RADD: Retrieval-Augmented Discrete Diffusion for Multi-Modal Knowledge Graph Completion

Anonymous ACL submission

Abstract

Most multi-modal knowledge graph completion (MMKGC) models use one embedding scorer to conduct both retrieval over the full entity set and final link prediction. We argue that this coupling is a core bottleneck: global high-recall search and local fine-grained disambiguation require different inductive biases. Therefore, we propose a **R**etrieval-**A**ugmented **D**iscrete **D**iffusion (RADD) framework to decouple retrieval and reranking for MMKGC. A relation-aware multimodal KGE retriever serves as both global retriever and distillation teacher, while a conditional discrete denoiser performs shortlist-level entity-identity generation for reranking. Training combines KGE supervision, denoising cross-entropy, and temperature-scaled distillation from the retriever to the denoiser. At inference, the designed Diff-Rerank first forms a top-K shortlist with the retriever and then reranks it with the denoiser, ensuring that recall is a strict requirement for precision. Experiments on three MMKGC benchmarks show that RADD achieves the best performance and consistent gains over strong unimodal, multimodal, and LLM-based baselines, while ablations further verify the contribution of each component.

1 Introduction

Knowledge graph completion (KGC) predicts missing head or tail entities in incomplete triples. Translation-based (Bordes et al., 2013), bilinear (Yang et al., 2015; Trouillon et al., 2016), and rotational (Sun et al., 2019; Chao et al., 2021) models provide strong foundations for structural KGC. Multi-modal knowledge graphs (MMKGs) further augment entities with image and text (Liu et al., 2019; Wang et al., 2023; Chen et al., 2024), offering richer evidence but causing uneven modality quality and relation-dependent signal reliability.

We argue that the central bottleneck in MMKGC is *decision allocation*, not representation fusion. As shown in Figure 1, a single scorer

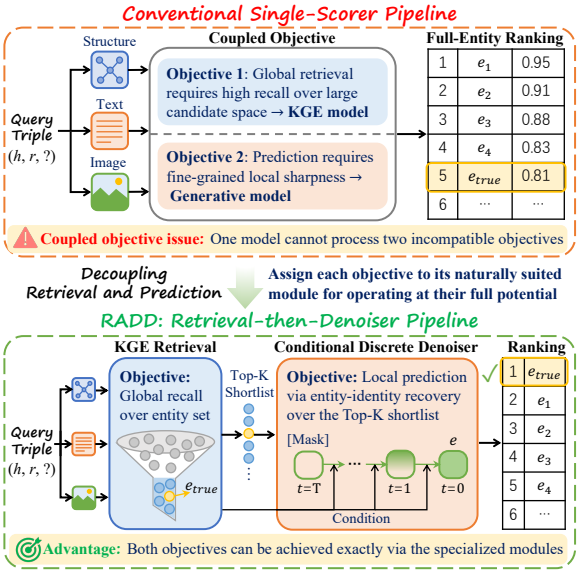

Figure 1: Conventional MMKGC uses one scorer for both high-recall search and fine-grained reranking, two objectives with conflicting inductive biases. RADD decouples them: a KGE retriever handles global search, while a discrete denoiser handles shortlist reranking.

has to simultaneously handle exhaustive full-set recall and fine-grained shortlist disambiguation. These tasks demand conflicting inductive biases: geometric smoothness for global search and conditional sharpness for local discrimination. **KGE-only models** (Sun et al., 2019) excel at the first role but fail at the second; **generative-only approaches** (Saxena et al., 2022; Huang et al., 2025) model richer conditional distributions but face vocabulary-scale sparsity without a retrieval stage. The prevailing strategy has been to improve input representations via relation-aware fusion of structural, visual, and textual embeddings (Lee et al., 2023; Zhang et al., 2024a, 2025), and more recently via LLM-based entity enrichment (Guo et al., 2024, 2025); yet all retain a single scorer at inference, leaving the tension unresolved. More specifically, our ablation studies quantify this:

on DB15K and MKG-W, both retriever-only and denoiser-only variants lag behind the full model by 9–12 points across all metrics.

To address this issue, we propose a **R**etrieval-**A**ugmented **D**iscrete **D**iffusion (RADD) framework that explicitly decouples the two objectives and assigns each to to its naturally suited module as in Figure 1. Considering that the KGC task is fundamentally a conditional recovery task: given one observed entity and a relation, the model must recover the missing entity that completes the triple. A relation-aware multimodal KGE retriever handles search over the full entity set, where geometric efficiency and recall breadth are essential. A conditional discrete denoiser then operates only after the retriever has pruned the candidate space, recovering entity identities directly in the discrete vocabulary without confronting full-vocabulary sparsity. Conditioning the denoiser on the observed entity and relation semantics (including multimodal signals) provides a principled and effective way to control generation toward triple-consistent candidates. The two components are coupled during training through temperature-scaled teacher-student distillation. At inference, the developed Diff-Rerank enforces the decomposition: the retriever commits to a hard top-K shortlist, and the denoiser reranks exclusively within it.

Our contributions are:

- We identify dual-objective coupling as a structural bottleneck in MMKGC and recast the task as retrieval-augmented discrete diffusion over entity vocabularies, adapting D3PM (Austin et al., 2021) to MMKGC.
- We propose a RADD framework comprising a relation-aware multimodal retriever, a conditional discrete denoiser, teacher-student distillation, and a Diff-Rerank inference mechanism that explicitly enforces retrieve-then-rerank decision allocation.
- Experiments on three benchmarks show the better performance of our model than existing baselines. The ablation studies further illustrate the effectiveness of each component.

2 Related Work

KGE-based KGC. Translation-based (Bordes et al., 2013), bilinear (Yang et al., 2015; Trouillon et al., 2016; Balazevic et al., 2019), and rotational (Sun et al., 2019; Chao et al., 2021; Tang et al., 2020) KGE models score triples via geometric energies over learned entity and relation embeddings. Graph neural-network (Zhu et al., 2021) and language-model scorers (Yao et al., 2019) show that the scoring mechanism design is at least as consequential as the representation. MMKGC adds uneven modality coverage and relation-dependent feature importance (Wang et al., 2017; Chen et al., 2024), further straining any single unified scorer.

Multi-modal fusion for KGC. Early MMKGC models inject visual or textual features directly into KGE objectives (Xie et al., 2017; Sergieh et al., 2018; Pezeshkpour et al., 2018; Wang et al., 2019; Lu et al., 2022). Subsequent work improves fusion quality via optimal transport (Cao et al., 2022), cross-modal attention (Li et al., 2023), hybrid transformers (Lee et al., 2023), modality-aware negative sampling (Xu et al., 2022; Zhang et al., 2023), adaptive weighting under incomplete observations (Zhang et al., 2024b,a), and relation-conditioned modality routing (Zhang et al., 2025). LLM-augmented variants (Guo et al., 2024, 2025) further narrow the performance gap by enriching entity representations with language model knowledge. All retain a single scorer at inference.

Diffusion models for structured prediction. Continuous DDPM (Ho et al., 2020) and discrete D3PM (Austin et al., 2021) established the theoretical foundations for denoising-based generation. D3PM's absorbing and uniform forward transitions over discrete state spaces are directly relevant to entity-identity recovery. Diffusion-Com (Huang et al., 2025) brings a related perspective to MMKGC by multimodal diffusion. RADD differs in two respects: it denoises discrete entity identities rather than continuous embeddings, and confines the denoiser to a KGE-supplied shortlist, avoiding full-vocabulary sparsity.

3 Task Definition

Let $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ be a multi-modal KG where each entity carries a structural embedding $\mathbf{e}^{(s)}$, a projected visual embedding $\mathbf{e}^{(v)}$, and a projected textual embedding $\mathbf{e}^{(t)} \in \mathbb{R}^d$. Given an incomplete triple $(h, r, ?)$ or $(?, r, t_e)$, the task is to rank all $|\mathcal{E}|$ candidates. RADD decomposes this into two coupled stages: a KGE-based retriever scoring every candidate and producing a top-K shortlist, and a conditional discrete denoiser reranking only

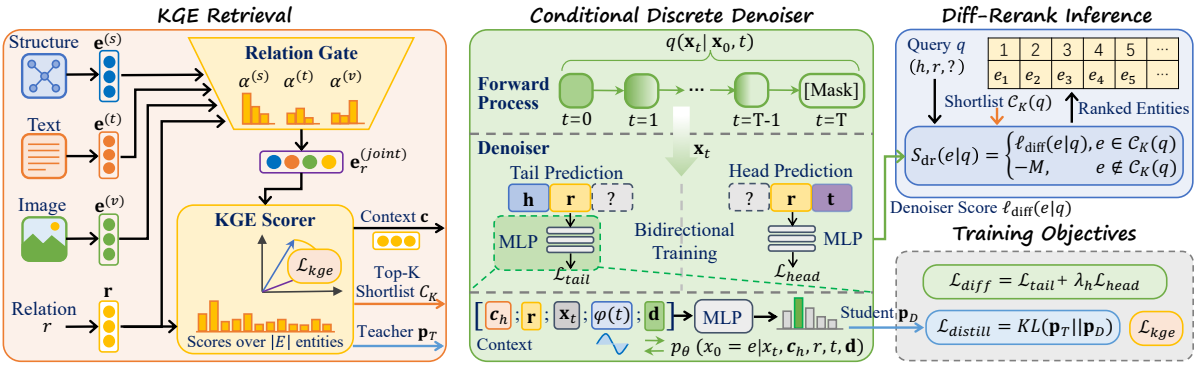

Figure 2: Overview of the RADD framework. The KGE Retriever (left) fuses structural, visual, and textual embeddings via a relation gate and scores all $|\mathcal{E}|$ candidates to form a top-K shortlist (recall stage). The Conditional Discrete Denoiser (centre) takes the retriever context $\mathbf{c}_h$, a noisy entity token x_t, a timestep embedding $\phi(t)$, and a direction embedding $\mathbf{d}$ as input, producing logits over entity identities (precision stage). During training, $\mathcal{L}_{\text{distill}}$ aligns the denoiser distribution p_D (student) with the frozen retriever's soft rankings p_T (teacher). At inference, Diff-Rerank hard-gates diffusion scores to the shortlist, enforcing recall as a strict requirement for reranking.

those K candidates. KGE supervision and knowledge distillation connect the two stages, ensuring joint optimization rather than post-hoc assembly.

4 Approach

Figure 2 shows an overview of the RADD framework. The KGE retriever fuses structural, visual, and textual signals to construct a top-K shortlist, and the conditional discrete denoiser reranks that shortlist by recovering entity identities from a corrupted token. The two modules are coupled during training via temperature-scaled distillation.

4.1 Relation-Aware Multimodal Retriever

The retriever serves two tightly coupled roles: it performs search over the full entity set to construct the shortlist seen by the denoiser, and it supplies the multimodal conditioning context used to resolve shortlist-level ambiguity. Both roles rely on the same relation-gated representation.

For entity e, the retriever constructs a relation-aware joint embedding $\mathbf{e}_r^{(\text{joint})}$ by fusing three modality-specific embeddings (structural $\mathbf{e}^{(s)}$, visual $\mathbf{e}^{(v)}$, textual $\mathbf{e}^{(t)}$) through a relation gate:

$$\mathbf{e}_r^{(\text{joint})} = \sum_{m \in \{s,v,t\}} \alpha_m^{(r)} \mathbf{e}^{(m)}, \tag{1}$$

where the relation-specific weights $\alpha_m^{(r)}$ allow each relation to up-weight its most informative modality. We reuse a RotatE-style scoring function (Sun et al., 2019) over complex-valued $\mathbf{e}^{(\text{joint})}$ as the retrieval score: complex rotation inherently models symmetric, antisymmetric, inverse, and compositional relation patterns, and allows closed-form scoring over all $|\mathcal{E}|$ candidates. The fused representation $\mathbf{e}_r^{(\text{joint})}$ is then reused as the conditioning component $\mathbf{c}_h$ in the denoiser input (Eq. 3), directly linking retrieval quality to denoising quality through a shared feature space.

4.2 Conditional Discrete Diffusion

The denoiser predicts entity identities as discrete indices into $\mathcal{E}$ rather than continuous embedding vectors, directly aligning denoising with the KGC objective: recovering the missing entity conditioned on the observed entity and relation. The conditioning signal is multimodal, fusing structural, visual, and textual features; the denoiser additionally receives a direction embedding to distinguish head from tail prediction. This discrete formulation eliminates quantization errors from post-hoc mapping of continuous outputs to entity indices, and cross-entropy over $\mathcal{E}$ yields clean gradients without continuous relaxations, simplifying joint training with the KGE retriever.

Following D3PM (Austin et al., 2021), we define a forward process over $\mathcal{E} \cup \{\texttt{[MASK]}\}$. Let $x_0 \in \mathcal{E}$ be the clean target entity. The forward process corrupts x_0 over T timesteps via a schedule-defined keep probability $p_{\text{keep}}(t)$:

$$q(x_t \mid x_0, t) = p_{\text{keep}}(t)\, \delta_{x_0} + p_{\text{mask}}(t)\, \delta_{\texttt{[MASK]}} + p_{\text{rep}}(t)\, \mathcal{U}(\mathcal{E} \setminus \{x_0\}), \tag{2}$$

where δ_z denotes a point mass at z, $\mathcal{U}(\mathcal{E} \setminus \{x_0\})$ is the uniform distribution over all entities except x_0, and $p_{\text{mask}}(t) + p_{\text{rep}}(t) = 1 - p_{\text{keep}}(t)$. Setting

$p_{\text{rep}}(t) = 0$ recovers mask-only corruption; setting $p_{\text{mask}}(t) = 0$ recovers random replacement.

The denoiser f_θ is parameterized as an MLP that takes as input the concatenation

$$\tilde{\mathbf{x}} = \big[\mathbf{c}_h;\ \mathbf{r};\ \mathbf{x}_t;\ \phi(t);\ \mathbf{d}\big], \tag{3}$$

where $\mathbf{c}_h = \mathbf{e}_r^{(\text{joint})}$ is the joint entity's embedding $\mathbf{r}$ is the relation embedding, $\mathbf{x}_t$ is the corrupted token x_t (or the learned mask embedding when $\mathbf{x}_t = \texttt{[MASK]}$), $\phi(t) \in \mathbb{R}^{d_t}$ is a sinusoidal timestep embedding, and $\mathbf{d} \in \mathbb{R}^{d_{\text{dir}}}$ is a learned direction embedding distinguishing head and tail prediction. The denoiser outputs logits over all $|\mathcal{E}|$ entities, reflecting the reverse model:

$$p_\theta(x_0 = e \mid \mathbf{x}_t, \mathbf{c}_h, \mathbf{r}, t, \mathbf{d}) = \text{softmax}(f_\theta(\tilde{\mathbf{x}})). \tag{4}$$

Since the target entity is unknown at inference, we set $\mathbf{x}_T = \texttt{[MASK]}$ and query the denoiser directly at $t = T$, obtaining a full entity distribution in a single forward pass without iterative sampling.

4.3 Bidirectional Diffusion Training

KGC requires both tail prediction $(h, r, ?)$ and head prediction $(?, r, t_e)$. These two subtasks are structurally asymmetric: relations typically impose stronger type constraints on tails, and head entities follow a more skewed frequency distribution, making head prediction consistently harder. A unidirectional denoiser would miss half of inference cases especially the harder head direction. The framework therefore trains both directions with an asymmetric loss, upweighting the head loss to compensate for this structural imbalance. The combined denoising loss is:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{tail}} + \lambda_h \mathcal{L}_{\text{head}}, \tag{5}$$

where each directional term is a cross-entropy loss over $\mathcal{E}$. The head-weighting coefficient $\lambda_h > 1$ compensates for the harder head-prediction task. The tail-only training variant serves as the corresponding ablation baseline.

4.4 Joint Optimization with Distillation

The full training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{kge}} + \mathcal{L}_{\text{diff}} + \lambda_d \mathcal{L}_{\text{distill}} + \lambda_r \mathcal{L}_{\text{rank}}. \tag{6}$$

$\mathcal{L}_{\text{kge}}$ is the standard negative-sampling KGE loss, and λ_d, λ_r are non-negative regularization weights. The distillation term transfers the retriever's soft predictions into the denoiser via temperature-scaled KL divergence over a candidate subset $\mathcal{C}$ containing the positive entity, random negatives, and KGE hard negatives:

$$\begin{aligned} \mathcal{L}_{\text{distill}} &= \text{KL}\big(p_T \,\|\, p_D\big)\,, \\ p_T(e) &= \frac{\exp(s_{\text{kge}}(e)/\tau)}{\sum_{e' \in \mathcal{C}} \exp(s_{\text{kge}}(e')/\tau)}, \\ p_D(e) &= \frac{\exp(f_\theta(e)/\tau)}{\sum_{e' \in \mathcal{C}} \exp(f_\theta(e')/\tau)}, \end{aligned} \tag{7}$$

where p_T and p_D are the temperature-τ distributions of the retriever (teacher) and denoiser (student) over $\mathcal{C}$, and $f_\theta(e)$ denotes the denoiser's output logit for entity e. By using the retriever's soft rankings as supervision targets, distillation seeds the denoiser with a concentrated prior over plausible candidates before shortlist-level discrimination begins. The retriever is frozen after a warm-up phase so that the teacher distribution is stationary when distillation begins.

4.5 Retrieval-Augmented Inference

We use a single inference strategy **Diff-Rerank**, because it most directly instantiates the proposed division of labor: the retriever is responsible for global retrieval, and the denoiser is responsible for local decision refinement. For a query q, the retriever first computes KGE scores over the full entity set and forms a shortlist

$$\mathcal{C}_K(q) = \text{TopK}_{e \in \mathcal{E}}\ s_{\text{kge}}(e \mid q). \tag{8}$$

This step preserves the retriever's strongest property: stable high-recall search over the entire entity set under noisy or incomplete multimodal evidence. The denoiser is then queried with $\mathbf{x}_T = \texttt{[MASK]}$ and the same query context to produce denoiser log-scores $\ell_{\text{diff}}(e \mid q) = \log p_\theta(\mathbf{x}_0 = e \mid \mathbf{x}_T, q)$. The final Diff-Rerank score is

$$S_{\text{dr}}(e \mid q) = \begin{cases} \ell_{\text{diff}}(e \mid q), & e \in \mathcal{C}_K(q), \\ -M, & e \notin \mathcal{C}_K(q), \end{cases} \tag{9}$$

where M is a large constant that hard-gates any entity outside the shortlist to last place. This constraint makes recall a strict requirement for final ranking: an entity can only be promoted by the denoiser if the retriever has already shortlisted it, preventing a high denoiser score from compensating for a retrieval miss. The result is a clean separation: the retriever is solely responsible for coverage, and the denoiser is solely responsible for precision within the retrieved set.

Table 1: Dataset statistics. Img and Txt count entities with at least one visual or textual feature.

Dataset	#Ent	#Rel	Train	Valid	Test	Img	Txt
DB15K	12,842	279	79,222	9,904	9,902	12,818	9,078
MKG-W	15,000	169	34,196	4,276	4,274	14,463	14,123
MKG-Y	15,000	28	21,310	2,665	2,663	14,244	12,305

5 Experiments

5.1 Datasets

We evaluate on three standard MMKGC benchmarks namely DB15K, MKG-W, and MKG-Y (Zhang et al., 2025), each with pre-extracted image and text features alongside structural triples as shown in Table 1. We further evaluate RADD-S on two traditional KGC benchmarks, FB15K-237 and NELL995, to verify that the structural component remains competitive independently.

5.2 Implementation Details

Modality embeddings are projected into the entity embedding space via a linear layer. **Retriever:** Embedding dimension 250, batch size 1024, 128 negatives per positive, KGE learning rate 10^{-4}; the KGE encoder is frozen at epoch 1000 as the distillation teacher. **Denoiser:** Learning rate 10^{-4}, T=100 diffusion steps, cosine noise schedule, replacement corruption probability 0.3 with linear annealing, bidirectional training with head-loss weight λ_h=2.0, EMA decay 0.9999. **Distillation and inference:** Temperature τ=0.7 with a 64-candidate training pool; Diff-Rerank uses a top-K=256 shortlist at inference. **Evaluation:** Following the standard filtered evaluation protocol (Bordes et al., 2013), we report Mean Reciprocal Rank (MRR) and Hits@k (k=1, 3, 10), all averaged over head and tail prediction queries.

5.3 Baselines

We compare against 27 baselines in four groups. **Unimodal KGC:** TransE (Bordes et al., 2013), TransD (Ji et al., 2015), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), PairRE (Chao et al., 2021), GC-OTE (Tang et al., 2020). **Multimodal KGC:** IKRL (Xie et al., 2017), TBKGC (Sergieh et al., 2018), TransAE (Wang et al., 2019), MMKRL (Lu et al., 2022), RSME (Wang et al., 2021), VBKGC (Zhang and Zhang, 2022), OTKGE (Cao et al., 2022), MoSE (Zhao et al., 2022), IMF (Li et al., 2023), QEB (Wang et al., 2023), VISTA (Lee et al., 2023), AdaMF-MAT (Zhang et al., 2024b), and MoMoK (Zhang et al., 2025), each using single scorer at inference. **Negative sampling:** KBGAN (Cai and Wang, 2018), MANS (Zhang et al., 2023), MM-RNS (Xu et al., 2022), and DHNS (Niu and Zhang, 2026) improve training-time sample quality. **LLM-augmented:** MKGL (Guo et al., 2024) and K-ON (Guo et al., 2025) use an LLM to enrich entity representations but apply a single scorer.

5.4 Research Questions

We structure the analysis around four questions: (RQ1) Does RADD outperforms the existing strong baselines? (RQ2) Does RADD reduce the structural asymmetry between head and tail prediction, and what design choices are responsible? (RQ3) Which components drive the gains and are they consistent across datasets? (RQ4) How sensitive is RADD to its key hyperparameters, and which design choices require careful tuning?

5.5 Results and Analysis

5.5.1 Main Results (RQ1)

As shown in Table 2, RADD performs best across all four metrics on DB15K and MKG-W, achieving MRR of 49.15 and 45.45 respectively, and leads in MRR and H@1 on MKG-Y. Against the strongest baseline on each dataset (MoMoK at 39.57 on DB15K, DHNS at 36.41 on MKG-W, and AdaMF-MAT at 38.57 on MKG-Y), the absolute MRR gains are +9.6, +9.0, and +0.8. RADD also exceeds the best LLM-augmented baseline K-ON (MRR 36.64 on DB15K) by +12.5 points, confirming that the retrieve-then-rerank decomposition is more effective than enriching representations alone. Notably, RADD achieves this with approximately 22 M trainable parameters, on par with MKGL (20 M) and over an order of magnitude fewer than K-ON (215 M), demonstrating that architectural role specialization delivers superior accuracy more efficiently than scaling up pretrained language models. On MKG-Y, RADD leads on MRR and H@1, the metrics most sensitive to top-rank precision.

The structure-only variant RADD-S already reaches 41.17, 36.48, and 38.96 MRR on DB15K, MKG-W, and MKG-Y, surpassing the strongest unimodal baselines on each dataset (GC-OTE, PairRE, and RotatE, respectively) by +9.3, +2.1, and +4.0 MRR. Strikingly, RADD-S's score of 41.17 on DB15K also exceeds every multimodal

Table 2: The comparison results (%) on three datasets. RADD-S denotes the structure-only variant of RADD. The best results are **bold** and the second-best underlined (best baseline underlined when RADD/RADD-S hold top-2).

Model		DB15K				MKG-W				MKG-Y			
		MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
Unimodal KGC	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
	TransD	21.52	8.34	29.93	44.24	26.84	19.65	31.48	42.68	26.39	17.01	33.05	40.41
	DistMult	23.03	14.78	26.98	40.59	20.95	15.89	22.88	36.80	25.04	19.32	27.80	39.95
	ComplEx	27.48	18.13	31.57	45.37	28.09	21.45	30.89	44.77	28.94	23.11	31.07	43.48
	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	35.83	43.89
	GC-OTE	31.85	22.11	36.52	51.18	33.92	26.55	35.96	46.05	32.95	26.77	36.44	44.08
	RADD-S	41.17	33.08	45.08	57.17	36.48	30.80	38.59	47.34	38.96	35.32	40.56	45.40
MMKGC	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
	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
	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
	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
	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
	VBKGC	30.61	19.75	37.18	49.44	30.61	24.91	33.01	40.88	37.04	33.76	38.75	42.30
	OTKGE	23.86	18.45	25.89	34.23	34.36	28.85	36.25	44.88	35.51	31.97	37.18	41.38
	MoSE	28.38	21.56	30.91	41.67	33.34	27.78	33.94	41.06	36.28	33.64	37.47	40.81
	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
	AdaMF	32.51	21.31	39.67	51.68	34.27	27.21	37.86	47.21	38.06	33.49	40.44	45.48
	AdaMF-MAT	35.14	25.30	41.11	52.92	35.85	29.04	39.01	48.42	38.57	34.34	40.59	45.76
	MoMoK	39.57	32.38	43.45	54.14	35.89	30.38	37.54	46.13	37.91	35.09	39.20	43.20
Negative Sampling	KBGAN	24.83	19.48	27.15	36.49	25.78	19.05	28.71	37.89	24.26	20.68	25.68	29.16
	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
	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
	DHNS	34.59	26.80	39.20	51.20	36.41	29.73	39.75	49.22	36.49	32.88	38.89	42.95
LLM-based MMKGC	MKGL	32.86	26.54	35.57	44.76	29.11	24.30	31.60	38.28	27.14	18.68	30.39	43.87
	K-ON	36.64	30.05	38.72	48.26	-	-	-	-	38.10	30.13	**42.77**	**53.59**
Ours	RADD	**49.15**	**41.00**	**53.15**	**64.27**	**45.45**	**39.88**	**47.58**	**55.83**	**39.40**	**35.75**	40.93	45.87

and LLM-augmented baseline in Table 2. Full RADD further improves over RADD-S by +3.9 on DB15K, +9.0 on MKG-W, and +0.4 on MKG-Y, confirming that multimodal conditioning provides incremental gains on datasets with higher modality coverage and relational diversity. The experiments on FB15K-237 and NELL995 in Appendix B further demonstrate that RADD's advantage holds without any multimodal input, establishing its generality for conventional KGC tasks.

Accordingly, RADD's improvements are substantial on DB15K and MKG-W, where relational diversity, triple density, and modality coverage are all high. On MKG-Y, the gain narrows to +0.8 MRR, as the simpler relational structure leaves less residual ambiguity for the denoiser to resolve. Rather than being a limitation, this behavior validates the central hypothesis: RADD's margin scales with shortlist disambiguation difficulty.

5.5.2 Head vs. Tail Entity Prediction (RQ2)

Table 3 separates head and tail prediction to locate the source of RADD's overall gains.

Head prediction. Competing methods score head MRR of 23.94–26.23 on DB15K and 18.50–18.56 on MKG-W, while their tail MRR reaches 47.94–51.68 and 51.66–52.80, yielding tail-to-head ratios of approximately 2.0× and 2.8×. RADD uniquely closes this structural gap, reducing the ratios to 1.27× and 1.55× through significant head-side improvements: +13.5 MRR (+51.4% relative) over MoMoK on DB15K and +17.1 MRR (+92.2% relative) over AdaMF-MAT on MKG-W. The H@1 gains are even more pronounced, at +68.5% relative on DB15K and +127% relative on MKG-W, and no other method comes close to these margins in either prediction direction. This recovery is driven by bidirectional diffusion training, which together counteract the many-to-one relational bias that systematically suppresses head-side recall in standard single-direction training.

Tail prediction. On DB15K, RADD's tail MRR (50.44) falls slightly below MoMoK (51.68). On MKG-W, RADD leads in tail MRR, H@1, and

Table 3: Head and tail prediction results (%) on DB15K and MKG-W. Best in **bold**, second-best underlined.

Model	DB15K								MKG-W							
	Head				Tail				Head				Tail			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
AdaMF-MAT	23.94	16.67	27.59	36.85	47.94	35.51	56.14	**70.42**	18.56	13.10	20.57	28.57	52.80	44.85	56.79	**67.69**
MoMoK	26.23	19.69	28.94	38.29	**51.68**	**43.18**	**56.29**	67.89	18.50	13.59	19.82	27.61	51.66	45.95	54.14	62.70
RADD	**39.71**	**33.18**	**42.35**	**52.02**	50.44	41.15	55.54	68.50	**35.67**	**30.86**	**37.48**	**44.06**	**55.24**	**48.90**	**57.67**	67.62

Table 4: Ablation results (%). Each ablated model removes or replaces one component of RADD. Best results in **bold**, and the most degraded variant per dataset is marked with superscript *.

Model		DB15K				MKG-W				MKG-Y			
		MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
Full model		**45.02**	**37.16**	**48.94**	**60.26**	**45.45**	**39.88**	**47.58**	**55.83**	**39.40**	**35.75**	**40.93**	**45.87**
Retriever	w/o Multimodality	41.17	33.08	45.08	57.17	36.48	30.80	38.59	47.34	38.96	35.32	40.56	45.40
	w/o Retrieval	35.56	27.77	39.69	50.53	35.23	29.97	37.44	45.33	38.44	35.05	40.07	44.61
Denoiser	w/o Denoising	32.83	24.27	36.91	49.65	35.05	27.99	38.29	47.96	38.74	34.57	40.59	45.85
	Tail-only Training	25.40*	19.31*	28.24*	37.19*	30.02*	25.90*	31.76*	38.11*	31.07*	27.62*	32.86*	37.83*
Training	w/o Distillation	39.93	32.00	43.56	55.30	35.60	29.35	38.39	47.09	39.08	35.35	40.69	45.49

H@3, while achieving comparable performance with AdaMF-MAT in H@10. Competitive baselines thus already capture most available tail-side signal; RADD recovers head-side recall without sacrificing tail performance.

Overall. Balanced bidirectional performance is RADD's most distinctive structural advantage among the compared models. Rather than further improving the already-strong tail side, the retrieve-then-rerank design with asymmetric bidirectional training substantially recovers head-side recall without sacrificing tail performance, and it is precisely this head-side recovery that drives the largest share of RADD's aggregate performance improvements in Table 2.

5.5.3 Ablation Study (RQ3)

Table 4 isolates the contribution of each component by ablating one module at a time. Three findings hold consistently across all three benchmarks.

Both the retriever and denoiser are necessary. Removing Denoiser reduces MRR to 32.83/35.05/38.74 (gaps of 12.2/10.4/0.7 points), while removing Retriever reaches only 35.56/35.23/38.44. Both single-module variants lag behind the full RADD, confirming that recall-focused retrieval and shortlist-level denoising are complementary. Removing the multimodal signals (RADD-S) further reduces MRR by $-3.9/-9.0/-0.4$, indicating that more modalities provide essential evidence beyond structural signals alone, with the effect amplified on datasets with higher modality coverage.

Bidirectional training is essential. Tail-only training incurs the largest overall degradation (ΔMRR: $-19.6/-15.4/-8.3$ on DB15K/MKG-W/MKG-Y). Per-direction inspection in Table 3 reveals that tail performance remains competitive under tail-only training, while head MRR collapses disproportionately, indicating that the bulk of the degradation is head-side. This illustrates that inverse-relation augmentation and the asymmetric head-loss weight λ_h are both necessary to counteract the structural many-to-one bias that suppresses head-side recall in standard training.

Distillation gain scales with retriever quality. Removing distillation reduces MRR by $-5.1/-9.9$ on DB15K/MKG-W but only -0.3 on MKG-Y. The gap tracks dataset richness: on richer datasets the KGE shortlist yields more actionable soft rankings for the denoiser; on MKG-Y, fewer relation types and lower modality diversity weaken shortlist confidence and reduce the distillation signal.

5.5.4 Hyperparameter Sensitivity (RQ4)

Figure 3 sweeps nine parameters and reveals a practically favorable tuning asymmetry: retrieval-side settings drive nearly all performance variance, while diffusion-side settings are broadly stable.

General parameters (Row 1). Top-K shortlist size is the most sensitive factor (MRR range across tested values: 4.4%): $K = 128$ focuses denoising on the most confident KGE candidates, while $K > 256$ admits marginal candidates the retriever cannot separate. Batch size improves MRR mono-

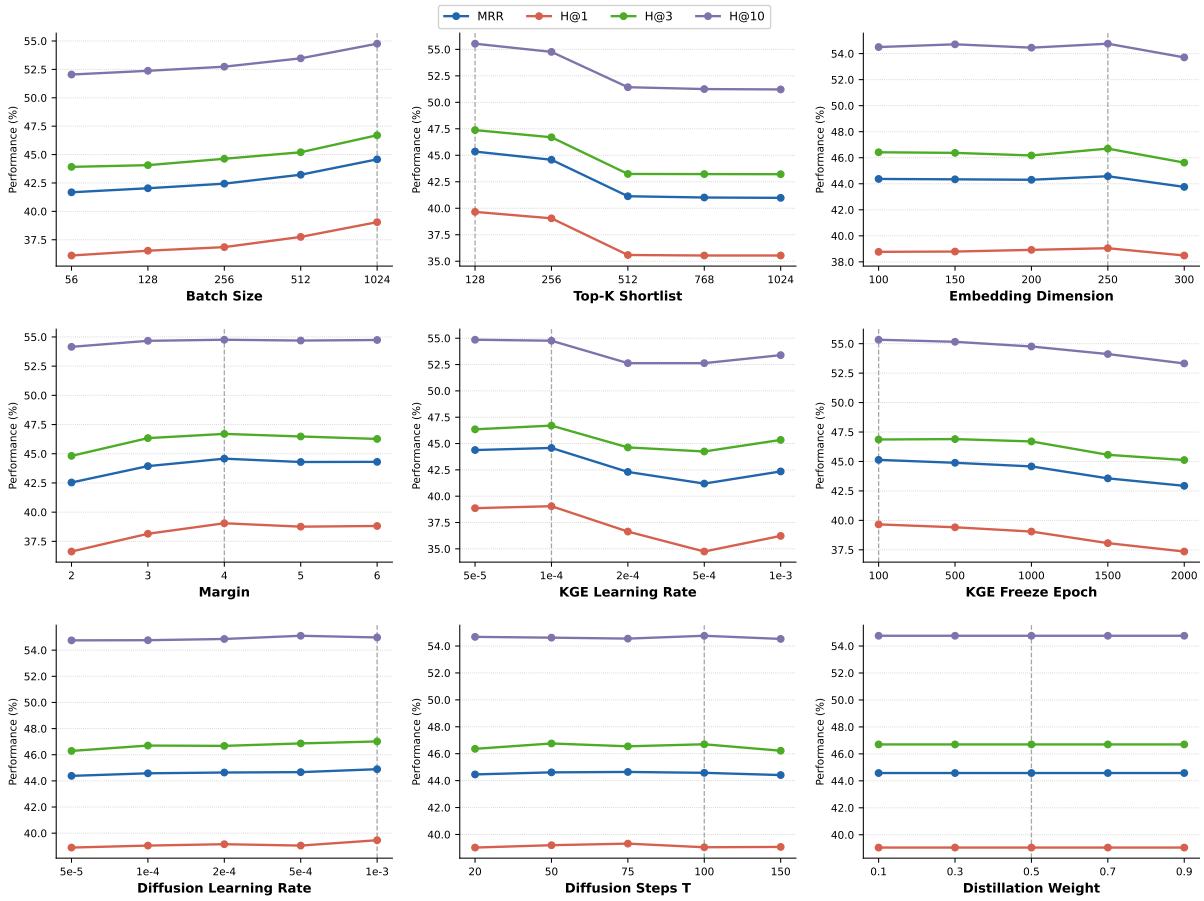

Figure 3: Hyperparameter sensitivity on MKG-W. Each subplot plots MRR, H@1, H@3, and H@10 (%) against one swept parameter. The dashed vertical line indicates the primary configuration used in the main experiments.

tonically (41.68→44.58), and embedding dimension is stable (MRR range: 0.8%).

KGE-specific parameters (Row 2). KGE learning rate shows the sharpest sensitivity (MRR range: 3.4%): raising it to 5×10^{-4} degrades H@1 by 3.6 pp with little effect on H@10, reflecting deteriorated top-rank geometry. Earlier retriever freezing is consistently better (epoch 100: 45.14% vs. epoch 2000: 42.93%), as it provides a stationary distillation target. Margin peaks at 4, with 2.0 pp loss at both extremes.

Diffusion-specific parameters (Row 3). All diffusion-side settings are nearly flat (MRR range <0.5% across all tested values): distillation weight shifts MRR by less than 0.1 pp over a 9× range, and step count T varies by only 0.2 pp over $T \in \{20, \ldots, 150\}$. Tuning effort can therefore be concentrated entirely on the retriever, while the denoiser performs near-optimally at its default configuration. Instance-level rank traces on representative queries from DB15K and MKG-W are presented in Appendix C, providing per-query evidence for the role specialization hypothesis.

6 Conclusion

We propose a Retrieval-Augmented Discrete Diffusion framework named RADD to address the challenge of dual-objective coupling in MMKGC by explicitly assigning full-set recall to a relation-aware multimodal KGE retriever and shortlist reranking to a conditional discrete denoiser over entity identities, with a Diff-Rerank mechanism at inference. Experiments on three benchmarks illustrate that RADD consistently outperforms strong unimodal, multimodal, and LLM-based baselines, and structure-only experiments confirm the decomposition holds without multimodal inputs. Ablations further verify that each component contributes distinctly, with gains scaling with shortlist quality and within-shortlist ambiguity. These results suggest that role specialization rather than representation scaling is the more effective path to accurate entity prediction for MMKGC.

7 Limitations

The most fundamental limitation of RADD is the hard recall. Diff-Rerank permanently excludes any entity not in the top-K shortlist, so retrieval misses are unrecoverable and overall performance is bounded by retriever recall. In terms of computational complexity, each query requires one KGE forward pass and one denoiser logit projection procedures, roughly doubling per-query latency relative to a single scorer. The retriever must warm up and stabilise before being frozen as the distillation teacher, introducing a serial dependency that increases total training time. Besides, embeddings are pre-extracted and fixed, and end-to-end fine-tuning of the encoders is a natural extension that may yield further gains on modality-rich datasets.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 5184–5193.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*.

Liwei Cai and William Yang Wang. 2018. Kbgan: Adversarial learning for knowledge graph embeddings. In *North American Chapter of the Association for Computational Linguistics*.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. In *Advances in Neural Information Processing Systems*.

Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. Pairre: Knowledge graph embeddings via paired relation vectors. In *Annual Meeting of the Association for Computational Linguistics*.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *Preprint*, arXiv:2402.05391.

Lingbing Guo, Zhongpu Bo, Zhuo Chen, Yichi Zhang, Jiaoyan Chen, Lan Yarong, Mengshu Sun, Zhiqiang Zhang, Yangyifei Luo, Qian Li, and 1 others. 2024. MKGL: Mastery of a three-word language. *Advances in Neural Information Processing Systems*, 37:140509–140534.

Lingbing Guo, Yichi Zhang, Zhongpu Bo, Zhuo Chen, Mengshu Sun, Zhiqiang Zhang, Wen Zhang, and Huajun Chen. 2025. K-ON: Stacking knowledge on the head layer of large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11745–11753.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

Wei Huang, Meiyu Liang, Peining Li, Xu Hou, Yawen Li, Junping Du, Zhe Xue, and Zeli Guan. 2025. Diffusioncom: Structure-aware multimodal diffusion model for multimodal knowledge graph completion. *Preprint*, arXiv:2504.06543.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 687–696.

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

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. Imf: Interactive multimodal fusion model for link prediction. In *The Web Conference*, pages 2572–2580.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. Mmkg: Multi-modal knowledge graphs. In *The Semantic Web*, pages 459–474.

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

Guanglin Niu and Xiaowei Zhang. 2026. Diffusion-based hierarchical negative sampling for multimodal knowledge graph completion. In *Database Systems for Advanced Applications*, pages 479–495, Singapore. Springer Nature Singapore.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding multimodal relational data for knowledge base completion. In *Conference on Empirical Methods in Natural Language Processing*, pages 3208–3218.

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. In *Annual Meeting of the Association for Computational Linguistics*, pages 2814–2828.

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.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *Annual Meeting of the Association for Computational Linguistics*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

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.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

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.

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

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *International Joint Conference on Artificial Intelligence*, pages 3140–3146.

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.

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 *International Conference on Learning Representations*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *Preprint*, arXiv:1909.03193.

Yichi Zhang, Mingyang Chen, and Wen Zhang. 2023. Modality-aware negative sampling for multi-modal knowledge graph embedding. In *International Joint Conference on Neural Networks*, pages 1–8.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2024a. Native: Multi-modal knowledge graph completion in the wild. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

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

Yichi Zhang and Wen Zhang. 2022. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *Preprint*, arXiv: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 *Conference on Empirical Methods in Natural Language Processing*, pages 10527–10536.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems*, pages 29476–29490.

Algorithm 1 RADD Training (one epoch, one batch)

Require: Mini-batch $\mathcal{B} = \{(h_i, r_i, t_{e,i})\}$, current epoch e, freeze threshold τ_f, head-loss weight λ_h, distillation temperature τ, weight λ_d

// Step 1: Retriever forward
for each $(h, r, t_e) \in \mathcal{B}$ **do**
  Compute relation-gated representation $\mathbf{e}^{(\text{joint})}$ via Eq. (1)
  Score all entities and sample adversarial negatives
  Accumulate $\mathcal{L}_{\text{kge}}$ (margin + adversarial)
end for
if $e \leq \tau_f$ **then**
  Update $\theta_{\text{ret}} \leftarrow \theta_{\text{ret}} - \eta \nabla \mathcal{L}_{\text{kge}}$ ▷ retriever frozen after epoch τ_f
end if
// Step 2: Denoiser forward (bidirectional)
for each $(h, r, t_e) \in \mathcal{B}$ **do**
  Sample $t' \sim \text{Uniform}\{1, \ldots, T\}$
  Tail prediction: corrupt $x_0{=}t_e$ to $x_{t'}$; set $\mathbf{c}_h = [\mathbf{e}_h^{(\text{joint})}; \mathbf{r}; \mathbf{e}_{x_{t'}}^{(\text{emb})}; \phi(t'); \mathbf{d}_{\text{tail}}]$
  $\mathcal{L}_{\text{tail}} \leftarrow \text{CE}\big(p_\theta(x_0 \mid \mathbf{c}_h),\ t\big)$
  Head prediction: corrupt $x_0{=}h$ to $x_{t'}$; set $\mathbf{c}_{t_e} = [\mathbf{e}_{t_e}^{(\text{joint})}; \mathbf{r}; \mathbf{e}_{x_{t'}}^{(\text{emb})}; \phi(t'); \mathbf{d}_{\text{head}}]$
  $\mathcal{L}_{\text{head}} \leftarrow \text{CE}\big(p_\theta(x_0 \mid \mathbf{c}_{t_e}),\ h\big)$
  $\mathcal{L}_{\text{diff}} \leftarrow \mathcal{L}_{\text{tail}} + \lambda_h \mathcal{L}_{\text{head}}$ (Eq. 5)
  $\mathcal{L}_{\text{distill}} \leftarrow \text{KL}\big(p_T(\mathcal{C}) \parallel p_D(\mathcal{C})\big)$ (Eq. 7)
end for
Update $\theta_{\text{den}} \leftarrow \theta_{\text{den}} - \eta \nabla(\mathcal{L}_{\text{diff}} + \lambda_d \mathcal{L}_{\text{distill}})$

Algorithm 2 Diff-Rerank Inference

Require: Query $(h, r, ?)$ or $(?, r, t_e)$, shortlist size K, diffusion steps T
Ensure: Ranked candidate list

// Stage 1: Retrieval
Compute retriever score $s(e)$ for all $e \in \mathcal{E}$
$\mathcal{S} \leftarrow \text{top-}K\big(\{s(e) : e \in \mathcal{E}\}\big)$ ▷ hard shortlist
// Stage 2: Denoising rerank
Set conditioning context $\mathbf{c}$ from query
Initialise $x_T = [\texttt{MASK}]$
for $t = T, T{-}1, \ldots, 1$ **do**
  Compute posterior $p_\theta(x_0 \mid x_t, \mathbf{c})$ restricted to $\mathcal{S}$
  Sample $x_{t-1} \sim p_\theta(x_{t-1} \mid x_t, \mathbf{c})$
end for
Score each $e_i \in \mathcal{S}$ by $p_\theta(e_i \mid x_1, \mathbf{c})$
return $\mathcal{S}$ reranked by denoiser scores

Table 5: Results on two traditional unimodal KGC datasets. RADD operates in structure-only mode. Best in **bold**, second-best underlined.

Model	FB15K-237				NELL995			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
TransE	33.10	23.40	36.90	52.60	23.90	7.90	34.80	50.80
DistMult	28.40	22.30	31.00	44.70	34.80	26.80	38.50	50.00
ComplEx	30.80	22.00	33.70	48.60	36.90	28.90	40.50	51.80
RotatE	33.60	24.00	37.40	53.00	35.80	26.90	40.40	52.60
RADD	**34.80**	**24.80**	**38.00**	**56.20**	**48.20**	**43.30**	**51.10**	**55.90**

A Training and Inference Algorithms

Algorithm 1 and Algorithm 2 describe the RADD training procedure and Diff-Rerank inference, respectively. The training procedure jointly optimises the retriever and denoiser via three losses: KGE margin loss $\mathcal{L}_{\text{kge}}$, bidirectional denoising cross-entropy $\mathcal{L}_{\text{diff}}$, and temperature-scaled distillation $\mathcal{L}_{\text{distill}}$ from the retriever's soft rankings to the denoiser. The retriever is frozen after epoch τ_f to provide a stationary distillation teacher. At inference, Diff-Rerank enforces a strict two-stage decomposition: the retriever commits to a hard top-K shortlist, and the denoiser reranks exclusively within it. The source codes will be released with MIT License upon publication for research.

B Traditional KGC Results

To assess whether RADD's advantage is rooted in its architecture or merely in its access to multimodal signals, Table 5 evaluates RADD in structure-only mode on two conventional KGC benchmarks where no visual or textual features are available. On FB15K-237, RADD reaches 34.80 MRR, surpassing the strongest classical baseline RotatE by +1.2, with consistent improvements across all metrics. On NELL995, the margin is substantially wider: RADD achieves 48.20 MRR, exceeding the best baseline ComplEx by +11.3 and raising H@1 by +14.4. These results illustrate that RADD's core benefit is architectural and modality-agnostic.

The contrast performance between the two datasets is informative. The smaller but uniform gain on FB15K-237 indicates that retrieve-then-rerank remains useful even when strong geometric baselines already provide competitive candidate ordering. The much larger gain on NELL995 suggests that shortlist-level denoising becomes more valuable when structural retrieval leaves greater residual ambiguity. From the RADD-S results in Table 2, these experiments confirm that the

Table 6: Filtered rank of the correct entity for representative DB15K and MKG-W queries. **Ret.** = retrieval-only, **Den.** = denoiser-only, **RADD** = full model.

Query	Correct Entity	Ret.	Den.	RADD
DB15K				
(?, *type*, *Island*)	*K. Balachander*	506	1066	**1**
(?, *country*, *United_States*)	*Huntington WV*	462	387	**1**
(*Larry Klein*, *spouse*, ?)	*Joni_Mitchell*	23	2	**1**
(*Stuyvesant HS*, *grades*, ?)	*Ninth_grade*	58	2	**1**
MKG-W				
(?, *P495*, *Q142*)	*Q15961525*	241	360	**1**
(?, *P31*, *Q11424*)	*Q897112*	236	747	**1**
(*Q1235040*, *P17*, ?)	*Q145*	146	3	**1**
(*Q207176*, *P131*, ?)	*Q25*	5	2	**1**

retrieve-then-rerank decomposition is the primary source of RADD's gains: it delivers consistent improvements even without multimodal evidence, and multimodal signals provide a further additive boost on top of that architectural foundation.

C Case Study

Table 6 reports the rank of the correct entity e^* for eight representative queries across DB15K and MKG-W, covering both head and tail prediction. RADD achieves rank 1 in every case, while neither single-module baseline does so consistently.

Head prediction: both modules are individually necessary. In all four head queries, retrieval-only and denoiser-only both produce low ranks (236–1066). Retrieval-only fails because KGE scoring over the full entity vocabulary cannot separate e^* from hundreds of structurally similar head candidates; for instance, DB15K's *rdf:type/Island* relation has 7638 confusable entities, creating a dense embedding neighbourhood that global similarity cannot resolve. Denoiser-only fails for a complementary reason: without a retrieval shortlist to bound its scope, the correct head entity is too weakly signalled to emerge at rank 1. RADD resolves the impasse: the retriever elevates e^* into the top-K shortlist, and the denoiser exploits shortlist-conditioned context to distinguish e^* from the remaining candidates.

Tail prediction: each module contributes complementary precision. Retrieval-only already places e^* at moderate rank (23–58 on DB15K; 5–146 on MKG-W), and denoiser-only reaches near-rank-1 in several cases (rank 2–3). Yet neither consistently achieves rank 1. RADD does so in all four cases: the retriever ensures e^* is included in the shortlist, and the denoiser elevates it above the retriever's residual errors, confirming that fine-grained shortlist conditioning, not denoising power alone, is the decisive factor.

Implication. These typical cases provide direct instance-level evidence for RADD's core hypothesis: effective MMKGC requires recall-oriented retrieval and precision-oriented shortlist denoising. Consistent rank-1 outcomes across diverse queries, two datasets, and both prediction directions confirm that role specialization is the structural source of RADD's aggregate advantages.