Zhang et al., 2024). Specifically, the number of negative samples K per positive triple is 32, the hidden dimension d is 256, the training batch size is 1024, the margin λ is 12, the temperature τ_{kgc} is 2.0, and the learning rate is set to $1e-4$. No extensive parameter tuning was conducted; theoretically, SNAG could achieve better performance with parameter optimization. ***(iv)*** The probability ρ of applying noise in GMNM is set at 0.2, with a noise ratio ϵ of 0.7. ***(v)*** For fairness in comparison, we excluded Ensemble-methods like MoSE (Zhao et al., 2022) and PLM-based methods

Table 2: Non-iterative MMEA results across three degrees of visual modality missing. Results are underlined when the baseline, equipped with the Gauss Modality Noise Masking (GMNM) module, surpasses its own original performance, and highlighted in **bold** when achieving SOTA performance.

	Models	R_{img}=0.4			R_{img}=0.6			Standard		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
DBP15K$_{ZH-EN}$	EVA	.623	.876	.715	.625	.877	.717	.683	.906	.762
	w/ GMNM	.629	.883	.724	.625	.881	.717	.680	.907	.760
	MCLEA	.627	.880	.715	.670	.899	.751	.732	.926	.801
	w/ GMNM	.652	.895	.740	.699	.912	.775	.754	.933	.819
	MEAformer	.678	.924	.766	.720	.938	.798	.776	.953	.840
	w/ GMNM	.680	.925	.767	.719	.939	.798	.777	.955	.841
	SNAG (Ours)	**.735**	**.945**	**.812**	**.757**	**.953**	**.830**	**.798**	**.963**	**.858**
DBP15K$_{JA-EN}$	EVA	.546	.829	.644	.552	.829	.647	.587	.851	.678
	w/ GMNM	.618	.876	.709	.625	.874	.714	.664	.902	.748
	MCLEA	.568	.848	.665	.639	.882	.723	.678	.897	.755
	w/ GMNM	.659	.901	.745	.723	.924	.795	.752	.935	.818
	MEAformer	.677	.933	.768	.736	.953	.815	.767	.959	.837
	w/ GMNM	.678	.937	.770	.738	.953	.816	.767	.958	.837
	SNAG (Ours)	**.735**	**.952**	**.814**	**.771**	**.961**	**.841**	**.795**	**.963**	**.857**
DBP15K$_{FR-EN}$	EVA	.622	.895	.719	.634	.899	.728	.686	.926	.771
	w/ GMNM	.628	.897	.725	.634	.900	.728	.686	.929	.772
	MCLEA	.622	.892	.722	.694	.915	.774	.734	.926	.805
	w/ GMNM	.663	.916	.756	.726	.934	.802	.759	.942	.827
	MEAformer	.676	.944	.774	.734	.958	.816	.776	.967	.846
	w/ GMNM	.678	.946	.776	.735	.965	.819	.779	.969	.849
	SNAG (Ours)	**.757**	**.963**	**.835**	**.790**	**.970**	**.858**	**.814**	**.974**	**.875**
OpenEA$_{EN-FR}$	EVA	.532	.830	.635	.553	.835	.652	.784	.931	.836
	w/ GMNM	.537	.829	.638	.554	.833	.652	.787	.935	.839
	MCLEA	.535	.842	.641	.607	.858	.696	.821	.945	.866
	w/ GMNM	.554	.848	.658	.624	.873	.714	.830	.950	.874
	MEAformer	.582	.891	.690	.645	.904	.737	.846	.862	.889
	w/ GMNM	.588	.895	.696	.647	.905	.738	.847	.963	.890
	SNAG (Ours)	**.621**	**.905**	**.721**	**.667**	**.922**	**.757**	**.848**	**.964**	**.891**
OpenEA$_{EN-DE}$	EVA	.718	.918	.789	.734	.921	.800	.922	.982	.945
	w/ GMNM	.728	.919	.794	.740	.921	.803	.923	.983	.946
	MCLEA	.702	.910	.774	.748	.912	.805	.940	.988	.957
	w/ GMNM	.711	.912	.782	.762	.928	.821	.942	.990	.960
	MEAformer	.749	.938	.816	.789	.951	.847	.955	.994	.971
	w/ GMNM	.753	.939	.817	.791	.952	.848	.957	.995	.971
	SNAG (Ours)	**.776**	**.948**	**.837**	**.810**	**.958**	**.862**	**.958**	.995	**.972**
OpenEA$_{D-W-V1}$	EVA	.567	.796	.651	.592	.810	.671	.859	.945	.890
	w/ GMNM	.597	.826	.678	.611	.826	.688	.870	.953	.900
	MCLEA	.586	.821	.672	.663	.854	.732	.882	.955	.909
	w/ GMNM	.604	.841	.689	.678	.869	.748	.889	.960	.915
	MEAformer	.640	.877	.725	.706	.898	.776	.902	.969	.927
	w/ GMNM	.656	.884	.738	.718	.905	.786	.904	.971	.929
	SNAG (Ours)	**.678**	**.897**	**.758**	**.728**	**.915**	**.796**	**.905**	.971	**.930**
OpenEA$_{D-W-V2}$	EVA	.774	.949	.838	.789	.953	.848	.889	.981	.922
	w/ GMNM	.787	.956	.848	.799	.958	.856	.892	.983	.924
	MCLEA	.751	.941	.822	.801	.950	.856	.929	.984	.950
	w/ GMNM	.766	.956	.836	.811	.965	.868	.938	.990	.957
	MEAformer	.807	.976	.869	.834	.980	.886	.939	.994	.960
	w/ GMNM	.833	.980	.886	.857	.983	.903	.942	.995	.962
	SNAG (Ours)	**.852**	**.986**	**.901**	**.870**	**.988**	**.913**	**.946**	**.996**	**.965**

like MKGformer (Chen et al., 2022b) due to significant parameter size differences (our model: 13M; MKGformer: over 200M).

MMEA: ***(i)*** Following Yang et al. (2019), Bag-of-Words (BoW) is employed for encoding relations (x^r) and attributes (x^a) into fixed-length vectors ($d_r = d_a = 1000$). This process entails sorting relations and attributes by frequency, followed by truncation or padding to standardize vector lengths, thus streamlining representation and prioritizing significant features. For any entity e_i, vector positions correspond to the presence or frequency of top-ranked attributes and relations, respectively. ***(ii)*** Following (Chen et al., 2020a; Lin et al., 2022), vision encoders Enc_v are selected as ResNet-152 (He et al., 2016) for DBP15K, and CLIP (Radford et al., 2021) for Multi-OpenEA. ***(iii)*** An alignment editing method is applied to minimize error accumulation (Sun et al., 2018). ***(iv)*** The head number N_h in MHCA is set to 1. The hidden layer dimensions d for all networks are unified into 300. The total epochs for baselines are set to 500 with an option for an additional 500 epochs of iterative training (Lin et al., 2022). Our training strategies incorporates a cosine warm-up schedule (15% of steps for LR warm-up), early stopping, and gradient accumulation, using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a consistent batch size of 3500. ***(v)*** The total learnable parameters of our model are comparable to those of baseline models. For instance, under the DBP15K$_{JA-EN}$ dataset: EVA has 13.27M, MCLEA has 13.22M, and our SNAG has 13.82M learnable parameters.

5 Experimental Results

Overall MKGC Results. As shown in Tab. 1, SNAG achieves SOTA performance across all metrics on three MKGC datasets, especially notable when compared with recent works like MANS (Zhang et al., 2023) and MMRNS (Xu et al., 2022) which all have refined the Negative Sampling techniques. Our Entity-level Modality Interaction approach for MMKG representation learning not only demonstrates a significant advantage but also benefits from the consistent performance enhancement provided by our Gauss Modality Noise Masking (GMNM) module, maintaining superior performance even in its absence.

Overall MMEA Results. As illustrated in the third segment of Tab. 2, our SNAG achieves SOTA performance across all metrics on seven standard MMEA datasets. Notably, in the latter four datasets of the OpenEA series (EN-FR-15K, EN-DE-15K, D-W-15K-V1, D-W-15K-V2) under the *Standard* setting where $R_{img} = 1.0$ indicating full image representation for each entity, our GMNM module maintains or even boosts performance. This suggests that strategic noise integration can lead to beneficial results, demonstrating the module's effectiveness even in scenarios where visual data is abundant and complete. This aligns with findings from related work (Chen et al., 2023b,a), which suggest that image ambiguities and multi-aspect

Table 3: Component Analysis for SNAG on MKGC datasets. The icon ◐ indicates the activation of the Gauss Modality Noise Masking (GMNM) module; ◑ denotes its deactivation. By default, GMNM's noise application probability ρ is set to 0.2, with a noise ratio ϵ of 0.7. Our Transformer-based structure serves as the default fusion method for SNAG. Alternatives include: "FC" (concatenating features from various modalities followed by a fully connected layer); "WS" (summing features weighted by a global learnable weight per modality); "AT" (leveraging an Attention network for entity-level weighting); "TS" (using a Transformer for weighting to obtain confidence scores $\tilde{w}^m$ for weighted summing); "w/ Only h^g" (using Graph Structure embedding for uni-modal KGC). "Dropout" is an experimental adjustment where Equation (1) is replaced with the Dropout function to randomly zero modal input features, based on a defined probability.

Variants	DB15K			MKG-W			MKG-Y		
	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
◐ **SNAG** (Full)	**.363**	**.274**	**.530**	**.373**	**.302**	**.503**	**.395**	**.354**	**.471**
◐ $\rho = 0.3, \epsilon = 0.6$	.361	.272	.528	.373	.302	.502	.393	.353	.468
◐ $\rho = 0.1, \epsilon = 0.8$	.360	.272	.525	.371	.299	.496	.391	.348	.463
◐ $\rho = 0.4, \epsilon = 0.4$	.358	.268	.526	.365	.296	.492	.388	.346	.458
◐ $\rho = 0.5, \epsilon = 0.2$	.360	.270	.528	.368	.299	.493	.389	.348	.457
◐ $\rho = 0.7, \epsilon = 0.2$	.359	.270	.526	.367	.299	.490	.387	.345	.456
◑ **SNAG**	.357	.269	.523	.365	.296	.490	.387	.345	.457
◑ - FC Fusion	.327	.210	.522	.350	.287	.467	.378	.340	.442
◑ - WS Fusion	.334	.218	.529	.361	.298	.480	.384	.345	.449
◑ - AT Fusion	.336	.225	.528	.361	.296	.481	.379	.343	.445
◑ - TS Fusion	.335	.221	.529	.358	.292	.472	.378	.344	.437
◑ - w/ Only h^g	.293	.179	.497	.337	.268	.467	.350	.291	.453
◑ - Dropout (0.1)	.349	.252	.527	.361	.297	.479	.382	.344	.446
◑ - Dropout (0.2)	.346	.249	.526	.359	.294	.478	.381	.343	.446
◑ - Dropout (0.3)	.343	.242	.524	.356	.290	.477	.381	.343	.445
◑ - Dropout (0.4)	.341	.238	.521	.356	.295	.467	.379	.341	.442

visual information can sometimes misguide the use of MMKGs. Unlike these studies that typically design models to refuse and combat noise, our SNAG accepts and intentionally integrates noise to better align with the inherently noisy conditions of real-world scenarios. Iterative training results further confirm the robustness of our approach as detailed in Appendix A.2.2.

Most importantly, as a versatile MMKG representation learning approach, it is compatible with both MMEA and MKGC tasks, illustrating its robust adaptability in diverse operational contexts.

Uncertainly Missing Modality. The first two segments from Tab. 2 present entity alignment performance with $R_{img} = 0.4, 0.6$, where 60%/40% of entities lack image data. These missing images are substituted with random image features following a normal distribution based on the observed mean and standard deviation across other entities' images (details in 3.2). This simulates uncertain modality absence in real-world scenarios. Our method outperforms baselines more significantly when the modality absence is greater (i.e., $R_{img} = 0.4$), with the GMNM module providing notable benefits. This demonstrates that intentionally introducing noise can increase training challenges while enhancing model robustness in realistic settings.

Ablation studies. In Table 3, we dissect the influence of various components on our model's performance, focusing on three key aspects: ***(i)*** **Noise Parameters:** The noise application probability ρ and noise ratio ϵ are pivotal. Optimal values of $\rho = 0.2$ and $\epsilon = 0.7$ were determined empirically, suggesting that the model tolerates up to 20% of entities missing images and that a modality-mask ratio of 0.7 acts as a soft mask. For optimal performance, we recommend empirically adjusting these parameters to suit other specific scenario. Generally, conducting a grid search on a smaller dataset subset can quickly identify suitable parameter combinations. ***(ii)*** **Entity-Level Modality Interaction:** Our exploration shows that absence of image information (w/ Only h^g) markedly reduces performance, emphasizing MKGC's importance; Weighted summing methods (WS, AT, TS) surpass simple FC-based approaches, indicating the superiority of nuanced modality integration; Using purely Transformer modality weights $\tilde{w}^m$ for weighting does not demonstrate a clear advantage over attention-based or globally learnable weight methods in MKGC. In contrast, our approach, which utilizes $\bar{h}^g$ (for DBP15K) and $\bar{h}^{avg}$ (for MKG-W and MKG-Y), significantly outperforms others, demonstrating its efficacy. ***(iii)*** **Modality-Mask vs. Dropout:** In assessing their differential impacts, we observe that even minimal dropout (0.1) adversely affects performance, likely because dropout to some extent distorts the original modal feature distribution, thereby hindering model optimization toward the alignment objective. Conversely, our modality-mask's noise is inherent, replicating the feature distribution seen when modality is absent, and consequently enhancing model robustness more effectively.

6 Conclusion

In this work, we introduce a unified noise-powered multi-modal knowledge graph representation framework that accepts and intentionally integrates noise, thereby aligning with the complexities of real-world scenarios. This initiative also stands out as the first in the MMKG domain to support both MKGC and MMEA tasks simultaneously, highlighting the adaptability of our approach.

7 Limitations

References & Definition. To aid quick comprehension of tasks within limited space, definitions and boundaries may lack full accuracy and completeness. Detailed explanations and related work are provided in the Appendix A.1 to elaborate on these concepts.

Benchmarks & Baselines. Due to page constraints, we selected datasets and benchmarks (e.g., Tab. 1 and Tab. 2) primarily from recent mainstream works, such as DB15K (Liu et al., 2019), MKG-W and MKG-Y (Xu et al., 2022). This selection may overlook older datasets like FB15K-237 (Toutanova et al., 2015), WN18 (Bordes et al., 2013), and WN9-IMG (Xie et al., 2017b).

For fairness in comparison, we excluded methods based on MoE or Ensemble approaches, such as MoSE (Zhao et al., 2022), and did not compare with PLM-based methods like MKGformer (Chen et al., 2022b) due to significant differences in parameter sizes (our model has only 13M parameters versus MKGformer's over 200M).

Future Applications. Our framework proposes a unified approach for MMKG representation learning, ideally positioned as an MMKG encoder for integrating into LLM training processes, potentially enhancing multi-modal entity embeddings. While our method theoretically supports diverse training objectives, due to the focused scope of this study, we did not validate this aspect experimentally. As the field progresses, we envision further integration of this unified framework into multi-modal knowledge pre-training, potentially supporting various downstream tasks like Multi-modal Knowledge Injection and Retrieval-Augmented Generation (RAG). Such developments could significantly benefit the community, particularly with the rapid advancements in Large Language Models.

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. Beit: BERT pre-training of image transformers. In *ICLR*. OpenReview.net.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: multi-modal knowledge graph embeddings via optimal transport. In *NeurIPS*.

Hao Chen, Jindong Wang, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. 2024. Learning with noisy foundation models. *Preprint*, arXiv:2403.06869.

Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020a. MMEA: entity alignment for multi-modal knowledge graph. In *KSEM (1)*, volume 12274 of *Lecture Notes in Computer Science*, pages 134–147. Springer.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022a. Multi-modal siamese network for entity alignment. In *KDD*, pages 118–126. ACM.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022b. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR*, pages 904–915. ACM.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z. Pan, Wenting Song, and Huajun Chen. 2023a. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *ACM Multimedia*, pages 3317–3327. ACM.

Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023b. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *ISWC*, volume 14265 of *Lecture Notes in Computer Science*, pages 121–139. Springer.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z. Pan, and Huajun Chen. 2023c. DUET: cross-modal semantic grounding for contrastive zero-shot learning. In *AAAI*, pages 405–413. AAAI Press.

Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. In *ACM SIGIR Forum*, volume 40, pages 64–69. ACM New York, NY, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Quan Fang, Xiaowei Zhang, Jun Hu, Xian Wu, and Changsheng Xu. 2023a. Contrastive multi-modal knowledge graph representation learning. *IEEE Trans. Knowl. Data Eng.*, 35(9):8983–8996.

Quan Fang, Xiaowei Zhang, Jun Hu, Xian Wu, and Changsheng Xu. 2023b. Contrastive multi-modal knowledge graph representation learning. *IEEE Trans. Knowl. Data Eng.*, 35(9):8983–8996.

Biao Gong, Xiaoying Xie, Yutong Feng, Yiliang Lv, Yujun Shen, and Deli Zhao. 2023. Uknow: A unified knowledge protocol for common-sense reasoning and vision-language pre-training. *CoRR*, abs/2302.06891.

Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing*, 461:598–607.

Lingbing Guo, Zhuo Chen, Jiaoyan Chen, and Huajun Chen. 2023. Revisit and outstrip entity alignment: A perspective of generative models. *CoRR*, abs/2305.14651.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Ningyuan Huang, Yash R. Deshpande, Yibo Liu, Houda Alberts, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2022. Endowing language models with multimodal knowledge graph representations. *CoRR*, abs/2206.13163.

Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Neftune: Noisy embeddings improve instruction finetuning. *CoRR*, abs/2310.05914.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. 2023. VISTA: visual-textual knowledge graph representation learning. In *EMNLP (Findings)*, pages 7314–7328. Association for Computational Linguistics.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Qian Li, Cheng Ji, Shu Guo, Zhaoji Liang, Lihong Wang, and Jianxin Li. 2023a. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. pages 987–999.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. 2023b. Graphadapter: Tuning vision-language models with dual knowledge graph. *CoRR*, abs/2309.13625.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023c. IMF: interactive multi-modal fusion model for link prediction. In *WWW*, pages 2572–2580. ACM.

Yancong Li, Xiaoming Zhang, Fang Wang, Bo Zhang, and Feiran Huang. 2022. Fusing visual and textual content for knowledge graph embedding via dual-track model. *Appl. Soft Comput.*, 128:109524.

Yangning Li, Jiaoyan Chen, Yinghui Li, Yuejia Xiang, Xi Chen, and Hai-Tao Zheng. 2023d. Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. In *ICASSP*, pages 1–5. IEEE.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *CoRR*, abs/2212.05767.

Ke Liang, Sihang Zhou, Yue Liu, Lingyuan Meng, Meng Liu, and Xinwang Liu. 2023a. Structure guided multi-modal pre-trained transformer for knowledge graph reasoning. *CoRR*, abs/2307.03591.

Shuang Liang, Anjie Zhu, Jiasheng Zhang, and Jie Shao. 2023b. Hyper-node relational graph attention network for multi-modal knowledge graph completion. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(2):62:1–62:21.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *COLING*, pages 2572–2584. International Committee on Computational Linguistics.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *AAAI*, pages 4257–4266. AAAI Press.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *ESWC*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer.

Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.*, 52(7):7480–7497.

Wenxin Ni, Qianqian Xu, Yangbangyan Jiang, Zongsheng Cao, Xiaochun Cao, and Qingming Huang. 2023. PSNEA: pseudo-siamese network for entity alignment between multi-modal knowledge graphs. In *ACM Multimedia*, pages 3489–3497. ACM.

Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022. Contrastive language-image pre-training with knowledge graphs. In *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.

Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In **SEM@NAACL-HLT*, pages 225–234. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*, pages 697–706. ACM.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC (1)*, volume 10587 of *Lecture Notes in Computer Science*, pages 628–644. Springer.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *IJCAI*, pages 4396–4402. ijcai.org.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proc. VLDB Endow.*, 13(11):2326–2340.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *ICLR (Poster)*. OpenReview.net.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, pages 1499–1509. The Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR (Poster)*. OpenReview.net.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Enqiang Wang, Qing Yu, Yelin Chen, Wushouer Slamu, and Xukang Luo. 2022. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Inf. Fusion*, 88:78–85.

Luyao Wang, Pengnian Qi, Xigang Bao, Chunlai Zhou, and Biao Qin. 2024a. Pseudo-label calibration semi-supervised multi-modal entity alignment. In *AAAI*, pages 9116–9124. AAAI Press.

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is visual context really helpful for knowledge graph? A representation learning perspective. In *ACM Multimedia*, pages 2735–2743. ACM.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *ACM Multimedia*, pages 2391–2399. ACM.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. 2024b. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. *CoRR*, abs/2401.17859.

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal data enhanced representation learning for knowledge graphs. In *IJCNN*, pages 1–8. IEEE.

Yuyang Wei, Wei Chen, Shiting Wen, An Liu, and Lei Zhao. 2023. Knowledge graph incremental embedding for unseen modalities. *Knowl. Inf. Syst.*, 65(9):3611–3631.

W. X. Wilcke, Peter Bloem, Victor de Boer, and R. H. van t Veer. 2023. End-to-end learning on multimodal knowledge graphs. *CoRR*, abs/2309.01169.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017a. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146. ijcai.org.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017b. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146. ijcai.org.

Baogui Xu, Chengjin Xu, and Bing Su. 2023a. Cross-modal graph attention network for entity alignment. In *ACM Multimedia*, pages 3715–3723. ACM.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *ACM Multimedia*, pages 3857–3866. ACM.

Derong Xu, Jingbo Zhou, Tong Xu, Yuan Xia, Ji Liu, Enhong Chen, and Dejing Dou. 2023b. Multimodal biological knowledge graph completion via triple co-attention mechanism. In *ICDE*, pages 3928–3941. IEEE.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *EMNLP/IJCNLP (1)*, pages 4430–4440. Association for Computational Linguistics.

Yichi Zhang, Mingyang Chen, and Wen Zhang. 2023. Modality-aware negative sampling for multi-modal knowledge graph embedding. *CoRR*, abs/2304.11618.

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. *Preprint*, arXiv:2402.15444.

Yichi Zhang and Wen Zhang. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *CoRR*, abs/2209.07084.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *EMNLP*, pages 10527–10536. Association for Computational Linguistics.

Jia Zhu, Changqin Huang, and Pasquale De Meo. 2023. DFMKE: A dual fusion multi-modal knowledge graph embedding framework for entity alignment. *Inf. Fusion*, 90:111–119.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *CoRR*, abs/2202.05786.

A Appendix

A.1 Supplementary for Related Work

Typically, a KG is considered multi-modal when it contains knowledge symbols expressed across various modalities, including, but not limited to, text, images, sound, or video. Current research primarily concentrates on the visual modality, assuming that other modalities can be processed similarly.

Multi-Modal Knowledge Graph Completion. Multi-modal Knowledge Graph Completion (MKGC) is crucial for inferring missing triples in existing MMKGs, involving three sub-tasks: Entity Prediction, Relation Prediction, and Triple Classification. Currently, most research in MKGC focuses on Entity Prediction, also widely recognized as Link Prediction, with two main methods emerging: **Embedding-based Approaches** build on conventional Knowledge Graph Embedding (KGE) methods (Bordes et al., 2013; Sun et al., 2019), adapted to integrate multi-modal data, enhancing entity embeddings. ***(i)*** **Modality Fusion Methods** (Wilcke et al., 2023; Wang et al., 2022; Huang et al., 2022) integrate multi-modal and structural embeddings to assess triple plausibility. Early efforts, like IKRL (Xie et al., 2017a), utilize multiple TransE-based scoring functions (Bordes et al., 2013) for modal interaction. RSME (Wang et al., 2021) employs gates for selective modal information integration. OTKGE (Cao et al., 2022) leverages optimal transport for fusion, while CMGNN (Fang et al., 2023b) implements a multi-modal GNN with cross-modal contrastive learning. HRGAT (Liang et al., 2023b) creates a hyper-node relational graph, CamE (Xu et al., 2023b) focuses on biological KGs with a triple co-attention module, VISITA (Lee et al., 2023) utilizes a transformer framework for relation and triple-level multi-modal information fusion. ***(ii)*** **Modality Ensemble Methods** train distinct models per modality, merging outputs for predictions. For example, MoSE (Zhao et al., 2022) utilizes structural, textual, and visual data to train three KGC models and employs, using ensemble strategies for joint predictions. Similarly, IMF (Li et al., 2023c) proposes an interactive model to achieve modal disentanglement and entanglement to make robust predictions. ***(iii)*** **Modality-aware Negative Sampling Methods** (Lu et al., 2022; Zhang and Zhang, 2022; Zhang et al., 2023; Xu et al., 2022) boost differentiation between correct and erroneous triples by incorporating multi-modal context for superior negative sample selection. MMKRL (Lu et al., 2022) introduces adversarial training to MKGC, adding perturbations to modal embeddings. Following this, VBKGC (Zhang and Zhang, 2022) and MANS (Zhang et al., 2023) develop fine-grained visual negative sampling to better align visual with structural embeddings for more nuanced comparison training. MMRNS (Xu et al., 2022) enhances this with relation-based sample selection. **Finetune-based Approaches** (Chen et al., 2022b; Liang et al., 2023a) exploit the world understanding capabilities of pre-trained Transformer models like BERT (Devlin et al., 2019) and VisualBERT (Li et al., 2019) for MKGC. These approaches reformat MMKG triples as token sequences for PLM processing (Liang et al., 2022), often framing KGC as a classification task. For example, MKGformer (Chen et al., 2022b) integrates multi-modal fusion at multiple levels, treating MKGC as a Masked Language Modeling (MLM) task, while SGMPT (Liang et al., 2023a) extends this by incorporating structural data and a dual-strategy fusion module.

Multi-Modal Entity Alignment. Entity Alignment (EA) is pivotal for KG integration, aiming to identify identical entities across different KGs by leveraging relational, attributive, and literal (surface) features. Multi-Modal Entity Alignment (MMEA) enhances this process by incorporating visual data, thereby improving alignment accuracy accuracy (Liu et al., 2019; Chen et al., 2020a; Ni et al., 2023; Xu et al., 2023a). Introduced in 2020, MMEA (Chen et al., 2020a) merges multiple modalities to align entities in MMKGs by minimizing the distance between their holistic embeddings. HMEA (Guo et al., 2021) represents MMKGs on the hyperbolic manifold, offering refined geometric interpretations. EVA (Liu et al., 2021) applies an attention mechanism to modulate the importance of each modality and introduces an unsupervised approach that utilizes visual similarities for alignment, reducing reliance on gold-standard labels. MSNEA (Chen et al., 2022a) leverages visual cues to guide relational feature learning. MCLEA (Lin et al., 2022) employs KL divergence to mitigate the modality distribution gap between uni-modal and joint embeddings. DFMKE (Zhu et al., 2023) employs a late fusion approach with modality-specific low-rank factors that enhance feature integration across vari-

Table 4: Statistics for the MKGC datasets, where the symbol definitions in the table header align with Definition 1.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{T}_{\mathcal{R}}$ (Train)$\rvert$	$\lvert\mathcal{T}_{\mathcal{R}}$ (Valid)$\rvert$	$\lvert\mathcal{T}_{\mathcal{R}}$ (Test)$\rvert$
DB15K	12842	279	79222	9902	9904
MKG-W	15000	169	34196	4276	4274
MKG-Y	15000	28	21310	2665	2663

ous knowledge spaces, complementing early fusion output vectors. MEAformer (Chen et al., 2023a) adjusts mutual modality preferences dynamically for entity-level modality fusion, addressing inconsistencies in entities' surrounding modalities. MoAlign (Li et al., 2023a), UMAEA (Chen et al., 2023b) PCMEA (Wang et al., 2024a) and DESAlign (Wang et al., 2024b) follow similar settings.

A-MMKG vs. N-MMKG. Drawing on the categorization proposed in (Zhu et al., 2022), we distinguish between two types of MMKGs: A-MMKG and N-MMKG. In A-MMKGs, images are attached to entities as attributes, while in N-MMKGs, images are treated as standalone entities interconnected with others. A-MMKGs are more prevalent in current research and applications within the semantic web community due to their accessibility and similarity to traditional KGs. Therefore, this paper will focus exclusively on A-MMKG, unless stated otherwise. For instance, in an MMKG, an attribute triple (e, a, v) in $\mathcal{T}_{\mathcal{A}}$ might associates an image as v to an entity e via an attribute a, typically denoted as *hasImage*.

A.2 Supplementary for Experiments

A.2.1 Datasets

In MMKG datasets like DBP15KJA-EN, where 67.58% of entities have images, the image association ratio (R_{img}) varies due to the data collection process (Chen et al., 2023a).

MKGC: ***(i)*** DB15K (Liu et al., 2019) is constructed from DBPedia (Lehmann et al., 2015), enriched with images obtained via a search engine. ***(ii)*** MKG-W and MKG-Y (Xu et al., 2022) are subsets of Wikidata (Vrandecic and Krötzsch, 2014) and YAGO (Suchanek et al., 2007) respectively. Text descriptions are aligned with the corresponding entities using the additional *sameAs* links provided by OpenEA benchmarks (Sun et al., 2020). Detailed statistics are available in Tab. 5 & 4.

MMEA: ***(i)*** Multi-modal DBP15K (Liu et al., 2021) extends DBP15K (Sun et al., 2017) by adding images from DBpedia and Wikipedia (Denoyer and Gallinari, 2006), covering three bilingual settings (DBP15K$_{\text{ZH-EN}}$, DBP15K$_{\text{JA-EN}}$, DBP15K$_{\text{FR-EN}}$) and featuring around 400K triples and 15K aligned entity pairs per setting. ***(ii)*** MMEA-UMVM (Chen et al., 2023b) includes two bilingual datasets (EN-FR-15K, EN-DE-15K) and two monolingual datasets (D-W-15K-V1, D-W-15K-V2) derived from Multi-OpenEA datasets ($R_{sa} = 0.2$) (Li et al., 2023d) and all three bilingual datasets from DBP15K (Liu et al., 2021). It offers variability in visual information by randomly removing images, resulting in 97 distinct dataset splits with different R_{img}. For this study, we focus on representative R_{img} values of $\{0.4, 0.6, maximum\}$ to validate our experiments. When $R_{img} = maximum$, the dataset corresponds to the original *Standard* dataset (as shown in Tab. 2). Note that for the Multi-modal DBP15K dataset, the "$maximum$" value is not 1.0.

Table 5: Statistics for the MMEA datasets. Each dataset contains 15,000 pre-aligned entity pairs ($|\mathcal{S}| = 15000$). Note that not every entity is paired with associated images or equivalent counterparts in the other KG. Additional abbreviations include: DB (DBpedia), WD (Wikidata), ZH (Chinese), JA (Japanese), FR (French), EN (English), DE (German).

Dataset	$\mathcal{G}$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{T}_{\mathcal{R}}\rvert$	$\lvert\mathcal{T}_{\mathcal{A}}\rvert$	$\lvert\mathcal{V}_{MM}\rvert$
DBP15K$_{\text{ZH-EN}}$	ZH	19,388	1,701	8,111	70,414	248,035	15,912
	EN	19,572	1,323	7,173	95,142	343,218	14,125
DBP15K$_{\text{JA-EN}}$	JA	19,814	1,299	5,882	77,214	248,991	12,739
	EN	19,780	1,153	6,066	93,484	320,616	13,741
DBP15K$_{\text{FR-EN}}$	FR	19,661	903	4,547	105,998	273,825	14,174
	EN	19,993	1,208	6,422	115,722	351,094	13,858
OpenEA$_{\text{EN-FR}}$	EN	15,000	267	308	47,334	73,121	15,000
	FR	15,000	210	404	40,864	67,167	15,000
OpenEA$_{\text{EN-DE}}$	EN	15,000	215	286	47,676	83,755	15,000
	DE	15,000	131	194	50,419	156,150	15,000
OpenEA$_{\text{D-W-V1}}$	DB	15,000	248	342	38,265	68,258	15,000
	WD	15,000	169	649	42,746	138,246	15,000
OpenEA$_{\text{D-W-V2}}$	DB	15,000	167	175	73,983	66,813	15,000
	WD	15,000	121	457	83,365	175,686	15,000

A.2.2 Iterative Training

Iterative training results further confirm the robustness of our approach, as shown in Tab. 6.

A.3 Metric Details

A.3.1 MMEA Metrics

(i) **MRR** (Mean Reciprocal Ranking ↑) is a statistic measure for evaluating many algorithms that produce a list of possible responses to a sample of queries, ordered by probability of correctness. In the field of EA, the reciprocal rank of a query

Table 6: Iterative MMEA results.

	Models	R_{img}=0.4			R_{img}=0.6			Standard		
		H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
DBP15K$_{ZH-EN}$	EVA	.696	.902	.773	.699	.903	.775	.749	.914	.810
	w/ GMNM	.708	.906	.780	.705	.911	.778	.752	.919	.813
	MCLEA	.719	.921	.796	.764	.941	.831	.818	.956	.871
	w/ GMNM	.741	.945	.818	.782	.954	.846	.830	.968	.882
	MEAformer	.754	.953	.829	.788	.958	.853	.843	.966	.890
	w/ GMNM	.763	.947	.832	.799	.959	.860	.845	.970	.891
	SNAG (Ours)	**.798**	**.957**	**.859**	**.821**	**.963**	**.876**	**.857**	**.972**	**.900**
DBP15K$_{JA-EN}$	EVA	.646	.888	.733	.657	.892	.743	.695	.904	.770
	w/ GMNM	.696	.910	.773	.700	.912	.776	.745	.916	.807
	MCLEA	.690	.922	.778	.756	.948	.828	.788	.955	.851
	w/ GMNM	.739	.937	.815	.796	.959	.858	.820	.969	.877
	MEAformer	.759	.957	.833	.808	.969	.868	.831	.972	.882
	w/ GMNM	.769	.953	.838	.817	.967	.872	.842	.974	.890
	SNAG (Ours)	**.808**	**.959**	**.864**	**.839**	**.975**	**.890**	**.861**	**.976**	**.904**
DBP15K$_{FR-EN}$	EVA	.710	.931	.792	.716	.935	.797	.769	.946	.834
	w/ GMNM	.714	.929	.794	.720	.932	.798	.777	.950	.841
	MCLEA	.731	.943	.814	.789	.958	.854	.814	.967	.873
	w/ GMNM	.759	.964	.840	.806	.974	.871	.837	.980	.893
	MEAformer	.763	.963	.842	.811	.976	.874	.844	.980	.897
	w/ GMNM	.779	.968	.847	.817	.974	.876	.852	.981	.899
	SNAG (Ours)	**.826**	**.976**	**.885**	**.852**	**.983**	**.904**	**.875**	**.987**	**.919**
OpenEA$_{EN-FR}$	EVA	.605	.869	.700	.619	.870	.710	.848	.973	.896
	w/ GMNM	.606	.870	.701	.621	.874	.713	.856	.971	.898
	MCLEA	.613	.889	.714	.702	.928	.785	.893	.983	.928
	w/ GMNM	.625	.902	.726	.707	.934	.790	.893	.983	.928
	MEAformer	.660	.913	.751	.729	.947	.810	.895	.984	.930
	w/ GMNM	.666	.916	.755	.741	.943	.815	.905	.984	.937
	SNAG (Ours)	**.692**	**.927**	**.778**	**.743**	**.945**	**.817**	**.907**	**.986**	**.939**
OpenEA$_{EN-DE}$	EVA	.776	.935	.833	.784	.937	.839	.954	.984	.965
	w/ GMNM	.779	.936	.837	.789	.938	.843	.955	.984	.966
	MCLEA	.766	.942	.829	.821	.956	.871	.969	.994	.979
	w/ GMNM	.779	.948	.840	.829	.959	.876	.971	.995	.980
	MEAformer	.803	.950	.854	.835	.958	.878	.963	.994	.976
	w/ GMNM	.807	.949	.856	.841	.961	.882	.975	.995	.982
	SNAG (Ours)	**.826**	**.962**	**.874**	**.859**	**.970**	**.899**	**.977**	**.998**	**.984**
OpenEA$_{D-W-V1}$	EVA	.647	.856	.727	.669	.860	.741	.916	.984	.943
	w/ GMNM	.663	.859	.735	.673	.862	.743	.927	.986	.950
	MCLEA	.686	.896	.766	.770	.941	.836	.947	.991	.965
	w/ GMNM	.699	.907	.778	.776	.946	.840	.949	.991	.966
	MEAformer	.718	.901	.787	.785	.934	.841	.943	.990	.962
	w/ GMNM	.728	.901	.793	.803	.942	.855	.956	.991	.970
	SNAG (Ours)	**.753**	**.930**	**.820**	**.808**	**.953**	**.864**	**.958**	**.993**	**.972**
OpenEA$_{D-W-V2}$	EVA	.854	.980	.904	.859	.983	.908	.925	.996	.951
	w/ GMNM	.866	.980	.909	.872	.981	.913	.948	.997	.969
	MCLEA	.841	.984	.899	.877	.990	.923	.971	.998	.983
	w/ GMNM	.845	.987	.902	.882	.992	.926	.973	.999	.984
	MEAformer	.886	.990	.926	.904	.992	.938	.965	.999	.979
	w/ GMNM	.902	.990	.936	.918	.993	.948	.975	.999	.985
	SNAG (Ours)	**.904**	**.994**	**.939**	**.924**	**.994**	**.952**	**.980**	.999	**.988**

entity (i.e., an entity from the source KG) response is the multiplicative inverse of the rank of the first correct alignment entity in the target KG. MRR is the average of the reciprocal ranks of results for a sample of candidate alignment entities:

$$\mathbf{MRR} = \frac{1}{|\mathcal{S}_{te}|} \sum_{i=1}^{|\mathcal{S}_{te}|} \frac{1}{\text{rank}_i} . \tag{11}$$

(ii) **Hits@N** describes the fraction of true aligned target entities that appear in the first N entities of the sorted rank list:

$$\text{Hits @N} = \frac{1}{|\mathcal{S}_{te}|} \sum_{i=1}^{|\mathcal{S}_{te}|} \mathbb{I}[\text{rank}_i \leqslant \text{N}] , \tag{12}$$

where rank_i refers to the rank position of the first correct mapping for the i-th query entities and $\mathbb{I} = 1$ if $\text{rank}_i \leqslant N$ and 0 otherwise. $\mathcal{S}_{te}$ refers to the testing alignment set.

A.3.2 MKGC Metrics

MKGC involves predicting the missing entity in a query, either $(h, r, ?)$ for tail prediction or $(?, r, t)$ for head prediction. To evaluate the performance, we use rank-based metrics such as mean reciprocal rank (MRR) and Hit@N (N=1, 3, 10), following standard practices in the field. ***(i)*** **MRR** is calculated as the average of the reciprocal ranks of the correct entity predictions for both head and tail predictions across all test triples:

$$\mathbf{MRR} = \frac{1}{|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} (\frac{1}{r_{h,i}} + \frac{1}{r_{t,i}}) . \tag{13}$$

(ii) **Hits@N** measures the proportion of correct entity predictions ranked within the top N positions for both head and tail predictions:

$$\mathbf{Hit@N} = \frac{1}{|\mathcal{T}_{test}|} \sum_{i=1}^{|\mathcal{T}_{test}|} (\mathbb{I}(r_{h,i} \leqslant N) + \mathbb{I}(r_{t,i} \leqslant N)) , \tag{14}$$

where $r_{h,i}$ and $r_{t,i}$ denote the rank positions in head and tail predictions, respectively.

Additionally, we employ a filter setting (Bordes et al., 2013) to remove known triples from the ranking process, ensuring fair comparisons and mitigating the impact of known information from the training set on the evaluation metrics.