BEYOND READABILITY: RETHINKING LLM-BASED MMKGC AS PARALLEL SEMANTIC RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal knowledge graph completion (MMKGC) remains challenging due to severe structural sparsity. Existing LLM-based approaches typically formulate MMKGC as a text generation task, where an entire token sequence must be produced before an entity can be identified. This paradigm is inefficient and error-prone: it emits a full sequence merely to name a single entity, and optimizing for linguistic coherence rather than entity discrimination often misaligns outputs with the target entity space. We argue that MMKGC is inherently a semantic ranking problem over candidate entities, and propose RIN, a decoupled inference framework that treats token-level prediction as an intermediate representation for entity ranking rather than a goal in itself. RIN invokes the LLM backbone only once to obtain a global semantic anchor, and then simulates the step-by-step token prediction process with a lightweight module, scoring the full candidate set without further backbone invocations. Concretely, RIN consists of three components: (1) an instruction fine-tuning module that reformulates MMKGC into a ranking-oriented objective and aligns multimodal graph inputs with LLM representations, (2) a one-shot inference module that replaces repeated LLM forward passes with a single pass followed by lightweight step-wise simulation, and (3) an entity-level consistency modeling module that captures token-level dependencies to produce globally grounded entity scores. Experiments on three benchmark datasets show that RIN outperforms 24 state-of-the-art methods, improving accuracy while achieving $5.9\times$ higher inference throughput than prior generation-based LLM pipelines. Code and data are available at: https://anonymous.4open.science/RIN-g8dXbN4pR6tJ/.

1 INTRODUCTION

Multimodal knowledge graph completion (MMKGC) Wang et al. (2019b) aims to infer missing entities in knowledge graphs by leveraging heterogeneous information, including structural relations, textual descriptions, and visual signals. It plays a critical role in a wide range of applications such as semantic search, recommendation, and knowledge reasoning Wang et al. (2019a); Dong et al. (2024); Gu et al. (2023). However, MMKGC remains highly challenging due to severe structural sparsity and the difficulty of effectively integrating multimodal information.

Recent advances in large language models (LLMs) OpenAI (2023) have introduced a new paradigm for MMKGC by exploiting their strong semantic understanding capabilities. Prior efforts, exemplified by LLMKGC Zhu et al. (2024), utilize prompt engineering to harness ChatGPT OpenAI (2023) for completion reasoning, while frameworks such as KGLlama Yao et al. (2025) and KoPA Zhang et al. (2024a) adopt instruction fine-tuning via triple verification to assess the veracity of candidate facts. However, these methods cast MMKGC as text generation, tightly coupling reasoning with the decoding process, which suffers from two fundamental limitations: (i) the absence of explicit boundary constraints on the entity space renders the generated text susceptible to being misaligned with the target entity space, and (ii) generating an entire entity name token by token incurs inference latency that the completion task itself does not require. In addition, existing LLM-based MMKGC approaches largely rely on textual inputs, with multimodal signals such as visual information not being fully utilized, and rarely exploit the local graph structure surrounding the query entity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To intuitively illustrate such limitations, we take the query triple (*snowman*, *color*, ?) as an example under the conventional character-level generation paradigm, as shown in Figure 1 (c). Suppose the model predicts the initial token "w" and then generates the high-frequency token "while". This constitutes a typical out-of-candidate-set prediction that violates the structural constraints of multimodal knowledge graphs (MMKGs) Liu et al. (2019) and lacks factual validity. In contrast, the proposed RIN framework explicitly decouples the inference process from readability constraints, and evaluates candidate entities according to the joint probability of their token sequences. Even if the unconstrained generation trajectory naturally drifts toward the out-of-graph token sequence "while", the scoring mechanism of RIN ensures that the ground-truth entity "white" achieves the highest ranking within the constrained candidate space. This design effectively anchors the semantic knowledge of LLMs to the structural boundaries of MMKGs, guarantees the factual correctness of predictions, and provides a solid foundation for single-pass inference over the full candidate set.

To address these limitations, we revisit the formulation of MMKGC and argue that it is fundamentally a semantic ranking problem over a discrete entity space, rather than a natural language generation task. Motivated by this observation, we propose RIN, a decoupled inference framework that separates semantic reasoning from token-level generation. RIN treats token-level prediction as an intermediate representation for entity ranking rather than a goal in itself, thereby discarding readability in favor of discrimination. By invoking the LLM backbone only once and simulating subsequent token prediction with a lightweight module, RIN avoids repeated full-model forward passes and scores the full candidate set efficiently.

RIN consists of three components. The instruction fine-tuning module reformulates the MMKGC task into a ranking-oriented objective and aligns the model with the multimodal graph distribution. The one-shot inference module encodes the query into a global semantic anchor and enables efficient scoring over candidate entities through lightweight step-wise simulation. The entity-level consistency modeling module further grounds predictions by aggregating token-level joint probabilities to produce globally grounded entity scores.

Our major contributions are summarized as follows.

- We revisit LLM-based MMKGC and reformulate it as a semantic ranking problem over a discrete entity space, revealing the limitations of generation-based approaches.
- We propose RIN, a decoupled inference framework that separates semantic reasoning from token-level generation, enabling efficient entity-level scoring over the full candidate set.
- We introduce an entity-level consistency modeling mechanism that captures token-level dependencies to produce globally grounded entity scores.
- Extensive experiments on three benchmark datasets demonstrate that RIN consistently outperforms 24 state-of-the-art methods, improving accuracy while achieving 5.9× higher inference throughput than generation-based LLM pipelines.

2 Related Works

Knowledge Graph Completion. Early knowledge graph completion methods learn low-dimensional representations of entities and relations by modeling triple structures. Representative translational models, such as TransE Bordes et al. (2013) and RotatE Sun et al. (2019), interpret relations as translations or rotations in vector space, while semantic matching models, including DistMult Yang et al. (2015) and ConvKB Nguyen et al. (2018), capture entity-relation interactions via bilinear or convolutional operations. To alleviate the limitations of purely structural modeling, multimodal knowledge graph completion (MMKGC) has been introduced to incorporate auxiliary information such as textual descriptions and visual signals. Representative methods, such as IKRL Xie et al. (2017), integrate visual features with structural embeddings to enrich entity representations. Despite these efforts, such approaches remain heavily dependent on observed graph topology and often degrade in sparsely connected regions.

LLM-based Knowledge Graph Completion. Recently, large language models (LLMs) OpenAI (2023) have been explored for knowledge graph completion by leveraging their pretrained semantic knowledge. Early efforts, such as LLMKGC Zhu et al. (2024), transform triples into natural language prompts and apply LLMs to predict missing entities in a generation-based manner. Subsequent methods, including KGLlama Yao et al. (2025) and KoPA Zhang et al. (2024a), further in-

vestigate instruction tuning and verification-based paradigms to improve completion accuracy over candidate entities. More recent work, such as K-ON Guo et al. (2025a), explores lightweight adaptation strategies that inject external knowledge into head layers of LLMs, while still relying primarily on text-centric modeling paradigms.

Discussion. Despite significant progress, existing MMKGC methods still encounter critical hurdles. First, structure-based models rely heavily on explicit graph topology, leading to a marked degradation in reasoning performance in sparsely connected regions. Second, current LLM-based completion paradigms, while leveraging vast world knowledge, are constrained by the sequential cost of token-by-token generation and tend to produce predictions misaligned with the target entity space. More fundamentally, generation-based formulations optimize the likelihood of token sequences rather than discriminating among candidate entities. Consequently, achieving accurate and efficient completion that is explicitly anchored to the entity space remains a formidable challenge.

3 METHODOLOGY

3.1 PRELIMINARIES

Multimodal Knowledge Graphs. A multimodal knowledge graph is defined as $G = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M})$, where $\mathcal{E}$ denotes the set of entities, $\mathcal{R}$ the set of relations, and $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ the set of triples. $\mathcal{M}$ denotes the modality space, which includes text and images. For each entity $e \in \mathcal{E}$, its observation under modality $m \in \mathcal{M}$ is denoted as e_m, where e_t and e_v represent textual and visual modalities, respectively.

MMKGC Task. Given an incomplete triple $(h, r, ?)$ or $(?, r, t)$, the MMKGC task aims to predict the missing entity by ranking candidate entities in $\mathcal{E}$.

Large Language Models. Large language models (LLMs) are autoregressive (AR) probabilistic models:

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i \mid x_{<i}; \theta), \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a token sequence and θ denotes model parameters.

3.2 LIMITATIONS OF GENERATION-BASED COMPLETION

Despite their strong semantic modeling capabilities, generation-based LLM completion paradigms exhibit a fundamental mismatch with the nature of MMKGC: the goal of MMKGC is not open-ended sequence generation, but semantic ranking over a fixed and discrete entity space. These formulations optimize the likelihood of token sequences rather than discriminating among candidate entities, and can only approximate the global ranking objective through inherently sequential, token-by-token generation paths. Consequently, they incur redundant computation and substantial decoding latency—candidate entities cannot be scored in parallel—while structural constraints are only implicitly enforced, and their predictions tend to be misaligned with the target entity space.

3.3 RIN FRAMEWORK OVERVIEW

As illustrated in Figure 1, RIN adopts a decoupled inference paradigm that separates semantic reasoning from token-by-token generation. Instead of sequential decoding, the model performs direct semantic ranking over candidate entities via a unified pipeline that invokes the LLM backbone only once per query.

Given an input query, RIN first encodes multimodal information into a semantic representation, then performs efficient candidate scoring, and finally refines predictions through consistency modeling to obtain entity-level scores.

- **Instruction Fine-tuning.** This module reformulates MMKGC as a ranking-oriented task by converting structured triples and multimodal signals into instruction-following inputs, aligning the model with the multimodal graph distribution.

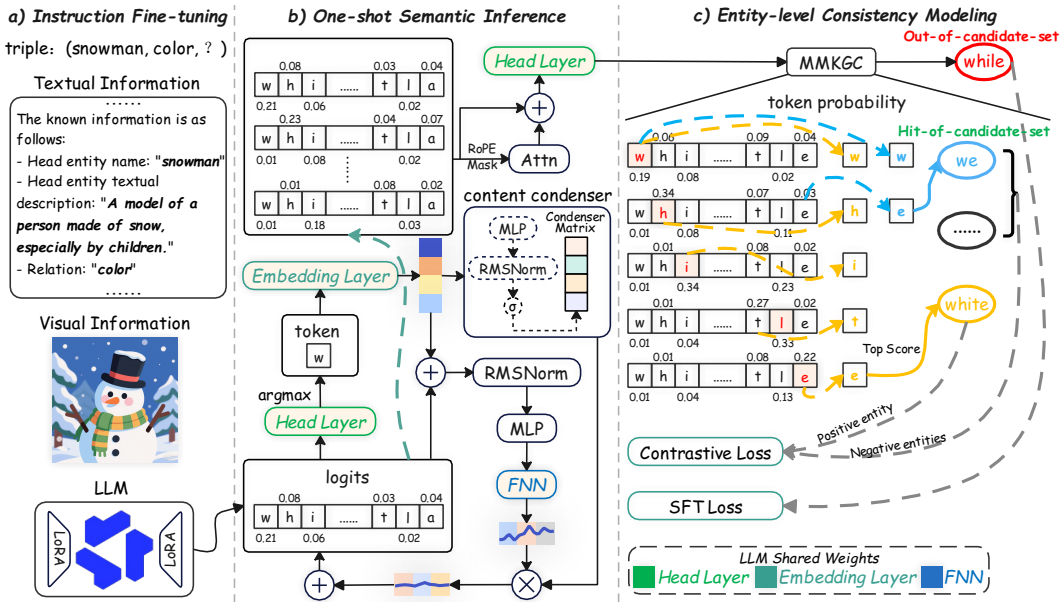

Figure 1: Architectural Overview of the RIN Model. For ease of exposition, we simplify the tokenization process by treating each character as an individual token.

- **One-shot Semantic Inference.** This module decouples reasoning from autoregressive decoding by performing a single-pass LLM inference to obtain a global semantic anchor, followed by a lightweight step-wise simulation that scores all candidate entities without repeated LLM calls.
- **Entity-level Consistency Modeling.** This module captures token-level dependencies to refine candidate predictions and aggregates them into globally grounded entity scores, improving structural validity and keeping predictions aligned with the target entity space.

Overall, RIN unifies multimodal alignment, decoupled inference, and consistency-aware scoring under a semantic ranking framework, enabling efficient and structurally grounded MMKGC.

3.4 Instruction Fine-tuning

As illustrated in Figure 1(a), this module bridges the gap between multimodal knowledge graphs and language models by reformulating MMKGC as an instruction-following task.

To this end, we design a predefined template (details are provided in Appendix 6.2) that transforms structured triples together with their associated multimodal information into natural language instructions. Specifically, the structured components (e.g., entities and relations), together with the local neighbor facts of the query entity, are combined with textual descriptions and visual features to form a unified instruction input. The target triple is excluded from the injected neighbor facts to prevent information leakage. This formulation enables the model to jointly encode structural and multimodal semantics within a consistent language space, thereby aligning the multimodal graph distribution with the input space of LLMs.

To adapt the pretrained LLM to this instruction-based formulation, we adopt LoRA Hu et al. (2022) for parameter-efficient fine-tuning:

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{AB}, \tag{2}$$

where $\mathbf{W}_0 \in \mathbb{R}^{|\mathcal{V}| \times d}$ denotes the frozen parameters of the LLM, and $\mathbf{A}$ and $\mathbf{B}$ are low-rank adaptation matrices. This design enables efficient optimization while preserving the original model knowledge.

3.5 ONE-SHOT SEMANTIC INFERENCE

Autoregressive LLM inference in MMKGC is inherently sequential, resulting in high latency and limited parallelism. We replace it with a single LLM forward pass followed by lightweight step-wise simulation.

Single-pass LLM Inference. Given an input query, we perform a single forward pass through the frozen LLM backbone M and take the hidden state at the last query token as the query representation:

$$\mathbf{h}_{base} = M(\text{query}), \tag{3}$$

which carries the global semantics of the query. The backbone is invoked only once per query, independent of the number of candidate entities, and model adaptation is achieved via LoRA. The step-wise predictions below are seeded directly from $\mathbf{h}_{base}$ through the shared prediction head, which yields the token distribution at the first position together with its most likely token:

$$\mathbf{p}_1 = \text{Head}_{\text{shared}}(\mathbf{h}_{base}), \qquad t_1 = \text{argmax}(\mathbf{p}_1). \tag{4}$$

Step-wise Prediction from the Query Representation. Building upon $(\mathbf{h}_{base}, t_1)$, we design a lightweight step-wise predictor that produces token distributions at all K positions. These predictions are not used as the model's output text; they serve as a fine-grained intermediate representation for entity-level scoring and are supervised only by a light auxiliary sequence loss during training.

An entity is naturally encoded as a token sequence, and reliably distinguishing among semantically related candidates requires composing the semantics of its constituent tokens step by step: a single holistic representation, despite providing rich global context, is insufficiently fine-grained for entity-level discrimination. The predictor therefore maintains a recurrent state that carries the query anchor and the token predicted at the previous step, and accumulates token-level semantic evidence across positions. Let $\mathbf{r}_1 = \mathbf{h}_{base}$; for each position $k \geq 2$, the input state is constructed as:

$$\mathbf{z}_{k-1} = \mathbf{r}_{k-1} + \text{Emb}_{\text{shared}}(t_{k-1}), \tag{5}$$

where $t_{k-1} = \text{argmax}(\mathbf{p}_{k-1})$ and $\text{Emb}_{\text{shared}}$ is the input embedding matrix of the backbone. We then apply a lightweight transformation to extract high-level semantic features:

$$\tilde{\mathbf{z}}_{k-1} = \text{MLP}_k(\text{RMSNorm}(\mathbf{z}_{k-1})), \tag{6}$$

where each step owns its MLP. To efficiently reuse pretrained knowledge, we leverage the shared FeedForward network from the backbone LLM:

$$\mathbf{h}_k = \text{FFN}_{\text{shared}}(\tilde{\mathbf{z}}_{k-1}). \tag{7}$$

A Content Condenser $\text{Cond}(\cdot)$ is further applied as a state-dependent sigmoid gate Qiu et al. (2025) to suppress noise while preserving informative features:

$$\mathbf{g}_k = \sigma(\text{RMSNorm}(\mathbf{W}_g \mathbf{z}_{k-1})) \in [0,1]^d, \qquad \mathbf{h}_k \leftarrow \mathbf{h}_k \odot \mathbf{g}_k, \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\mathbf{W}_g$ is a learnable gating matrix shared across positions, and $\odot$ denotes element-wise multiplication. Token distributions and the recurrent state are then updated as:

$$\mathbf{p}_k = \text{Head}_{\text{shared}}(\mathbf{h}_k), \qquad t_k = \text{argmax}(\mathbf{p}_k), \qquad \mathbf{r}_k = \mathbf{h}_k + \mathbf{h}_{base}. \tag{9}$$

The step-wise dependency across positions is retained by design—it underlies the accumulation of token-level semantic evidence—and is implemented outside the LLM backbone, incurring negligible computational overhead. The step predictions $\mathbf{L} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]$ are used only to seed the next step and to provide auxiliary supervision; entity scoring is performed on the refined representations below.

Consistency-aware Refinement. To compensate for the limited sequential dependencies captured by the predictor, we introduce a lightweight attention-based refinement module.

Let $\mathbf{H} = [\mathbf{h}_{base}, \mathbf{h}_1, \ldots, \mathbf{h}_{K-1}]$ denote the sequence of query and step representations. We first inject relative positional information using RoPE Su et al. (2021):

$$\mathbf{H}_{rope} = \text{RoPE}(\text{RMSNorm}(\mathbf{H}), p_{1:K}), \tag{10}$$

where $p_{1:K}$ denotes position indices. We then apply causal self-attention to model directional dependencies:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}, \tag{11}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are all obtained from $\mathbf{H}_{rope}$, and $\mathbf{M}$ is a causal mask. Finally, we fuse multi-source information via residual aggregation:

$$\mathbf{H}_{final} = \omega_1\mathbf{A} + \omega_2\mathbf{H} + \omega_3\mathbf{h}_{base}, \tag{12}$$

where $\omega_1, \omega_2, \omega_3$ are fixed fusion weights. Entity scores are produced from the refined representations:

$$\mathbf{P} = \text{Head}_{\text{shared}}(\mathbf{H}_{final}), \tag{13}$$

which yields the token distributions used for entity scoring; the step predictions $\mathbf{p}_k$ are used only for seeding and auxiliary supervision. This refinement enables the model to approximate the cumulative effect of autoregressive dependency modeling within a decoupled and efficient inference framework.

3.6 Entity-level Consistency Modeling

Given the refined token representations $\mathbf{H}_{final}$, we project them into the vocabulary space $\mathcal{V}$ using a shared prediction head, yielding a logits matrix $\mathbf{L} \in \mathbb{R}^{K\times|\mathcal{V}|}$, where K is the maximum entity length and $|\mathcal{V}|$ is the vocabulary size.

Entity-level Scoring. To avoid unconstrained text generation and keep predictions aligned with the target entity space, we perform entity-level scoring over a predefined candidate set. Each candidate entity e is represented as a token sequence $\{t_1, \ldots, t_m\}$ ($m \le K$). We compute its score by aggregating token-level log-probabilities across positions:

$$\text{Score}(e) = \sum_{i=1}^{m} \log P(t_i \mid \mathbf{L}_i), \tag{14}$$

where $\mathbf{L}_i$ denotes the logits at position i, and $P(t_i \mid \mathbf{L}_i)$ is obtained via softmax over the vocabulary.

This formulation explicitly anchors predictions to the target entity space and ensures structural validity.

Training Objective. To optimize entity discrimination, we adopt a self-adversarial contrastive loss:

$$\mathcal{L}_{\text{KGC}} = -\log\sigma(\text{Score}(e^+)) - \sum_{j=1}^{|\mathcal{N}|} w_j \log\left(1 - \sigma(\text{Score}(e_j^-))\right), \tag{15}$$

where the weights w_j are computed via temperature-scaled softmax:

$$w_j = \frac{\exp(\text{Score}(e_j^-)/T)}{\sum_{j'=1}^{|\mathcal{N}|}\exp(\text{Score}(e_{j'}^-)/T)}. \tag{16}$$

In practice, the log-sigmoid terms are evaluated using the numerically stable identities $\log\sigma(s) = -\text{softplus}(-s)$ and $\log(1-\sigma(s)) = -\text{softplus}(s)$.

In addition, we incorporate a supervised token-level objective to stabilize training under lightweight step-wise prediction:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{K}\sum_{k=1}^{K} \log P(y_k \mid \mathbf{L}_k), \tag{17}$$

where y_k is the ground-truth token at position k.

The final objective is a weighted combination:

$$\mathcal{L} = \mathcal{L}_{\text{KGC}} + \lambda\mathcal{L}_{\text{SFT}}. \tag{18}$$

4 EXPERIMENT

This section evaluates the proposed RIN framework on the multimodal link prediction task. After detailing the experimental setup, we address the following research questions:

- **RQ1:** Does RIN outperform existing state-of-the-art methods on MMKGC benchmarks?
- **RQ2:** To what extent do the individual components and specific architectural designs of RIN contribute to its overall predictive performance?
- **RQ3:** How much efficiency gain does RIN achieve compared to autoregressive and traditional MMKGC baselines, in terms of training cost and inference throughput?
- **RQ4:** How sensitive is the proposed framework to the negative sampling size?

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate RIN on three benchmark datasets to test its robustness across heterogeneous settings. DB15K Liu et al. (2019) integrates visual data retrieved from search engines with DBpedia entities Lehmann et al. (2015), while MKG-W and MKG-Y Xu et al. (2022) are subsets derived from Wikidata Vrandecic & Krötzsch (2014) and YAGO Suchanek et al. (2007), respectively.

Baselines. We compare RIN against a broad range of baselines covering four representative categories in MMKGC.

(i) *Structure-based methods*, which model relational structures using only graph topology, including TransE Bordes et al. (2013), DistMult Yang et al. (2015), RotatE Sun et al. (2019), PairRE Chao et al. (2021), and other embedding-based variants such as KBGAN Cai & Wang (2018).

(ii) *Text-enhanced methods*, which incorporate textual descriptions or semantic features, including IKRL Xie et al. (2017), TransAE Wang et al. (2019c), RSME Wang et al. (2021), and MANS Zhang et al. (2023).

(iii) *Multimodal MMKGC models*, which jointly exploit visual and textual information, including TBKGC Sergieh et al. (2018), MMKRL Lu et al. (2022), IMF Li et al. (2023), QEB Wang et al. (2023), VISTA Lee et al. (2023), and MMRNS Xu et al. (2022), AdaMF-MAT Zhang et al. (2024b), MyGO Zhang et al. (2025), M-REFT Wang et al. (2025), APKGC Jian et al. (2025), MCKGC Gao et al. (2025), and LBMKGC Guo et al. (2025b).

(iv) *LLM-based approaches*, which leverage large language models for knowledge graph reasoning, including KGLlama Yao et al. (2025), GPT-3.5 Zhu et al. (2024), and K-ON Guo et al. (2025a).

Metrics. We follow standard evaluation protocols in prior work and report ranking-based metrics, including Mean Reciprocal Rank (MRR) and Hits@K ($K = 1, 3, 10$). All results are evaluated under the filtered setting Bordes et al. (2013) for fair comparison with existing methods.

Implementation details. We utilize Qwen2.5-VL-7B-Instruct Bai et al. (2025) as the backbone and conduct all experiments on $8\times$ NVIDIA H20 GPUs. We employ the AdamW Kingma & Ba (2015) optimizer with a learning rate of 1×10^{-4}, while setting the batch size to 8 and the negative sampling size $|N|$ to 128.

4.2 MAIN RESULTS (RQ1)

We evaluate RIN on the three benchmark datasets, with results reported in Table 1. Overall, RIN achieves the best score on every metric across all three datasets, outperforming 24 baselines spanning structure-based, text-enhanced, multimodal, and LLM-based methods. Relative to the strongest competing model on each metric, RIN improves MRR, Hits@1, Hits@3, and Hits@10 by 2.7%, 3.9%, 3.4%, and 1.0% on DB15K; by 11.4%, 20.8%, 8.4%, and 1.5% on MKG-W; and by 14.1%, 13.6%, 12.5%, and 8.6% on MKG-Y. The improvements are particularly pronounced on Hits@1 for MKG-W and MKG-Y, indicating a stronger capability in precise entity ranking under multimodal settings.

We further compare RIN with representative LLM-based baselines, including KG-Llama-7b Yao et al. (2025), GPT-3.5 Turbo Zhu et al. (2024), and K-ON Guo et al. (2025a). To ensure a fair

Table 1: Knowledge Graph Completion Models Comparison

Category	Model	DB15K				MKG-W				MKG-Y			
		MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Structure-based	TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.37
	DistMult	23.03	14.78	26.28	39.59	20.99	15.93	22.28	30.86	25.04	19.33	27.80	35.95
	RotatE	29.28	17.87	36.12	49.66	33.67	26.80	36.68	46.73	34.95	29.10	38.35	45.30
	PairRE	31.13	21.62	35.91	49.30	34.40	28.24	36.71	46.04	32.01	25.53	35.84	43.89
	KBGAN	25.73	9.91	36.95	51.93	29.47	22.21	34.87	40.64	29.71	22.81	34.88	40.21
Text-enhanced	IKRL	26.82	14.09	34.93	49.09	32.36	26.11	34.75	44.07	33.22	30.37	34.28	38.26
	TransAE	28.09	21.25	31.17	41.17	30.00	21.23	34.91	44.72	28.10	25.31	29.10	33.03
	RSME	29.76	24.15	32.12	40.29	29.23	23.36	31.97	40.43	34.44	31.78	36.07	39.09
	MANS	28.82	16.87	36.58	49.26	30.88	24.89	33.63	41.78	29.03	25.25	31.35	34.49
Multimodal	TBKGC	28.40	15.61	37.03	49.86	31.48	25.31	33.98	43.24	33.99	30.47	35.27	40.07
	MMKRL	26.81	13.85	35.07	49.39	30.10	22.16	34.09	44.69	36.81	31.66	39.79	45.31
	IMF	32.25	24.20	36.00	48.19	34.50	28.77	36.62	45.44	35.79	32.95	37.14	40.63
	QEB	28.18	14.82	36.67	51.55	32.38	25.47	35.06	45.32	34.37	29.49	36.95	42.32
	VISTA	30.42	22.49	33.56	45.94	32.91	26.12	35.38	45.61	30.45	24.87	32.39	41.53
	MMRNS	32.68	23.01	37.86	51.01	35.03	28.59	37.49	47.47	35.93	30.53	39.07	45.47
	AdaMF-MAT	35.14	25.30	41.11	52.92	35.77	28.87	38.85	48.71	38.57	34.34	40.59	45.76
	MyGO	37.72	30.08	41.26	52.21	36.10	29.78	38.54	47.75	38.44	35.01	39.84	44.19
	M-REFT	37.60	30.30	41.00	51.50	37.20	30.80	39.90	48.90	37.30	34.50	39.20	42.90
	APKGC	36.40	28.20	41.30	52.70	37.40	30.60	40.40	50.10	-	-	-	-
	MCKGC	39.79	31.92	43.80	54.66	36.88	31.32	38.92	47.43	38.92	35.49	40.57	45.21
	LBMKGC	37.23	27.78	42.75	54.71	38.46	31.20	41.78	51.46	40.03	33.89	43.11	50.81
LLM-based	KGLlama	-	13.46	-	-	-	20.20	-	-	-	-	-	-
	GPT-3.5	-	21.71	-	-	-	22.66	-	-	-	-	-	-
	K-ON	38.10	30.13	42.77	53.59	36.64	30.05	38.72	48.26	33.59	28.23	36.21	42.74
LLM-based (Ours)	RIN	**40.85**	**33.15**	**45.27**	**55.24**	**42.84**	**37.83**	**45.27**	**52.22**	**45.69**	**40.31**	**48.48**	**55.20**
		+2.7%	+3.9%	+3.4%	+1.0%	+11.4%	+20.8%	+8.4%	+1.5%	+14.1%	+13.6%	+12.5%	+8.6%

and consistent comparison, K-ON results on DB15K and MKG-W are adopted from the original paper, and those on MKG-Y are reproduced under the same protocol. As shown in Table 1, RIN consistently outperforms these LLM-based methods on all three datasets; for example, it improves MRR by 7.2%, 16.9%, and 36.0% and Hits@1 by 10.0%, 25.9%, and 42.8% over K-ON on DB15K, MKG-W, and MKG-Y, respectively. This suggests that replacing token-by-token generation with a single forward pass that scores all candidate entities directly leads to more accurate and stable predictions for MMKGC tasks.

Table 2: Ablation Study

Settings	DB15K				MKG-W				MKG-Y			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
RIN	**40.85**	**33.15**	**45.27**	**55.24**	**42.84**	**37.83**	**45.27**	**52.22**	**45.69**	**40.31**	48.48	**55.20**
RIN w/o cond	-	-	-	-	-	-	-	-	45.67	40.08	**48.62**	55.10
RIN w/o attn	38.71	31.09	43.79	53.19	40.75	35.57	43.63	50.18	42.89	37.51	45.47	52.42
RIN w/o $\mathcal{L}_{\text{KGC}}$	32.38	25.70	35.94	45.56	39.78	34.59	42.49	48.46	40.82	35.97	43.34	50.71
RIN w/o $\mathcal{L}_{\text{SFT}}$	37.60	30.50	41.99	51.73	42.14	36.58	45.05	51.70	43.91	37.76	47.38	54.74
RIN w/o e_v	40.73	33.10	44.47	54.38	42.46	37.51	45.00	51.73	44.97	39.82	47.29	54.56

4.3 Ablation Study (RQ2)

We ablate the key design choices in RIN, as summarized in Table 2. The full model achieves 40.85% MRR and 33.15% Hits@1 on DB15K.

Decoupled Single-Pass Inference. Removing the attention refinement module (attn) lowers DB15K MRR to 38.71% and Hits@1 to 31.09%. More critically, removing the content condenser (cond) causes training non-convergence on DB15K and MKG-W; on MKG-Y, where it still converges, MRR remains essentially unchanged at 45.67%. These results indicate that semantic compression is critical for training stability, while dependency modeling remains essential for reliable candidate-level prediction.

Entity-level Optimization Strategy. Removing $\mathcal{L}_{\text{KGC}}$ causes the most significant performance degradation, lowering DB15K MRR to 32.38% and Hits@1 to 25.70%. Removing $\mathcal{L}_{\text{SFT}}$ also leads to a clear decline, to 37.60% MRR and 30.50% Hits@1, confirming that explicit entity-level supervision is crucial for ranking accuracy.

Multimodal Instruction Design. Removing the visual feature e_v causes a moderate drop on DB15K, with MRR at 40.73% and Hits@1 at 33.10%, while the decline is slightly larger on MKG-W and MKG-Y, where MRR falls to 42.46% and 44.97%. Replacing the multimodal backbone Qwen2.5-

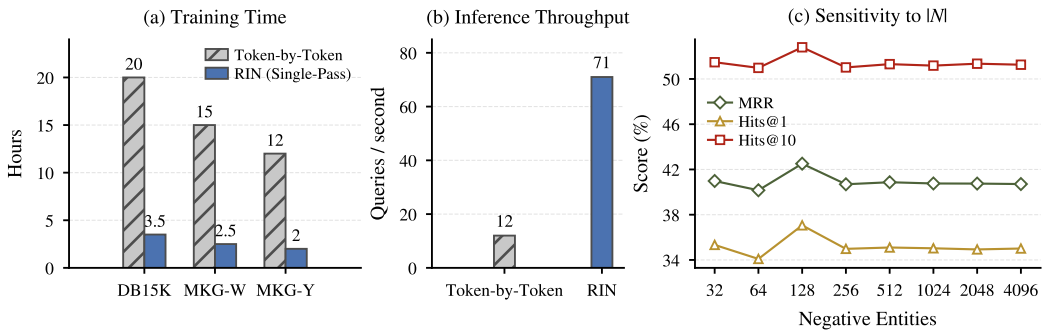

Figure 2: Efficiency and sensitivity analysis.

VL-7B with its text-only counterpart Qwen2.5-7B decreases MRR to 40.73% on DB15K, 42.65% on MKG-W, and 45.38% on MKG-Y, showing that RIN remains effective even without a multimodal backbone.

Overall, decoupled inference and entity-level optimization are the main contributors to performance gains, while multimodal instruction design provides a necessary foundation.

4.4 Efficiency and Sensitivity Analysis (RQ3 and RQ4)

We compare RIN against a token-by-token LLM baseline. As shown in Figure 2(a,b), on DB15K and under the same number of training epochs, RIN reduces the training time from approximately 20 hours to 3.5 hours, corresponding to an $\sim$83% reduction. It also improves inference throughput from 12 to 71 queries per second, yielding a $5.9\times$ speedup. These results demonstrate that the single-pass design substantially reduces the training and inference cost of LLM-based MMKGC.

To assess web-scale scalability, we expand the candidate pool of each benchmark to 3M entities. As reported in Appendix 6.1, RIN outperforms LBMKGC on every metric across all three datasets under this setting. Moreover, its candidate-scoring cost grows linearly with the number of candidate entities, making exhaustive evaluation feasible at substantially larger candidate scales.

We further investigate the effect of the negative sampling size $|N|$ by varying it from 32 to 4096. As shown in Figure 2(c), a moderate sampling size around $|N| = 128$ yields the best overall performance. Increasing $|N|$ beyond this point leads to a broad performance plateau with only minor fluctuations, indicating that substantially larger negative sets provide limited additional benefit. These results suggest that moderate negative sampling offers a favorable balance between informative contrastive signals and training efficiency.

5 Conclusion and Limitation

We propose RIN, a single-pass inference framework for LLM-based Multimodal Knowledge Graph Completion (MMKGC) that reformulates token-by-token generation as semantic ranking over candidate entities. This design reduces inference latency, keeps predictions aligned with the target entity space, and achieves state-of-the-art performance on three benchmarks.

A key limitation lies in structural modeling. RIN relies mainly on the semantic priors of LLMs and cross-modal entity descriptions, together with the local neighbor facts injected during instruction fine-tuning; it does not explicitly propagate graph topology, so capturing complex structural dependencies remains limited compared with dedicated structure-based methods.

Future work will explore dynamic knowledge graphs and the effective integration of structural priors to enhance the structural awareness of RIN.

AI USE STATEMENT

Generative AI tools were used solely for language editing of the manuscript. All outputs were reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

To support reproducibility, we publicly release the source code of our method on GitHub, including the complete implementation, training and evaluation pipelines. The dataset used in this work is also provided with an accessible link.

ETHICS STATEMENT

This work presents a method for multimodal knowledge graph completion and does not involve human-subject studies or the collection of new personal data. Our experiments use publicly available benchmark datasets and follow their original licenses and terms of use.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2787–2795, 2013.

Liwei Cai and William Yang Wang. KBGAN: adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1470–1480. Association for Computational Linguistics, 2018.

Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. Pairre: Knowledge graph embeddings via paired relation vectors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 4360–4369. Association for Computational Linguistics, 2021.

Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 2417–2429. Association for Computational Linguistics, 2024.

Yuxiao Gao, Fuwei Zhang, Zhao Zhang, Xiaoshuang Min, and Fuzhen Zhuang. Mixed-curvature multi-modal knowledge graph completion. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 11699–11707. AAAI Press, 2025.

Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18941–18951. IEEE, 2023.

Lingbing Guo, Yichi Zhang, Zhongpu Bo, Zhuo Chen, Mengshu Sun, Zhiqiang Zhang, Wen Zhang, and Huajun Chen. K-ON: stacking knowledge on the head layer of large language model. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 11745–11753. AAAI Press, 2025a.

Yuan Guo, Qian Ma, Hui Li, Qiao Ning, Furui Zhan, Yu Gu, Ge Yu, and Shikai Guo. Lbmkgc: Large model-driven balanced multimodal knowledge graph completion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Yue Jian, Xiangyu Luo, Zhifei Li, Miao Zhang, Yan Zhang, Kui Xiao, and Xiaoju Hou. APKGC: noise-enhanced multi-modal knowledge graph completion with attention penalty. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 15005–15013. AAAI Press, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. VISTA: visual-textual knowledge graph representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 7314–7328. Association for Computational Linguistics, 2023.

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. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6 (2):167–195, 2015.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pp. 2572–2580. ACM, 2023.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. MMKG: multi-modal knowledge graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503, pp. 459–474, 2019.

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

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pp. 327–333. Association for Computational Linguistics, 2018.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pp. 225–234. Association for Computational Linguistics, 2018.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy (eds.), *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pp. 697–706. ACM, 2007.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

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

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. Is visual context really helpful for knowledge graph? A representation learning perspective. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 2735–2743. ACM, 2021.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. KGAT: knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 950–958. ACM, 2019a.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 2391–2399. ACM, 2023.

Yunpeng Wang, Bo Ning, Xin Wang, Chengfei Liu, and Guanyu Li. 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 2025, Padua, Italy, July 13-18, 2025*, pp. 1176–1185. ACM, 2025.

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

Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*, pp. 1–8. IEEE, 2019c.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Image-embodied knowledge representation learning. In Carles Sierra (ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 3140–3146. ijcai.org, 2017.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 3857–3866. ACM, 2022.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. Exploring large language models for knowledge graph completion. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025.

Yichi Zhang, Mingyang Chen, and Wen Zhang. Modality-aware negative sampling for multi-modal knowledge graph embedding. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pp. 1–8. IEEE, 2023.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Wen Zhang, and Huajun Chen. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 233–242. ACM, 2024a.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 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, 20-25 May, 2024, Torino, Italy*, pp. 17120–17130. ELRA and ICCL, 2024b.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 13322–13330. AAAI Press, 2025.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llms for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web (WWW)*, 27(5):58, 2024.

6 Appendix

6.1 Web-Scale Scalability and Inference Complexity

Web-scale benchmarks for MMKGC remain scarce. We therefore stress-test RIN by expanding the candidate pool of each benchmark to 3M entities, using additional entities randomly sampled from the Wikidata entity dump (`wikidata_dump_entities_20220207`). We evaluate RIN and the strong baseline LBMKGC Guo et al. (2025b) under the same setting. Table 3 reports MRR, Hits@1, Hits@3, and Hits@10 for the original 15K candidate pools and the 3M expanded pools on MKG-Y, MKG-W, and DB15K.

As the candidate pool grows by roughly $200\times$, the MRR of LBMKGC decreases by 25.6% on average, whereas RIN decreases by only 11.9%. Accordingly, RIN's MRR margin over LBMKGC widens from 3.6–5.7 points at the 15K scale to 7.2–12.0 points at the 3M scale. RIN also outperforms LBMKGC on every metric across all datasets under both settings. The contrast is particularly pronounced on DB15K: LBMKGC's Hits@1 drops from 27.78% to 13.68%, whereas RIN declines only to 27.92%. These results suggest that LLM-based semantic representations are more robust to interference from substantially enlarged candidate spaces than embedding-based scoring.

Scoring Strategy and Inference Complexity For every query, RIN scores all entities in the knowledge graph and applies no heuristic pruning to narrow the candidate set. A single forward pass yields the logits matrix $\mathbf{L} \in \mathbb{R}^{K\times|\mathcal{V}|}$ produced by the shared prediction head, where K is the maximum entity length and $\mathcal{V}$ is the vocabulary; the scores of all $|E|$ candidate entities, defined as the joint probabilities of their token sequences, are then obtained from this matrix in one parallel gather operation, without any loop over entities. The inference cost is therefore linear in $|E|$: the LLM forward pass and the lightweight step-wise modules are independent of the number of entities, and scoring all candidates costs $\mathcal{O}(|E| \cdot K)$ with $K \ll |E|$, implemented as a GPU-parallel indexing operation. Computation does not bottleneck on entity scale, which makes the 3M-candidate evaluation above feasible and deployment on web-scale knowledge graphs practical.

Table 3: Web-scale scalability. 15K denotes the original candidate pool of each benchmark and 3M the Wikidata-expanded pool. All metrics are reported in %.

Dataset	Pool	RIN				LBMKGC			
		MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
MKG-Y	15K	45.69	40.31	48.48	55.20	40.03	33.89	43.11	50.81
	3M	39.54	32.86	43.93	50.19	30.96	22.62	36.05	44.20
MKG-W	15K	42.84	37.83	45.27	52.22	38.46	31.20	41.78	51.46
	3M	38.71	33.11	41.60	49.14	31.51	23.12	35.65	46.09
DB15K	15K	40.85	33.15	45.27	55.24	37.23	27.78	42.75	54.71
	3M	35.75	27.92	40.12	50.83	23.80	13.68	28.77	42.63

6.2 Specifications of the Predefined Prompt Template

To ensure the transparency and reproducibility of our framework, we provide the comprehensive prompt template employed for the Multimodal Knowledge Graph Completion (MMKGC) task.

Design Philosophy. The prompt is architected to facilitate a synergistic reasoning process between the LLM's internalized knowledge and the provided external evidence. We adopt an *expert role-playing* strategy to prime the model for high-precision fact-checking, and inject the local neighbor facts of the query entity so that structural evidence is available beyond the multimodal inputs. To keep predictions aligned with the target entity space, the prompt further constrains the output to the name of the predicted tail entity only, avoiding free-form text that may drift beyond the candidate entities.

Placeholder Definitions. The dynamic components of the template, denoted by braces, are instantiated as follows:

- {`hname`}: The surface name of the head (query) entity.
- {`rname`}: The relation of the query triple, mapped into a natural language phrase to activate the model's linguistic priors.
- {`hdesc`}: The textual description of the head entity extracted from the KG, truncated to a fixed length (`trim_len`) to balance semantic richness and computational efficiency.
- {`h_img_status`}: An indicator specifying the validity of the visual modality, ensuring the model ignores placeholder images in data-sparse scenarios.
- {`h_neighbor_text`}: The local neighbor facts of the head entity, each given as a (relation, connected entity) pair.

Full Template Structure The complete implementation is presented in the following box:

Multimodal Knowledge Graph Completion Prompt

You are an expert in multimodal knowledge graph completion. A knowledge graph stores facts as triples (head entity, relation, tail entity). Given an incomplete triple, your job is to predict the single most likely tail entity.

Incomplete triple to complete:
("{hname}", "{rname}", ?)

Information about the head entity:
- Name: "{hname}"
- Description: "{hdesc[:self.trim_len]}"
- Image: Picture 1 [status: {h_img_status}]
- Known neighboring facts (relation, connected entity):
{h_neighbor_text}

Guidance:
- If the image status is "placeholder", the image carries no meaningful information and must be ignored.
- Reason jointly over the relation semantics, the head entity's description, its neighboring facts, the image (only if meaningful), and general knowledge.

Output ONLY the name of the predicted tail entity. Do NOT add any explanation, reasoning, or extra text.

Answer: