ROLE-AWARE HIERARCHICAL SUB-ENTITY ALIGNMENT AND DIFFICULTY-WEIGHTED NEGATIVE SAMPLING FOR MULTI-MODAL KNOWLEDGE GRAPH COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal knowledge graph completion (MMKGC) aims to exploit visual, textual, and structural evidence to infer missing facts, but existing methods still suffer from two key limitations: (i) they fuse modalities mostly at the entity level, ignoring relation- and role-specific fine-grained alignment between sub-entities (e.g., phrases and image regions), and (ii) hard negative sampling often amplifies false-negative bias when plausible but unobserved triples are treated as negative samples. Therefore, we propose RHDN, a unified framework that jointly addresses these issues. RHDN introduces a role-aware hierarchical alignment module that decomposes entities into multi-modal sub-entities, constructs relation- and role-conditioned optimal transport couplings between visual and textual evidence, and aggregates coupling-weighted pairwise interactions with hierarchical role prototypes to obtain relation-aware entity representations. On the training side, we decouple negative difficulty from false-negative risk using model scores and multimodal role-consistency evidence, respectively, and combine them through KL-regularized reweighting. Extensive experiments on several multi-modal KG benchmarks show that RHDN consistently outperforms strong baselines.

1 INTRODUCTION

Multi-modal knowledge graphs (MMKGs) extend conventional knowledge graphs (KGs) by associating entities with heterogeneous information such as images and textual descriptions Liu et al. (2019). By combining structured triples with multimodal evidence, MMKGs provide richer entity semantics and have been applied to recommender systems Sun et al. (2020), large language models Liu et al. (2025), and medical diagnosis Li et al. (2025a). However, MMKGs remain incomplete, making multi-modal knowledge graph completion (MMKGC) important for inferring missing facts. Compared with traditional KGC methods that mainly rely on graph structure Li & Huang (2026a), MMKGC jointly exploits structural, visual, and textual signals for link prediction Li & Huang (2026b).

Despite recent progress, existing MMKGC methods still face two limitations. First, most approaches perform cross-modal fusion at the coarse entity level, compressing visual regions and textual descriptions into global embeddings that are shared across different relations and semantic roles Gao et al. (2025); Shang et al. (2024b). This overlooks that only a subset of multimodal evidence may be relevant to a specific relation role. For example, when predicting the award of a film, a phrase such as "Academy Award for Best Picture" can be more informative than visually salient but relation-irrelevant regions. As illustrated in Figure 1, coarse fusion may therefore allow dominant but irrelevant evidence to overwhelm fine-grained relation-relevant cues. Second, MMKG incompleteness introduces false-negative bias during training. Corrupted triples used as negatives may correspond to unobserved true facts, especially when strong multimodal encoders assign high plausibility scores to semantically valid but missing triples Wei et al. (2024). Conventional hard negative sampling may mistakenly emphasize such candidates and impose large penalties, resulting in misleading supervision. Therefore, an effective MMKGC model should preserve informative hard negatives while reducing the influence of likely false negatives.

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.

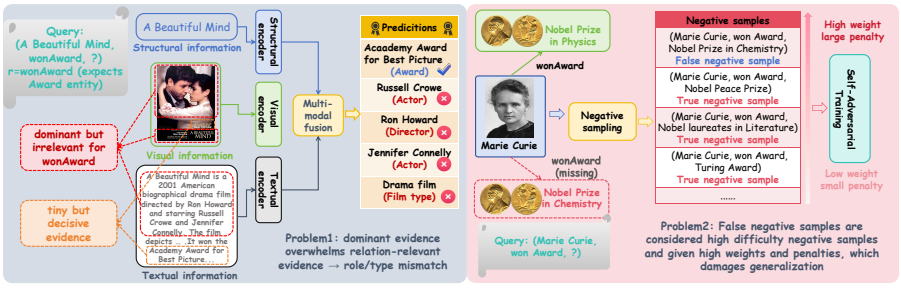


Figure 1: Two remaining problems in current MMKGC methods.

To address these issues, we propose **RHDN**, a unified framework for **R**ole-aware **H**ierarchical sub-entity alignment and **D**ifficulty-weighted **N**egative sampling. RHDN consists of two complementary components. First, **R-HAlign** constructs relation- and role-conditioned visual/textual sub-entity distributions and computes an optimal transport coupling between them. The transport plan is then used to aggregate pair-specific cross-modal interactions, while hierarchical role prototypes further inject relation- and cluster-level semantic guidance, yielding fine-grained relation-aware entity representations. Second, **FDNS** explicitly decouples negative difficulty from false-negative risk: model scores measure hardness, whereas relation- and role-conditioned multimodal evidence estimates false-negative risk. A KL-regularized reweighting objective then preserves informative hard negatives while suppressing candidates that are more likely to be unobserved true facts. Together, the two modules improve both multimodal representation quality and the reliability of training supervision.

Our main contributions are summarized as follows:

- We propose **R-HAlign**, a role-aware hierarchical sub-entity alignment module that constructs relation- and role-conditioned optimal transport couplings and aggregates coupling-weighted pairwise cross-modal interactions, enabling fine-grained relation-aware multimodal representation learning.
- We develop **FDNS**, a false-negative-aware difficulty-weighted negative sampling strategy that decouples model-based difficulty from multimodal evidence-based false-negative risk and combines them through KL-regularized reweighting.
- We integrate R-HAlign and FDNS into the end-to-end **RHDN** framework. Experiments on three widely used MMKGC benchmarks demonstrate strong and consistent improvements over unimodal, multimodal, and negative-sampling baselines, with further analyses validating the contributions of the key components.

2 Related Work

2.1 Unimodal KGC Methods.

Due to space limitations, we include a review of current unimodal knowledge graph completion methods in the Appendix.

2.2 Multi-Modal KGC Methods.

To overcome the limitation of unimodal KGC in modeling non-structural information, MMKGC has been widely studied. Existing methods typically integrate visual, textual, and structural features into unified entity representations for triple scoring. Early approaches, such as IKRL Xie et al. (2017) and TransAE Wang et al. (2019), extend translation-based KGC models with multimodal features. More recent methods design stronger fusion mechanisms: MoSE Zhao et al. (2022) ensembles modality-specific predictors; IMF Li et al. (2023) decouples and recombines modality-specific embeddings; VISTA Lee et al. (2023) adopts transformer-based encoders for visual-textual fusion; LAFA Shang et al. (2024a) introduces link-aware fusion to reduce irrelevant visual noise; and NativE Zhang et al. (2024a) enables relation-adaptive fusion across arbitrary modalities. Further, SNAG

Chen et al. (2025) improves robustness via noise masking, AdaMF Zhang et al. (2024b) learns adaptive modality weights with adversarial training, and MoMoK Zhang et al. (2025a) employs relation-guided mixture-of-experts for multi-view fusion. Recent methods such as MyGO Zhang et al. (2025b), C^{2}RS Shu et al. (2025), and IMVIA Li et al. (2025b) further enhance fine-grained semantics, cross-modal consistency, and relation-context-aware feature selection. Despite these advances, most methods still rely on coarse entity-level fusion and insufficiently model relation- and role-specific fine-grained cross-modal alignment, which may limit representation quality and final link prediction performance.

2.3 Negative Sampling strategies in MMKGC

Negative sampling is crucial for training MMKGC models. While random replacement is simple, it often produces overly easy negatives with limited training value. Recent MMKGC studies therefore exploit multimodal information to construct harder negatives. For example, MMRNS Xu et al. (2022) uses knowledge-guided cross-modal attention to generate semantically relevant negatives; MANS Zhang et al. (2023) aligns structural embeddings with visual features to identify challenging candidates; and DHNS Niu & Zhang (2025) employs a diffusion model to generate negatives with diverse difficulty levels. Although these methods improve over random sampling, they generally treat high-scoring candidates as informative hard negatives, without distinguishing them from false negatives, i.e., unobserved but true facts. Such ambiguity may introduce misleading supervision and degrade model performance.

3 Methodology

3.1 Problem Definition

A knowledge graph (KG) is denoted as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ denote the entity, relation, and triple sets, respectively. A multi-modal KG (MMKG) further associates entities with auxiliary modalities such as images and textual descriptions, and is denoted as $\mathcal{G}' = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{M})$. Given a query $(h, r, ?)$ or $(?, r, t)$, multi-modal knowledge graph completion (MMKGC) aims to rank the missing entity by jointly exploiting structural and multi-modal evidence. We use a scoring function $f : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$, where a larger value indicates a more plausible triple. Our goal is to address two complementary issues in this setting: how to identify relation- and role-relevant fine-grained multimodal evidence, and how to prevent plausible but unobserved facts from being overly penalized during negative-sample training.

3.2 Overview

As illustrated in Fig. 2, **RHDN** jointly improves multimodal representation learning and negative supervision. We first encode structural, visual, and textual information into a coarse multimodal entity representation. Since coarse fusion cannot identify relation- and role-relevant sub-entity evidence, **R-HAlign** constructs relation- and role-conditioned visual/textual distributions, computes a many-to-many optimal transport (OT) coupling, and aggregates pair-specific cross-modal interactions according to the transport plan. The resulting coupling-sensitive representation is further integrated with relation- and cluster-level role prototypes for link prediction. During training, **FDNS** decouples negative difficulty from false-negative risk: model scores measure hardness, while multimodal role-consistency evidence independently estimates false-negative risk. A KL-regularized reweighting scheme then preserves informative hard negatives while reducing the influence of high-risk candidates.

3.3 Multimodal Encoders and Coarse Entity Representation

Modality-specific encoders. For each entity $e \in \mathcal{E}$, we encode visual, textual, and structural information as

$$\mathbf{E}_e^v = [\mathbf{v}_{e,1}; \ldots; \mathbf{v}_{e,N_e^v}] \in \mathbb{R}^{N_e^v \times d}, \mathbf{E}_e^x = [\mathbf{x}_{e,1}; \ldots; \mathbf{x}_{e,N_e^x}] \in \mathbb{R}^{N_e^x \times d}, \mathbf{E}_e^s \in \mathbb{R}^d, \tag{1}$$

where $\mathbf{v}_{e,i}$ and $\mathbf{x}_{e,j}$ denote the i-th visual region and j-th textual token or phrase, respectively. We employ ViT-, BERT-, and RGAT-based encoders for the three modalities.

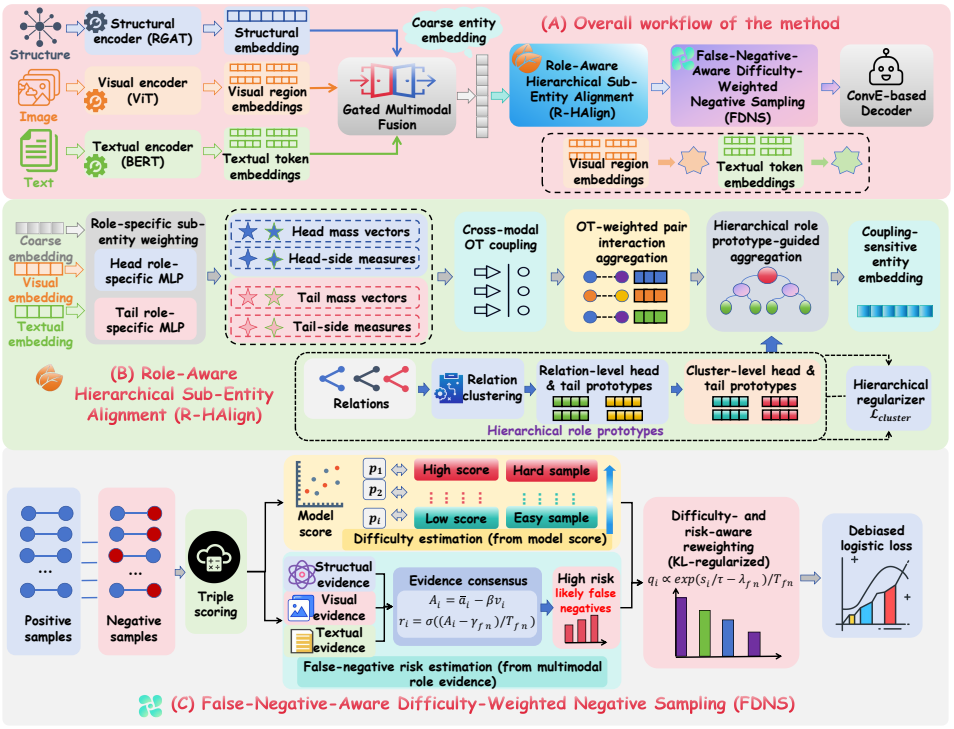


Figure 2: The overall framework of RHDN.

The region- and token-level features are retained for subsequent fine-grained alignment. Meanwhile, we also construct a coarse entity representation to summarize the overall multimodal semantics and provide a stable entity-level anchor.

Coarse multimodal fusion. We first obtain modality-level representations by

$$\mathbf{u}_e^{\text{vis}} = \frac{1}{N_e^v}\sum_{i=1}^{N_e^v}\mathbf{v}_{e,i}, \quad \mathbf{u}_e^{\text{text}} = \frac{1}{N_e^x}\sum_{j=1}^{N_e^x}\mathbf{x}_{e,j}, \quad \mathbf{u}_e^{\text{str}} = \mathbf{E}_e^s. \tag{2}$$

Since the reliability and usefulness of different modalities vary across entities, we use an adaptive gating mechanism rather than uniform averaging:

$$\alpha_e^{(m)} = \frac{\exp(\mathbf{w}_m^\top \mathbf{g}_e)}{\sum_{m'\in\{\text{vis,text,str}\}}\exp(\mathbf{w}_{m'}^\top \mathbf{g}_e)}, \quad \mathbf{g}_e = \text{MLP}_g[\mathbf{u}_e^{\text{vis}} \| \mathbf{u}_e^{\text{text}} \| \mathbf{u}_e^{\text{str}}]. \tag{3}$$

The coarse entity representation is then

$$\mathbf{b}_e = \sum_{m\in\{\text{vis,text,str}\}} \alpha_e^{(m)}\mathbf{u}_e^{(m)}. \tag{4}$$

This representation captures global entity semantics, but remains relation-agnostic. We therefore use it only as an anchor and perform relation-specific reasoning over the retained sub-entity features in the next stage.

3.4 R-HALIGN: ROLE-AWARE HIERARCHICAL SUB-ENTITY ALIGNMENT

R-HAlign transforms coarse multimodal evidence into relation- and role-specific entity representations. Its central idea is to first determine *which* visual and textual sub-entities are relevant to the current relation role, and then model *how* those selected sub-entities correspond across modalities.

Hierarchical role prototypes. A relation imposes different semantic requirements on its head and tail entities. For example, the head of *starring* is typically a film, whereas its tail is typically a person. To explicitly encode this asymmetry, we maintain role-specific prototypes.

Let $\mathcal{C}$ denote a set of relation clusters and $C : \mathcal{R} \to \mathcal{C}$ map relation r to its cluster. For each relation and cluster, we introduce learnable head/tail role prototypes $\mathbf{u}_r^{(h)}, \mathbf{u}_r^{(t)} \in \mathbb{R}^d$ and $\mathbf{u}_c^{(h)}, \mathbf{u}_c^{(t)} \in \mathbb{R}^d$. Relation-level prototypes capture relation-specific semantics, while cluster-level prototypes provide a coarser semantic prior shared by related relations. We encourage consistency between the two levels through

$$\mathcal{L}_{\text{cl}} = \sum_{r \in \mathcal{R}} \left(\|\mathbf{u}_r^{(h)} - \mathbf{u}_{C(r)}^{(h)}\|_2^2 + \|\mathbf{u}_r^{(t)} - \mathbf{u}_{C(r)}^{(t)}\|_2^2 \right). \tag{5}$$

Relation- and role-aware sub-entity weighting. The hierarchical prototypes specify what semantic role an entity is expected to play, but different visual regions and textual tokens contribute unequally to that role. We therefore construct a role-conditioned context for the head entity:

$$\mathbf{c}_r^{(h)} = \text{MLP}^{(h)} \left([\mathbf{r} \| \mathbf{u}_r^{(h)} \| \mathbf{b}_h] \right), \tag{6}$$

where $\mathbf{r}$ is the relation embedding and $\mathbf{b}_h$ provides entity-level context.

We use this context to independently normalize visual and textual evidence:

$$\pi_{h,i}^{(r)} = \frac{\exp((\mathbf{W}_v \mathbf{v}_{h,i})^\top \mathbf{c}_r^{(h)})}{\sum_{k=1}^{N_h^v} \exp((\mathbf{W}_v \mathbf{v}_{h,k})^\top \mathbf{c}_r^{(h)})}, \rho_{h,j}^{(r)} = \frac{\exp((\mathbf{W}_x \mathbf{x}_{h,j})^\top \mathbf{c}_r^{(h)})}{\sum_{k=1}^{N_h^x} \exp((\mathbf{W}_x \mathbf{x}_{h,k})^\top \mathbf{c}_r^{(h)})}. \tag{7}$$

Thus, $\boldsymbol{\pi}_h^{(r)} \in \Delta_{N_h^v}$ and $\boldsymbol{\rho}_h^{(r)} \in \Delta_{N_h^x}$ define two valid relation-conditioned probability measures:

$$\nu_{h,r}^{\text{vis}} = \sum_{i=1}^{N_h^v} \pi_{h,i}^{(r)} \delta_{\mathbf{v}_{h,i}}, \qquad \nu_{h,r}^{\text{text}} = \sum_{j=1}^{N_h^x} \rho_{h,j}^{(r)} \delta_{\mathbf{x}_{h,j}}. \tag{8}$$

The tail-side measures are constructed analogously using the tail-role prototype. These distributions identify relation-relevant evidence within each modality, but do not yet specify the correspondence between individual visual and textual sub-entities. We address this with OT.

Cross-modal OT-guided pairwise interaction. Given the role-aware visual and textual measures, we seek a soft many-to-many correspondence between their sub-entities. For the head entity, the pairwise transport cost is

$$C_{ij}^h = \|\mathbf{v}_{h,i} - \mathbf{x}_{h,j}\|_2^2. \tag{9}$$

We then compute the entropy-regularized transport plan

$$\mathbf{T}^{h,*} = \arg \min_{\mathbf{T}^h \in \Pi(\boldsymbol{\pi}_h^{(r)}, \boldsymbol{\rho}_h^{(r)})} \left[\sum_{i,j} T_{ij}^h C_{ij}^h - \varepsilon_{\text{OT}} \mathcal{H}(\mathbf{T}^h) \right], \tag{10}$$

where $\Pi(\boldsymbol{\pi}, \boldsymbol{\rho})$ denotes the set of transport plans with marginals $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$, and $\mathcal{H}(\mathbf{T}) = -\sum_{i,j} T_{ij} \log T_{ij}$. Eq. equation 10 is solved using Sinkhorn iterations. A key design choice is that we do not directly pool visual and textual features using the transport weights. Such a separable aggregation can collapse to marginal-weighted pooling and discard the internal coupling structure. Instead, we first construct an interaction representation for each visual–textual pair:

$$\begin{aligned} \mathbf{p}_{h,ij}^{(r)} &= \phi_{\text{pair}} \left([\mathbf{v}_{h,i} \| \mathbf{x}_{h,j} \| \mathbf{v}_{h,i} \odot \mathbf{x}_{h,j} \| |\mathbf{v}_{h,i} - \mathbf{x}_{h,j}| \| \mathbf{c}_r^{(h)}] \right), \\ \mathbf{m}_h^{(r)} &= \sum_{i,j} T_{ij}^{h,*} \mathbf{p}_{h,ij}^{(r)}, \widetilde{\mathbf{z}}_h^{(r)} = \mathbf{W}_m \mathbf{m}_h^{(r)} + \mathbf{W}_b \mathbf{b}_h. \end{aligned} \tag{11}$$

Here, ϕ_{pair} maps each cross-modal pair into a d-dimensional interaction embedding. Because the interaction contains non-separable terms such as $\mathbf{v}_{h,i} \odot \mathbf{x}_{h,j}$, the representation in Eq. equation 11 depends on the internal structure of $\mathbf{T}^{h,*}$ rather than only on its marginals. Hence, OT contributes explicit pairwise correspondence information to the entity representation. The tail-side representation $\widetilde{\mathbf{z}}_t^{(r)}$ is computed analogously. To further encourage semantically coherent cross-modal correspondences, we regularize positive triples by the expected transport cost:

$$\mathcal{L}_{\text{OT}}(h, r, t) = \sum_{i,j} T_{ij}^{h,*} C_{ij}^h + \sum_{i,j} T_{ij}^{t,*} C_{ij}^t. \tag{12}$$

Thus, the transport plan plays two complementary roles: it explicitly enters representation learning through Eq. equation 11, while $\mathcal{L}_{\mathrm{OT}}$ encourages cross-modal consistency.

Hierarchical role aggregation. The OT representation captures instance-specific cross-modal correspondence, whereas the hierarchical prototypes encode relation-level semantic priors. We combine the two sources of information as

$$\mathbf{z}_h^{(r)} = \sigma\left(\mathbf{W}_z^{(h)}[\widetilde{\mathbf{z}}_h^{(r)} \| \mathbf{u}_r^{(h)} \| \mathbf{u}_{C(r)}^{(h)}]\right), \tag{13}$$

with an analogous formulation for $\mathbf{z}_t^{(r)}$. We further encourage the resulting instance-level representation to remain consistent with the corresponding relation role:

$$\mathcal{L}_{\mathrm{pro}}(h, r, t) = \|\mathbf{z}_h^{(r)} - \mathbf{u}_r^{(h)}\|_2^2 + \|\mathbf{z}_t^{(r)} - \mathbf{u}_r^{(t)}\|_2^2. \tag{14}$$

Decoder. Finally, a ConvE-style decoder transforms the relation-aware head representation and relation embedding into a query representation $\mathbf{c}_{h,r}$. The plausibility of a candidate tail is

$$f(h, r, t) = \mathbf{c}_{h,r}^\top \mathbf{z}_t^{(r)}. \tag{15}$$

The resulting score is used both for link prediction and for estimating negative difficulty in FDNS.

3.5 FDNS: False-Negative-Aware Difficulty-Weighted Negative Sampling

R-HAlign improves the quality of multimodal representations, but stronger representations can also assign high scores to plausible triples that are missing from an incomplete KG. Treating all such high-scoring candidates as hard negatives may therefore introduce misleading supervision. FDNS addresses this issue by explicitly distinguishing *difficulty* from *false-negative risk*.

Difficulty estimation. For a positive triple (h, r, t), let $\mathcal{N}(h, r, t) = \{x_i\}_{i=1}^N$ denote the candidate negatives obtained by head or tail corruption after removing duplicates and observed training positives. For each candidate $x_i = (h_i, r, t_i)$, we compute

$$s_i = f(h_i, r, t_i), \qquad p_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^N \exp(s_j/\tau)}, \tag{16}$$

where $\tau > 0$ controls how strongly the distribution emphasizes high-scoring negatives. A large p_i therefore indicates that x_i is difficult under the current model, but does not determine whether it is actually a valid negative.

Score-decoupled false-negative risk. To obtain an independent risk signal, we evaluate whether the corrupted entity is supported by multiple modalities under its expected semantic role. Let e_i denote the corrupted entity and $\kappa_i \in \{h, t\}$ its role in candidate x_i. For every available modality $m \in \mathcal{M}_i \subseteq \{\mathrm{str}, \mathrm{vis}, \mathrm{text}\}$, we compute

$$a_i^{(m)} = \frac{1}{2}\left[1 + \cos\left(\mathbf{u}_{e_i}^{(m)}, \mathbf{u}_r^{(\kappa_i)}\right)\right], \tag{17}$$

which measures how well modality m supports the expected relation role.

A plausible unobserved fact should ideally receive not only high support but also consistent support across available modalities. We therefore summarize multimodal evidence by

$$\bar{a}_i = \frac{1}{|\mathcal{M}_i|}\sum_{m\in\mathcal{M}_i} a_i^{(m)}, v_i = \frac{1}{|\mathcal{M}_i|}\sum_{m\in\mathcal{M}_i}(a_i^{(m)} - \bar{a}_i)^2, A_i = \bar{a}_i - \beta v_i, \qquad \beta \geq 0. \tag{18}$$

The first term rewards strong multimodal support, whereas the disagreement penalty prevents a single highly activated modality from dominating the risk estimate. We map the evidence-consensus score into a smooth false-negative risk:

$$r_i = P_{\mathrm{FN}}(x_i) = \sigma\left(\frac{A_i - \gamma_{\mathrm{fn}}}{T_{\mathrm{fn}}}\right), \tag{19}$$

where γ_{fn} is a risk threshold and $T_{\mathrm{fn}} > 0$ controls smoothness. Crucially, r_i is not a deterministic function of s_i. Thus, two candidates with similar model scores can receive different risks when their multimodal role evidence differs.

Difficulty–risk reweighting. We next combine the two signals without discarding the useful ranking induced by hard-negative mining. Specifically, we seek a distribution $\mathbf{q} \in \Delta_N$ that remains close to the difficulty distribution $\mathbf{p}$ while penalizing high-risk candidates:

$$\mathbf{q}^* = \arg\min_{\mathbf{q}\in\Delta_N} \left[\mathrm{KL}(\mathbf{q}\|\mathbf{p}) + \lambda_{\mathrm{fn}} \sum_{i=1}^{N} q_i r_i \right]. \tag{20}$$

The closed-form solution is

$$q_i = \frac{p_i \exp(-\lambda_{\mathrm{fn}} r_i)}{\sum_j p_j \exp(-\lambda_{\mathrm{fn}} r_j)} = \frac{\exp(s_i/\tau - \lambda_{\mathrm{fn}} r_i)}{\sum_j \exp(s_j/\tau - \lambda_{\mathrm{fn}} r_j)}. \tag{21}$$

This formulation preserves high-scoring candidates when their risk is low, while selectively reducing the influence of multimodally plausible candidates.

Debiased link-prediction loss. Finally, we optimize

$$\mathcal{L}_{\mathrm{link}} = -\log \sigma(f(h,r,t)) - \sum_{x_i \in \mathcal{N}(h,r,t)} \mathrm{sg}(q_i) \log\left(1 - \sigma(s_i)\right), \tag{22}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Detaching q_i prevents the model from trivially manipulating the weighting distribution to reduce the negative loss. In this way, FDNS changes the strength of supervision rather than allowing the risk estimator itself to become an optimization shortcut.

3.6 Overall Objective

RHDN is trained end-to-end by combining link prediction with the alignment and prototype regularizers:

$$\mathcal{L}_{\mathrm{total}} = \sum_{(h,r,t)\in\mathcal{T}} [\mathcal{L}_{\mathrm{link}} + \lambda_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}} + \lambda_{\mathrm{pro}}\mathcal{L}_{\mathrm{pro}}] + \lambda_{\mathrm{cl}}\mathcal{L}_{\mathrm{cl}}, \tag{23}$$

where λ_{OT}, λ_{pro}, and λ_{cl} balance OT consistency, instance–prototype alignment, and hierarchical role regularization, respectively. Together, R-HAlign improves relation- and role-aware multimodal representation, while FDNS provides more reliable negative supervision for learning these representations.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocol. To demonstrate the effectiveness of our method, we conduct experiments on three widely used public MMKG benchmarks: **DB15K** Liu et al. (2019), **MKG-W** Xu et al. (2022), and **MKG-Y** Xu et al. (2022). Detailed statistics for these datasets are provided in Appendix. To evaluate the performance of our method on MMKGC, we conduct experiments on the link prediction task. Specifically, we adopt ranking-based evaluation metrics—such as Mean Reciprocal Rank (**MRR**) and **Hit@K** (K = 1, 3, 10)—to quantitatively assess link prediction performance, considering both head entity prediction $(?, r, t)$ and tail entity prediction $(h, r, ?)$. All reported results are averaged over three independent runs with different random seeds.

Baselines and implementation details. To conduct a comprehensive comparison, we select more than ten representative methods across three categories as experimental baselines. The first category is unimodal KGC methods, including: TransE Wang et al. (2014), DistMult Yang et al. (2015), ComplEx Trouillon et al. (2017), and RotatE Sun et al. (2019). The second category is MMKGC methods, including: IKRL Xie et al. (2017), TransAE Wang et al. (2019), MoSE Zhao et al. (2022), IMF Li et al. (2023), VISTA Lee et al. (2023), NativE Zhang et al. (2024a), SNAG Chen et al. (2025), AdaMF Zhang et al. (2024b), MoMoK Zhang et al. (2025a), MyGO Zhang et al. (2025b), C^{2}RS Shu et al. (2025), and IMVIA Li et al. (2025b). The third category is negative sampling strategies specifically designed for MMKGC, including: MMRNS Xu et al. (2022), MANS Zhang et al. (2023), and DHNS Niu & Zhang (2025). Detailed implementation settings and hyperparameter configurations are provided in Appendix.

4.2 MAIN RESULTS

As shown in Table 1, on all three multimodal KGC benchmarks (DB15K, MKG-W, MKG-Y), RHDN delivers consistent gains over unimodal KGC, MMKGC, and negative sampling baselines. On DB15K, RHDN attains the best MRR, Hits@1, and Hits@10 while remaining highly competitive on Hits@3, outperforming strong baselines such as MoMoK and C^{2}RS in most metrics. On the more challenging MKG-W, RHDN sets a new state of the art across all four metrics (MRR, Hits@1/3/10), outperforming recent robust multimodal models such as SNAG, IMVIA, and C^{2}RS by a clear margin. Similarly, on MKG-Y, RHDN achieves the highest performance across the board, demonstrating robust generalization compared to specialized negative-sampling approaches like DHNS and MMRNS. Overall, these results confirm that the fine-grained cross-modal alignment via R-HAlign effectively captures subtle semantic discrepancies, while the false-negative-aware sampling (FDNS) prevents the model from over-penalizing plausible unobserved facts, enabling RHDN to learn more accurate structural regularities than existing approaches.

Table 1: EXPERIMENTAL RESULTS ON DB15K, MKG-W, AND MKG-Y DATASETS. BEST RESULTS ARE IN **BOLD**, AND SECOND BEST RESULTS ARE UNDERLINED.

Method	DB15K				MKG-W				MKG-Y			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE	24.86	12.78	31.48	47.07	29.19	21.06	33.20	44.23	30.73	23.45	35.18	43.87
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
ComplEx	27.48	18.37	31.57	45.37	24.93	19.09	26.69	36.73	28.71	22.26	32.12	40.93
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
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
MoSE	28.38	21.56	30.91	41.67	33.34	27.78	-	-	36.28	33.64	-	-
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
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
NativE	37.16	28.01	42.33	54.13	36.58	29.56	38.57	47.81	39.04	34.79	40.33	45.91
SNAG	36.30	27.40	41.10	53.00	37.30	30.20	40.50	50.30	39.50	35.40	41.10	47.10
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
MoMoK	39.57	32.38	**43.45**	54.14	35.89	30.38	-	-	37.91	35.09	-	-
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
C^{2}RS	39.65	31.93	43.16	54.57	39.75	32.85	42.37	52.84	39.26	35.05	41.19	46.45
IMVIA	38.29	30.34	42.29	53.55	35.62	30.22	37.70	46.35	37.97	35.15	39.56	43.47
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
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
DHNS	34.36	23.34	41.56	53.72	35.68	28.98	38.68	48.12	39.11	34.70	41.23	46.66
RHDN	**40.24**	**33.16**	42.41	**56.10**	**40.92**	**33.64**	**42.97**	**54.19**	**40.38**	**36.20**	**41.82**	**48.53**

4.3 ABLATION STUDY

We conduct component-level and mechanism-level ablations on DB15K, MKG-W, and MKG-Y. At the component level, **w/o Role Prototypes** removes the hierarchical role prototypes and their regularization; **w/o FDNS** replaces FDNS with standard self-adversarial negative sampling; and **w/o R-HAlign** retains only coarse gated multimodal fusion. For OT, **w/o OT (Aggregation + Loss)** removes both coupling-sensitive aggregation and $\mathcal{L}_{\mathrm{OT}}$; **Independent Coupling** replaces the optimal plan with $\mathbf{T}^{\mathrm{ind}} = \boldsymbol{\pi}\boldsymbol{\rho}^{\top}$ while preserving the same marginals; **w/o OT Loss** removes only $\mathcal{L}_{\mathrm{OT}}$; and **w/o OT Aggregation** removes only coupling-sensitive aggregation. Results are reported in Table 2. Removing any major component consistently degrades performance. The largest drop occurs

Table 2: Ablation study of RHDN on DB15K, MKG-W, and MKG-Y. Best results are shown in bold.

Variant	DB15K				MKG-W				MKG-Y			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
RHDN (Full Model)	**40.24**	**33.16**	**42.41**	**56.10**	**40.92**	**33.64**	**42.97**	**54.19**	**40.38**	**36.20**	**41.82**	**48.53**
Component-level ablations												
w/o Role Prototypes	39.85	32.80	42.10	55.65	40.21	33.05	42.30	53.50	39.95	35.85	41.50	47.90
w/o FDNS	38.56	31.10	40.85	54.10	38.45	31.20	40.55	51.80	38.80	34.50	40.10	46.50
w/o R-HAlign	36.80	29.45	39.10	52.20	36.12	28.95	38.20	49.35	36.50	32.10	38.85	44.20
OT mechanism-level ablations												
w/o OT (Aggregation + Loss)	39.12	31.95	41.50	54.80	39.55	32.40	41.65	52.60	39.10	34.95	40.60	47.15
Independent Coupling	39.57	32.45	41.79	55.24	40.11	32.87	42.13	53.36	39.62	35.29	41.05	47.65
w/o OT Loss	39.83	32.74	42.01	55.51	40.29	33.04	42.27	53.53	39.80	35.46	41.23	47.91
w/o OT Aggregation	39.35	32.18	41.68	55.02	39.78	32.58	41.84	52.91	39.34	35.07	40.82	47.39

without R-HAlign, indicating that coarse entity-level fusion alone is insufficient for relation- and

role-specific multimodal reasoning. Removing FDNS also causes a clear decrease, especially in MRR and Hits@1, while removing role prototypes yields a smaller but consistent drop, showing the complementary benefits of reliable negative supervision and hierarchical role semantics. The OT-specific ablations further clarify its contribution. Replacing the optimal plan with the independent coupling degrades performance despite preserving the same visual and textual marginals, indicating that the gain cannot be explained by marginal weighting alone. Moreover, removing OT-based pairwise aggregation causes a larger drop than removing only $\mathcal{L}_{\mathrm{OT}}$. These results suggest that coupling-sensitive representation learning is the main source of improvement, while the OT consistency loss provides additional regularization.

4.4 Case Study

We further conduct a qualitative case study on DB15K by comparing RHDN with C^{2}RS. Fig. 3 reports the Top-5 predictions for two representative queries and annotates each candidate with its type and semantic role. For the tail prediction query $(\texttt{Ocean's_Eleven}, \texttt{Producer}, ?)$, RHDN ranks the ground-truth producer `Jerry_Weintraub` at #1, and the remaining Top-5 candidates are also person entities with producer-related roles. In contrast, C^{2}RS returns several relation-inconsistent candidates, including actors, an organization, and a film entity, indicating that coarse fusion can be dominated by relation-irrelevant but salient evidence. For the head prediction query $(?, \texttt{Capital}, \texttt{Santiago})$, RHDN ranks the correct answer `Chile` at #1 and produces a type-consistent candidate list consisting of country entities. By contrast, C^{2}RS ranks `Chile` only at #5 and includes mismatched entities such as a mountain range and a city. These results show that RHDN improves both ranking accuracy and candidate quality by enforcing stronger relation- and role-consistency in multimodal reasoning.

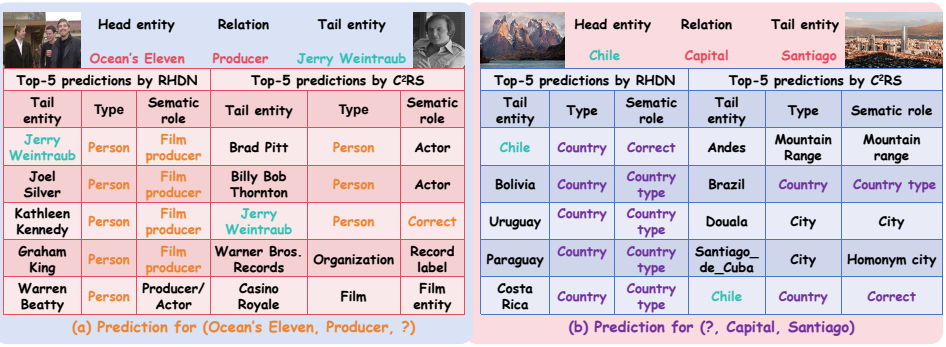


Figure 3: Case study on the DB15K dataset comparing RHDN and C^{2}RS.

5 Conclusions

This paper addresses two key challenges in MMKGC: (i) coarse entity-level fusion that ignores relation- and role-specific fine-grained cross-modal alignment, and (ii) false-negative bias in hard negative sampling on incomplete MMKGs. We proposed **RHDN**, which couples **R-HAlign** and **FDNS**. R-HAlign constructs relation- and role-conditioned optimal transport couplings between visual and textual sub-entities and aggregates coupling-weighted pairwise interactions together with hierarchical role prototypes, producing relation-specific head/tail representations for decoding. FDNS derives a KL-regularized reweighting scheme that emphasizes difficult negatives while suppressing candidates with high false-negative risk. Experiments on DB15K, MKG-W, and MKG-Y show that RHDN consistently outperforms strong unimodal, multimodal, and negative-sampling baselines, while ablations and sensitivity studies verify the complementary benefits of fine-grained alignment and debiased sampling. Overall, RHDN offers a robust and effective framework for MMKGC with controlled computational overhead.

AI USE STATEMENT

Generative AI tools were used solely to assist with manuscript formatting and language polishing. They were not used to generate research ideas, design the methodology, conduct experiments, analyze experimental results, or produce scientific conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and does not involve human-subject experiments or the collection of private or sensitive personal data. The proposed method is intended for research on multi-modal knowledge graph completion. As model predictions may still contain errors, they should not be treated as verified factual statements without appropriate validation. The authors take responsibility for the integrity of the experimental results and the presentation of this work.

REPRODUCIBILITY STATEMENT

We provide the complete formulation of RHDN in the main paper and detailed implementation settings, hyperparameters, evaluation protocols, and additional diagnostic experiments in the appendix. All hyperparameters are selected on the validation set, and the test set is used only for final evaluation. These details are provided to facilitate reproducibility of our experimental results.

REFERENCES

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data: Application to word-sense disambiguation. *Machine learning*, 94(2):233–259, 2014.

Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Jeff Z Pan, Huajun Chen, and Wen Zhang. Noise-powered multi-modal knowledge graph representation framework. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 141–155, 2025.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yao Dong, Lei Wang, Ji Xiang, Xiaobo Guo, and Yuqiang Xie. Rotatect: Knowledge graph embedding by rotation and coordinate transformation in complex space. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4918–4932, 2022.

Chang Gao, Chengjie Sun, Lili Shan, Lei Lin, and Mingjiang Wang. Rotate3d: Representing relations as rotations in three-dimensional space for knowledge graph embedding. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 385–394, 2020.

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

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

Haoning Li and Qinghua Huang. Dans-kgc: Diffusion based adaptive negative sampling for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22913–22921, 2026a.

Haoning Li and Qinghua Huang. Complete fuzzy knowledge representation with knowledge graph embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b.

Qing Li, Zehao Li, Jingjing Song, Jianshuo Bao, Jin Yang, and Zhuhong You. Inkrat: Interpretable diagnosis prediction models based on cross-modal knowledge graph semantic retrieval fusion. *Information Fusion*, pp. 103546, 2025a.

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*, pp. 2572–2580, 2023.

Zhen Li, Jibin Wang, Zhuo Chen, Kun Wu, Meng Ai, Leike An, Liqiang Wang, and Haoxuan Li. Unifying within and across: Intra-modality multi-view fusion and inter-modality alignment for knowledge graph completion. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Ding Wang, and Botian Shi. Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 981–992, 2025.

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

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4710–4723, 2019.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Guanglin Niu and Xiaowei Zhang. Diffusion-based hierarchical negative sampling for multimodal knowledge graph completion. *arXiv preprint arXiv:2501.15393*, 2025.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Bin Shang, Yinliang Zhao, Jun Liu, and Di Wang. Lafa: Multimodal knowledge graph completion with link aware fusion and aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8957–8965, 2024a.

Bin Shang, Yinliang Zhao, Jun Liu, and Di Wang. Lafa: Multimodal knowledge graph completion with link aware fusion and aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8957–8965, 2024b.

Yulou Shu, Wengen Li, Jiaqi Wang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. C 2 rs: Multimodal knowledge graph completion with cross-modal consistency and relation semantics. *IEEE Transactions on Artificial Intelligence*, 2025.

Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1405–1414, 2020.

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

Théo Trouillon, Christopher R Dance, Éric Gaussier, Johannes Welbl, Sebastian Riedel, and Guillaume Bouchard. Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research*, 18(130):1–38, 2017.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

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. IEEE, 2019.

Zihan Wei, Ke Wang, Fengxia Li, and Yina Ma. M3kgr: A momentum contrastive multi-modal knowledge graph learning framework for recommendation. *Information Sciences*, 676:120812, 2024.

Han Xiao, Minlie Huang, Yu Hao, and Xiaoyan Zhu. Transa: An adaptive approach for knowledge graph embedding. *arXiv preprint arXiv:1509.05490*, 2015.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Image-embodied knowledge representation learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 3140–3146. International Joint Conferences on Artificial Intelligence Organization, 2017.

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

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

Yichi Zhang, Mingyang Chen, and Wen Zhang. Modality-aware negative sampling for multimodal knowledge graph embedding. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2023.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 91–101, 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)*, pp. 17120–17130, 2024b.

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

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13322–13330, 2025b.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10527–10536, 2022.

A Related Works

A.1 Unimodal KGC Methods

Unimodal KGC methods focus solely on modeling the structured knowledge within the knowledge graph. Existing unimodal KGC approaches can be broadly categorized into four types: 1)

Translation-based methods, which model relations as translation operations in low-dimensional space (e.g., TransE Wang et al. (2014), TransA Xiao et al. (2015), and TransR Lin et al. (2015)); 2) Rotation-based methods, which represent relations as rotation operations in complex or high-dimensional vector spaces (e.g., RotatE Sun et al. (2019), RotatE3D Gao et al. (2020), and RotateCT Dong et al. (2022)); 3) Semantic matching methods, which evaluate triples based on latent semantic similarity (e.g., RESCAL Bordes et al. (2014), DistMult Yang et al. (2015), and HolE Nickel et al. (2016)); and 4) Neural network–based methods, which leverage neural architectures to capture the graph topology of KGs (e.g., ConvE Dettmers et al. (2018), RGCN Schlichtkrull et al. (2018), and KBGAT Nathani et al. (2019)). Although significant progress has been made in these approaches, they face inherent limitations in modeling multimodal information associated with KGs, which restricts their applicability in real-world scenarios involving rich multimodal data.

B IMPLEMENTATION DETAILS

Software and hardware. We implement RHDN using PyTorch 2.3.1 with Python 3.10 and CUDA 12.1. All experiments are conducted on NVIDIA RTX A6000 GPUs with 48 GB memory. Each individual training run uses a single GPU, while independent runs and ablation experiments are executed in parallel across four GPUs. Mixed-precision training is enabled to reduce memory consumption. Unless otherwise specified, all experiments use three random seeds $\{42, 3407, 2027\}$.

Multi-modal feature extraction. We use a two-layer RGAT as the structural encoder, with four attention heads and hidden dimension $d = 200$. For visual information, we adopt a pretrained ViT-B/16 encoder. Each entity image is resized to 224×224 and encoded into patch-level representations. To obtain a fixed and computationally manageable number of visual sub-entities, the patch feature map is adaptively pooled into a 4×4 grid, resulting in at most $N_v = 16$ visual region embeddings per entity.

For textual information, we use BERT-base-uncased as the textual encoder. Entity descriptions are tokenized with a maximum sequence length of 64, and at most $N_x = 32$ non-special token representations are retained as textual sub-entities. Longer descriptions are truncated, whereas shorter sequences are padded and masked accordingly. The pretrained ViT and BERT encoders are kept frozen during RHDN training. Visual, textual, and structural representations are projected into the common $d = 200$ dimensional space through learnable linear projections followed by layer normalization.

The global visual and textual representations used in coarse multimodal fusion are obtained by mean pooling over the corresponding valid region and token embeddings. The resulting structural, visual, and textual representations are then combined by the gated multimodal fusion module described in Sec. 3.3.

Handling missing modalities. Not every entity in the benchmark datasets is associated with complete visual and textual information. We therefore perform modality-aware computation rather than imputing unavailable modalities with artificial observations. When the visual or textual modality is missing, it is excluded from the gated fusion normalization and the remaining modality weights are renormalized. If only structural information is available, the coarse representation reduces to the structural representation.

For R-HAlign, cross-modal OT coupling is performed only when both visual and textual sub-entities are available. If either modality is absent, the fine-grained OT interaction term for that entity is skipped and the coarse entity representation is used through the residual projection $\mathbf{W}_b\mathbf{b}_e$. The corresponding OT consistency loss is set to zero for that entity side. Similarly, the FDNS evidence set $\mathcal{M}_i$ contains only the modalities available for candidate entity e_i, so that missing modalities do not artificially decrease its evidence-consensus score.

Relation clustering and role prototypes. The hierarchical role module uses both relation-level and cluster-level head/tail prototypes. To obtain relation clusters without using validation or test labels, we first warm up the structural link-prediction model for 20 epochs using training triples only. We then apply k-means clustering to the resulting relation embeddings. We use 20 random

initializations for k-means and retain the solution with the lowest within-cluster sum of squared distances. The cluster assignments are fixed during subsequent joint RHDN training.

The number of clusters $|\mathcal{C}|$ is selected on the validation set and is set to 16, 16, and 8 for DB15K, MKG-W, and MKG-Y, respectively, as reported in Table 4. Relation-level head and tail prototypes are initialized from the corresponding relation embeddings with small Gaussian perturbations $\mathcal{N}(0, 0.01^2)$. Each cluster-level prototype is initialized as the mean of the relation-level prototypes assigned to that cluster. All relation- and cluster-level prototypes are subsequently optimized jointly with the rest of RHDN.

OT-based pairwise interaction. For each head or tail entity, R-HAlign constructs relation- and role-aware visual and textual marginal distributions and computes an entropy-regularized OT coupling using Sinkhorn iterations. The Sinkhorn regularization coefficient is $\varepsilon_{\mathrm{OT}} = 0.05$, and we use 10 Sinkhorn iterations for all three datasets.

The pairwise interaction encoder ϕ_{pair} in Eq. equation 11 is implemented as a two-layer MLP:

$$5d \to d \to d, \tag{24}$$

because its input concatenates $\mathbf{v}_i$, $\mathbf{x}_j$, $\mathbf{v}_i \odot \mathbf{x}_j$, $|\mathbf{v}_i - \mathbf{x}_j|$, and the role-conditioned context $\mathbf{c}_r$. A GELU activation and the dataset-specific dropout rate in Table 4 are applied between the two linear layers. The same pairwise encoder is shared between head-side and tail-side interaction aggregation.

Negative candidate construction. For each positive triple, we sample $N = 64$ candidate negatives. Head and tail corruption are performed with equal probability. The corrupted entity is sampled uniformly from the entity vocabulary, and duplicates as well as triples already observed in the training graph are removed from the negative candidate set. Validation and test triples are not consulted when constructing training negatives, thereby avoiding information leakage from held-out splits.

All remaining candidates are first scored by the current RHDN model. Their model scores determine the difficulty distribution p_i, whereas their relation- and role-conditioned multimodal evidence determines the false-negative risk r_i. The final FDNS weights q_i are computed according to Eq. equation 21. We detach q_i from the computation graph before applying the weighted negative logistic loss, as indicated by the stop-gradient operator in Eq. equation 22.

Optimization and model selection. All trainable parameters are optimized using Adam. The dataset-specific batch sizes, learning rates, dropout rates, and loss coefficients are listed in Table 4. We additionally use a weight decay of 10^{-5} and clip the global gradient norm to 1.0. Models are trained for at most 500 epochs and evaluated on the validation set every five epochs. Training is terminated early if validation MRR does not improve for 30 consecutive validation evaluations. The checkpoint obtaining the highest validation MRR is used for final evaluation on the test set.

The pretrained ViT and BERT parameters remain frozen throughout training, whereas RGAT, modality projections, gated fusion, role-conditioned attention, relation/cluster prototypes, pairwise interaction layers, and the ConvE-based decoder are jointly optimized.

Hyperparameter selection. All model-specific hyperparameters are selected exclusively according to validation-set MRR. We adopt a two-stage search strategy: an initial coarse search identifies a promising region, followed by local refinement around the best coarse configuration for sensitive parameters. The final dataset-specific settings are reported in Table 4. Importantly, the values listed in Table 4 correspond to the final refined configurations and therefore need not coincide exactly with the discrete points of the initial coarse grid. The complete search protocol is described in Appendix C. No hyperparameter is selected using test-set performance.

Filtered evaluation. We evaluate both head prediction $(?, r, t)$ and tail prediction $(h, r, ?)$ under the standard filtered link-prediction protocol. For each query, all entities are ranked according to their plausibility scores. Before computing the rank of the target entity, all other candidates that correspond to known true triples in the training, validation, or test split are removed, while the target triple itself is retained. MRR and Hits@K ($K \in \{1, 3, 10\}$) are then computed over both head and tail queries. The same evaluator is used for RHDN and all reproduced baselines.

C Experimental Setup and Hyperparameter Settings

C.1 Experimental Setup

We use the official train/validation/test splits of DB15K, MKG-W, and MKG-Y, whose statistics are summarized in Table 3. Model selection and hyperparameter tuning are performed exclusively on the validation set, while the test set is reserved for final evaluation. For each training run, we select the checkpoint with the highest validation MRR and use this checkpoint for test evaluation.

Table 3: Statistics of three MMKGC benchmark datasets.

Dataset	Entity	Relation	Image	Text	Train	Valid	Test
DB15K	12842	279	12818	9078	79222	9902	9904
MKG-W	15000	169	14463	14123	34196	4276	4274
MKG-Y	15000	28	14244	12305	21310	2665	2663

C.2 Hyperparameter Settings

We determine the hyperparameters using a two-stage validation-based search. In the first stage, we perform a coarse search over commonly used values to identify promising regions. For FDNS, the initial search considers

$$\begin{aligned} \tau &\in \{0.3, 0.5, 0.6, 0.65, 0.8\}, \\ \gamma_{\mathrm{fn}} &\in \{0.50, 0.55, 0.60, 0.65, 0.70\}, \\ T_{\mathrm{fn}} &\in \{0.05, 0.10, 0.15, 0.20\}, \\ \lambda_{\mathrm{fn}} &\in \{0.1, 0.3, 0.5, 0.7\}, \\ \beta &\in \{0, 0.2, 0.4, 0.6, 0.8\}. \end{aligned} \tag{25}$$

For R-HAlign, we search

$$\begin{aligned} \lambda_{\mathrm{OT}} &\in \{0.05, 0.10, 0.15, 0.20, 0.30\}, \\ \lambda_{\mathrm{proto}} &\in \{0.05, 0.10, 0.20, 0.30\}, \\ \lambda_{\mathrm{cluster}} &\in \{0.01, 0.05, 0.10, 0.20\}, \\ |\mathcal{C}| &\in \{4, 8, 16, 32\}. \end{aligned} \tag{26}$$

In the second stage, parameters showing clear sensitivity are locally refined around the best coarse-grid configuration using smaller intervals. In particular, this refinement is applied to γ_{fn}, T_{fn}, and β, allowing the final validation-selected values to lie between the initial coarse-grid points. This explains, for example, the refined values $\gamma_{\mathrm{fn}} = 0.62$ and 0.58, $T_{\mathrm{fn}} = 0.12$, and $\beta = 0.50$ in Table 4. All remaining optimization parameters, including the learning rate, batch size, and dropout rate, are likewise selected according to validation performance. The test set is never used for hyperparameter selection.

The final dataset-specific configurations are summarized in Table 4. These settings are used for the main experiments and serve as the reference configurations for the post-hoc sensitivity analysis.

D Other studies

D.1 Hyperparameter Sensitivity Analysis

To examine the robustness of RHDN to its key hyperparameters, we conduct one-at-a-time sensitivity analyses for the difficulty temperature τ, false-negative penalty λ_{fn}, OT consistency weight λ_{OT}, and relation-cluster number $|\mathcal{C}|$. All sensitivity experiments are conducted on the *validation set*; the test-set results reported in the main paper are not used in this analysis.

Table 4: Final hyperparameter settings on the three MMKGC datasets, selected according to validation performance.

Hyperparameters	**DB15K**	**MKG-W**	**MKG-Y**
Embedding dimension d	200	200	200
Batch size	1024	512	512
Optimizer	Adam	Adam	Adam
Learning rate	5×10^{-4}	1×10^{-4}	1×10^{-4}
Dropout rate	0.3	0.4	0.3
FDNS-related hyperparameters			
Difficulty temperature τ	0.50	0.65	0.60
FN-risk threshold γ_{fn}	0.62	0.58	0.60
FN-risk temperature T_{fn}	0.10	0.12	0.10
False-negative penalty λ_{fn}	0.30	0.50	0.50
Evidence disagreement weight β	0.40	0.50	0.40
R-HAlign-related hyperparameters			
OT consistency weight λ_{OT}	0.10	0.15	0.20
Prototype alignment weight λ_{proto}	0.10	0.10	0.20
Cluster regularization weight $\lambda_{\mathrm{cluster}}$	0.05	0.10	0.20
Number of relation clusters $\lvert\mathcal{C}\rvert$	16	16	8
Sinkhorn regularization $\varepsilon_{\mathrm{OT}}$	0.05	0.05	0.05
Sinkhorn iterations	10	10	10

For each sweep, only the parameter under study is varied, while all remaining parameters are fixed to the corresponding dataset-specific settings in Table 4. Thus, the selected default point in different sweeps corresponds to the same underlying model configuration for a given dataset. The sweep ranges are chosen around the selected operating regions for post-hoc robustness analysis and therefore need not exactly coincide with the initial coarse search grids used during hyperparameter selection.

FDNS hardness temperature τ. The temperature τ controls the concentration of the difficulty distribution over candidate negatives. Very small values place excessive weight on a few high-scoring candidates, whereas overly large values weaken the hard-negative effect. The validation curves show that intermediate dataset-specific values provide a more favorable balance between hardness and training stability.

False-negative penalty λ_{fn}. The coefficient λ_{fn} controls the strength of risk-aware negative suppression. A small value provides limited protection against high-risk candidates, whereas an excessively large value may also suppress informative hard negatives. The selected settings therefore reflect a balance between preserving difficult supervision and reducing the influence of likely false negatives.

OT consistency weight λ_{OT}. The coefficient λ_{OT} balances cross-modal consistency regularization against the link-prediction objective. Insufficient OT regularization provides limited alignment guidance, while an overly strong constraint can over-emphasize cross-modal consistency at the expense of the primary prediction objective. The validation results support moderate dataset-specific settings.

Relation clustering granularity $|\mathcal{C}|$. The number of relation clusters controls the granularity of semantic sharing among hierarchical role prototypes. Too few clusters may force semantically different relations to share overly coarse prototypes, whereas too many clusters weaken the benefit of higher-level semantic sharing. The selected values in Table 4 provide a suitable trade-off across the three datasets.

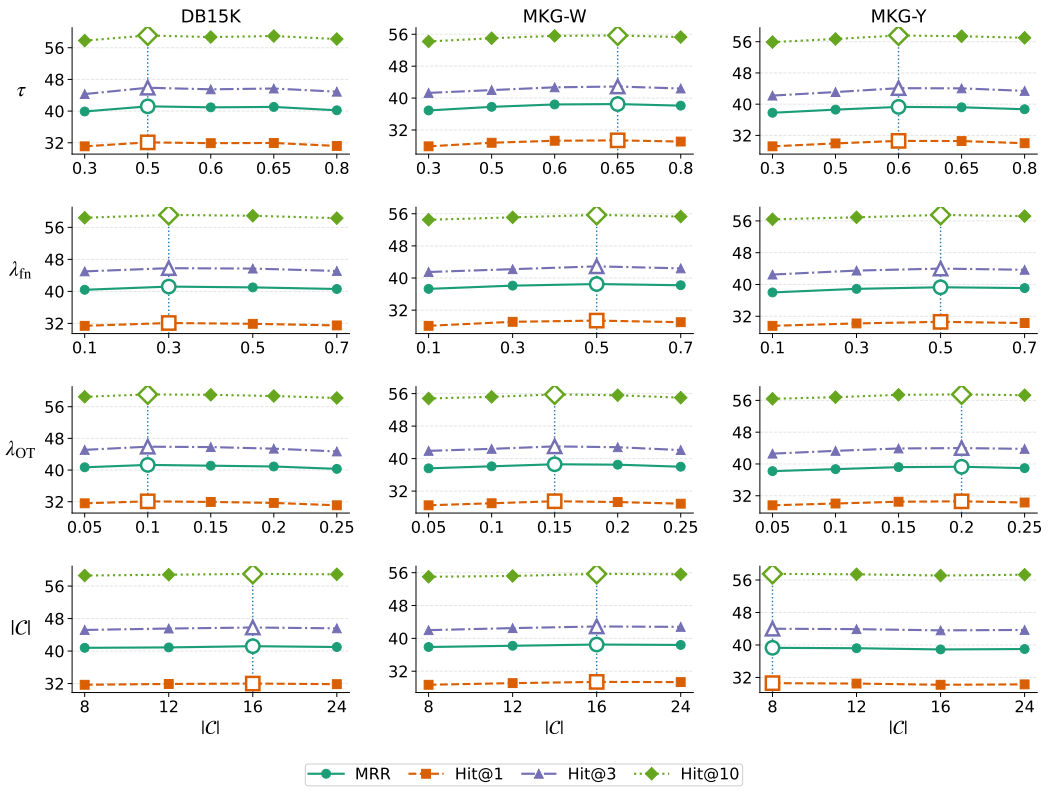


Figure 4: Validation-set sensitivity analysis of RHDN on DB15K, MKG-W, and MKG-Y. Each sweep varies one hyperparameter while fixing all others to the dataset-specific settings in Table 4. The vertical dashed lines indicate the selected default values.

D.2 DIRECT VALIDATION AND MECHANISM ANALYSIS OF FDNS

The main ablation study verifies the overall effectiveness of FDNS, but does not by itself determine whether the improvement arises from selectively suppressing likely false negatives or simply from reducing the weights of high-scoring negatives. We therefore conduct controlled diagnostic experiments on DB15K to explicitly examine the proposed difficulty–risk decoupling mechanism.

Controlled pseudo-false-negative protocol. We randomly hide 5% of the DB15K training triples from the observed graph before training and use them only for diagnostic evaluation. During negative corruption, a candidate that matches one of these hidden triples is regarded as a *pseudo false negative* (pseudo-FN). The hidden triples are never exposed to FDNS as false-negative labels during training.

For mechanism-level diagnostics, we use a fixed trained checkpoint and freeze both the candidate pools and their model scores. All weighting strategies are then evaluated on exactly the same candidates, so that differences in the reported statistics are caused by the weighting mechanism rather than by different trained models or different negative pools. Among the remaining candidates, high-scoring negatives that do not match either observed or hidden positives are treated as genuine hard negatives for this controlled analysis, while low-scoring candidates are treated as easy negatives. We note that, due to KG incompleteness, these non-pseudo negatives cannot be guaranteed to be true negatives; the terminology is used only to distinguish them from the explicitly held-out pseudo-FNs.

Selective suppression of pseudo-FNs. We first compare the average normalized weights assigned to different types of negative candidates. In addition to standard self-adversarial sampling, we include a **Score-only Risk** baseline, which uses the model score for both difficulty estimation and

Table 5: Average normalized negative weights under the controlled pseudo-FN protocol on DB15K. Score-only Risk estimates both difficulty and risk from the model score, whereas FDNS estimates false-negative risk from multimodal role-consistency evidence.

Method	**Pseudo-FNs** ↓	**Hard negatives**	**Easy negatives**	**Hard/FN ratio** ↑
Self-adversarial	0.214	0.221	0.081	1.03
Score-only Risk	0.198	0.204	0.080	1.03
FDNS	**0.170**	**0.219**	0.080	**1.29**

Table 6: Pseudo-FN ratio among highly weighted negative candidates on DB15K. Lower values indicate that fewer pseudo-FNs receive strong negative supervision.

Method	Top-10% ↓	Top-20% ↓
Self-adversarial	27.6%	21.3%
Score-only Risk	26.9%	20.8%
FDNS	**13.7%**	**11.6%**

false-negative risk:

$$r_i^{\text{score}} = \sigma\left(\frac{s_i - \gamma_s}{T_s}\right). \tag{27}$$

This baseline isolates whether an independent multimodal risk signal is necessary beyond nonlinear transformation of the model score.

As shown in Table 5, self-adversarial sampling assigns similar weights to pseudo-FNs and hard negatives. Score-only Risk provides only limited separation because both difficulty and risk are derived from the same model score. In contrast, FDNS selectively reduces the average weight of pseudo-FNs while largely preserving the weight assigned to hard negatives, increasing the hard-to-FN weight ratio from 1.03 to 1.29. This behavior is consistent with the intended separation between negative difficulty and false-negative risk.

Pseudo-FNs among highly weighted negatives. We further examine the negative candidates receiving the strongest training penalties. Specifically, we rank candidates within each frozen candidate pool according to their normalized negative weights and compute the pseudo-FN ratio among the top-10% and top-20% candidates.

Table 6 shows that score-only risk changes the composition of highly weighted negatives only marginally relative to self-adversarial sampling. By contrast, FDNS reduces the pseudo-FN proportion from 27.6% to 13.7% among the top-10% weighted negatives and from 21.3% to 11.6% among the top-20%. Since all methods operate on the same frozen scores and candidate pools, this difference directly reflects the effect of the independent multimodal risk signal on negative reweighting.

Exact-score-matched analysis. The previous analyses may still be influenced by differences in model difficulty between pseudo-FNs and hard negatives. We therefore construct a stricter score-controlled subset. For every pseudo-FN, we select a hard negative from the *same candidate pool* whose model score is closest to that of the pseudo-FN and satisfies

$$|s_{\text{FN}} - s_{\text{hard}}| \le \epsilon, \qquad \epsilon = 0.005. \tag{28}$$

Each hard negative is matched at most once. This approximately controls for model difficulty and allows us to test whether the false-negative risk estimator provides information beyond the prediction score itself.

Table 7 further controls for model difficulty by matching pseudo-FNs and hard negatives with nearly identical scores. Under this setting, Score-only Risk assigns almost identical risks to the two groups, yielding an AUROC close to random chance and a weight ratio near one. In contrast, FDNS assigns higher risk to pseudo-FNs than to score-matched hard negatives and achieves substantially better risk discrimination. Consequently, pseudo-FNs receive approximately 11% lower negative weights

Table 7: Exact-score-matched analysis on DB15K. Pseudo-FNs and hard negatives are matched within the same candidate pool with $|s_{\mathrm{FN}} - s_{\mathrm{hard}}| \leq 0.005$. Risk AUROC measures the ability of the risk signal to distinguish pseudo-FNs from score-matched hard negatives.

Method	$\lvert\Delta\mathbf{s}\rvert$	$\mathbf{r}_{\mathrm{FN}}$	$\mathbf{r}_{\mathrm{hard}}$	**Risk AUROC** ↑	$\mathbf{q}_{\mathrm{FN}}/\mathbf{q}_{\mathrm{hard}}$ ↓
Score-only Risk	0.0021	0.742	0.738	50.9	0.998
FDNS	0.0021	**0.781**	**0.394**	**79.6**	**0.890**

Table 8: Mechanism-level ablation of FDNS on DB15K.

Variant	MRR	Hits@1	Hits@3	Hits@10
w/o FN Risk	38.56	31.10	40.85	54.10
w/o Difficulty	39.18	31.94	41.37	54.88
Score-only Risk	39.31	32.06	41.52	55.01
w/o Evidence Disagreement	39.76	32.67	42.02	55.48
w/o Role Conditioning	39.52	32.39	41.76	55.19
FDNS (Full)	**40.24**	**33.16**	**42.41**	**56.10**

than equally difficult hard negatives. This result directly supports that FDNS introduces a risk signal beyond the model score itself.

Mechanism-level ablation of FDNS. Finally, we decompose the revised FDNS mechanism on DB15K. Unlike the preceding diagnostic experiments, which use a frozen checkpoint, the following variants are trained independently under the same training and evaluation protocol.

w/o FN Risk removes the risk penalty and reduces FDNS to standard difficulty-based self-adversarial weighting. **w/o Difficulty** replaces the difficulty distribution with a uniform distribution while retaining multimodal risk suppression. **Score-only Risk** replaces the multimodal risk estimator with Eq. equation 27. **w/o Evidence Disagreement** sets $\beta = 0$ in Eq. equation 18, such that false-negative risk depends only on the mean multimodal role support. Finally, **w/o Role Conditioning** removes the head/tail role distinction when estimating multimodal support by replacing $\mathbf{u}_r^{(h)}$ and $\mathbf{u}_r^{(t)}$ with their shared prototype

$$\bar{\mathbf{u}}_r = \frac{1}{2}\left(\mathbf{u}_r^{(h)} + \mathbf{u}_r^{(t)}\right). \tag{29}$$

As shown in Table 8, removing either difficulty modeling or false-negative risk estimation consistently degrades performance, indicating that the two factors play complementary roles. Using score-only risk also performs clearly worse than the full model, supporting the need to estimate false-negative risk from a signal independent of model difficulty. In addition, removing either inter-modality disagreement modeling or head/tail role conditioning causes consistent degradation, suggesting that both multimodal evidence consensus and role-specific semantics contribute to more reliable risk estimation.

Overall, the controlled diagnostics and mechanism-level ablations consistently support the intended behavior of FDNS. Model scores characterize how difficult negative candidates are, while relation- and role-conditioned multimodal evidence provides an independent estimate of false-negative risk. Their combination through KL-regularized reweighting allows FDNS to preserve informative hard negatives while reducing the influence of plausible unobserved facts.

D.3 Behavioral Analysis of R-HAlign

To examine whether R-HAlign exhibits relation- and role-aware fine-grained alignment, we conduct complementary case-level and dataset-level diagnostics. All case-level results reported below are extracted from the same trained checkpoint under deterministic evaluation, using the same entity inputs, sub-entity extraction procedure, and preprocessing pipeline. Accordingly, repeated entity–relation–role configurations share exactly the same attention and OT statistics across tables.

Table 9: Relation-aware alignment for the same entity *A Beautiful Mind* under different query relations.

Query relation	Top attended textual phrases	Top attended visual regions	Strongest cross-modal OT pairs
wonAward	"Academy Award" (0.38); "Best Picture" (0.35); "directed by Ron Howard" (0.10)	award/trophy cue (0.42); laurel/banner area (0.34); title/poster area (0.10)	trophy cue → "Academy Award" (0.193); trophy cue → "Best Picture" (0.151); laurel/banner area → "Best Picture" (0.136)
starring	"starring Russell Crowe" (0.31); "Jennifer Connelly" (0.27); "Academy Award / Best Picture" (0.22)	Russell Crowe face (0.33); Jennifer Connelly face (0.27); background/banner area (0.16)	Russell Crowe face → "starring Russell Crowe" (0.162); Jennifer Connelly face → "Jennifer Connelly" (0.130); award/trophy cue → "Academy Award / Best Picture" (0.073)

Table 10: Qualitative head–tail alignment patterns under the same relation *starring*. The head-side statistics are identical to the corresponding *starring* configuration in Table 9.

Role / Entity	Top attended textual phrases	Top attended visual regions	Strongest cross-modal OT pairs
Head / *A Beautiful Mind*	"starring Russell Crowe" (0.31); "Jennifer Connelly" (0.27); "Academy Award / Best Picture" (0.22)	Russell Crowe face (0.33); Jennifer Connelly face (0.27); background/banner area (0.16)	Russell Crowe face → "starring Russell Crowe" (0.162); Jennifer Connelly face → "Jennifer Connelly" (0.130); award/trophy cue → "Academy Award / Best Picture" (0.073)
Tail / *Russell Crowe*	"Russell Crowe" (0.37); "actor" (0.25); "starred in A Beautiful Mind" (0.21)	face close-up (0.41); upper-body region (0.24); caption/name area (0.17)	face close-up → "Russell Crowe" (0.189); upper-body region → "actor" (0.077); caption/name area → "Russell Crowe" (0.065)

D.3.1 Relation-Aware Alignment

We first analyze the same entity, *A Beautiful Mind*, under two query relations, *wonAward* and *starring*. If R-HAlign is relation-aware, the evidence emphasized for the same entity should change according to the queried relation.

As shown in Table 9, the alignment pattern changes substantially across the two relations. For *wonAward*, the model emphasizes award-related phrases and trophy/banner regions, whereas for *starring*, attention shifts toward actor names and face regions. This qualitative example illustrates that R-HAlign can adapt fine-grained multimodal evidence to the current relation rather than relying on a fixed entity-level representation.

D.3.2 Role-Aware Alignment

We next fix the relation as *starring* and inspect its two semantic roles using the head entity *A Beautiful Mind* and the tail entity *Russell Crowe*. For consistency, the head-side statistics in Table 10 are exactly the same as the *starring* configuration in Table 9.

As shown in Table 10, the movie-side representation focuses on actor-related evidence associated with the film, whereas the person-side representation emphasizes the identity and actor-related cues of *Russell Crowe*. This comparison provides an intuitive illustration of the asymmetric head/tail behavior induced by role conditioning. However, because the two sides correspond to different entities, this example alone is not intended as a controlled proof that the difference is caused solely by the role mechanism. We therefore quantify role-dependent behavior with controlled dataset-level diagnostics in Sec. D.3.3.

D.3.3 Dataset-Level Validation

We further evaluate relation- and role-aware behavior on the DB15K validation set using controlled dataset-level diagnostics. Relation Selectivity measures the average Jensen–Shannon divergence between attention distributions of the same entity under different relations, while Role Selectivity measures the change in attention induced by different head/tail role conditions. Prototype Purity@5 measures the percentage of the top-5 nearest neighbors of each relation-role prototype that match its expected semantic type or role category.

These dataset-level metrics provide a more controlled assessment of relation- and role-dependent behavior than the individual examples above. In particular, Role Selectivity is used as the primary

quantitative evidence for role-dependent modulation, whereas Table 10 serves only as a qualitative illustration.

D.4 Complexity Analysis

We analyze the computational complexity of RHDN during training. Let $|\mathcal{T}|$ denote the number of training triples, B the batch size, d the embedding dimension, N the number of candidate negatives per positive triple, N_v and N_x the maximum numbers of retained visual regions and textual tokens, respectively, and I the number of Sinkhorn iterations. Since the pretrained visual and textual features are extracted before RHDN training, we focus on the additional cost introduced by R-HAlign, the decoder, and FDNS. For clarity, we give a conservative upper bound without assuming cross-candidate computation sharing.

Role-aware sub-entity weighting. For each scored triple, R-HAlign first constructs role-conditioned contexts and computes attention over visual and textual sub-entities. The context MLP and feature projections involve d-dimensional transformations, giving

$$C_{\text{role}} = O\left((N_v + N_x)d^2 + d^2\right). \tag{30}$$

The lower-order attention normalization cost $O((N_v + N_x)d)$ is absorbed by the above term.

Optimal transport and pairwise interaction. Constructing the visual–textual cost matrix requires

$$O(N_v N_x d), \tag{31}$$

while I Sinkhorn iterations require

$$O(I N_v N_x). \tag{32}$$

Importantly, the revised R-HAlign applies the pair encoder $\phi_{\text{pair}} : 5d \to d \to d$ to every visual–textual pair. A two-layer MLP of this form costs $O(d^2)$ per pair, yielding

$$C_{\text{pair}} = O(N_v N_x d^2). \tag{33}$$

Transport-weighted aggregation additionally requires $O(N_v N_x d)$, which is dominated by the pair-encoding term. The subsequent hierarchical role aggregation involves d-dimensional linear projections and therefore costs $O(d^2)$.

Combining these operations, the complexity of R-HAlign for one scored triple is

$$\begin{aligned} C_{\text{align}} = O\Big((N_v + N_x)d^2 + N_v N_x d^2 \\ + N_v N_x (d + I) + d^2\Big). \end{aligned} \tag{34}$$

Equivalently, the dominant additional term introduced by fine-grained pairwise interaction is $O(N_v N_x d^2)$. The head and tail computations differ only by constant factors, which are omitted in the asymptotic notation.

ConvE-style decoding. Given the relation-aware entity representations, the ConvE-style decoder applies convolution followed by projection and candidate matching. Let n_f denote the number of convolutional filters and k the kernel size. The cost for one scored triple is

$$C_{\text{dec}} = O(n_f k^2 d + d^2), \tag{35}$$

where the d^2 term accounts for the projection layer.

FDNS reweighting. For each positive triple, FDNS considers N candidate negatives. Their model scores are already included in the triple-scoring cost above. For false-negative risk estimation, each

candidate is compared with the corresponding role prototype using at most three available modality representations. Since each cosine similarity costs $O(d)$, risk estimation requires $O(Nd)$ in total. Computing the difficulty distribution, evidence statistics, sigmoid transformation, and normalized reweighting is linear in N. Hence,

$$C_{\mathrm{FDNS}} = O(Nd). \tag{36}$$

Overall complexity. Each positive triple requires scoring one positive and N candidate negatives. A conservative per-positive-triple upper bound is therefore

$$C_{\mathrm{triple}} = (N+1)\,(C_{\mathrm{align}} + C_{\mathrm{dec}}) + O(Nd). \tag{37}$$

Substituting the above terms gives

$$\begin{aligned} C_{\mathrm{triple}} = O\Big((N+1)\big[&(N_v+N_x)d^2 + N_vN_xd^2 \\ &+ N_vN_x(d+I) + n_fk^2d + d^2\big] + Nd\Big). \end{aligned} \tag{38}$$

Accordingly, the complexity of one training epoch is

$$\begin{aligned} C_{\mathrm{epoch}} = O\Big(|\mathcal{T}|\Big\{(N+1)\big[&(N_v+N_x)d^2 + N_vN_xd^2 \\ &+ N_vN_x(d+I) + n_fk^2d + d^2\big] + Nd\Big\}\Big). \end{aligned} \tag{39}$$

The corresponding per-batch complexity is $O(B\,C_{\mathrm{triple}})$. In practice, N_v, N_x, N, and I are fixed predefined budgets, so the training cost remains linear in the number of training triples $|\mathcal{T}|$. Compared with coarse entity-level fusion, the main additional cost of RHDN arises from the pairwise interaction encoder, $O(N_vN_xd^2)$, while FDNS introduces only an additional $O(Nd)$ reweighting cost after candidate scoring.

D.5 Visualization Study

D.5.1 Visualizing OT Coupling and Key Token–Region Alignments

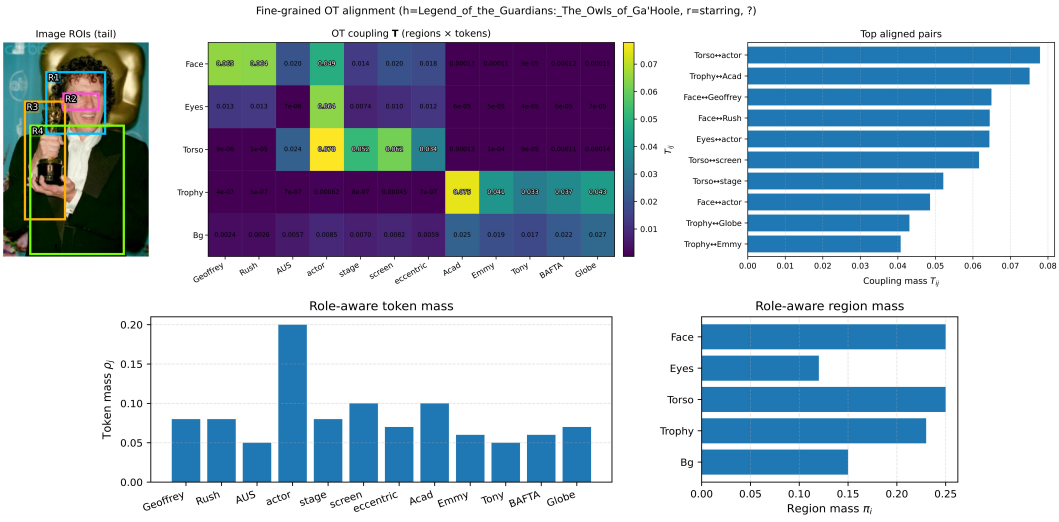


Figure 5: Visualization .

To improve the interpretability and diagnosability of multimodal knowledge graph completion (MMKGC), we visualize the *optimal transport* (OT) coupling between textual evidence and visual evidence, together with the induced *key token–region alignment pairs*. In MMKGC, even when a model predicts the correct entity, it remains unclear *why* the prediction is made and whether the model truly leverages relation-consistent cross-modal evidence. This issue is particularly salient under noisy or weakly informative modalities (e.g., irrelevant image backgrounds or off-topic textual snippets). Therefore, our visualization aims to answer two questions: (i) whether the model aligns *relation-relevant textual cues* to *semantically plausible visual regions*; and (ii) whether the model can distinguish *critical evidence* from *non-critical/noisy evidence*, thereby substantiating the effectiveness of our fine-grained alignment and noise-robust evidence allocation.

Construction. Given a query $(h, r, ?)$ (we use $r = \texttt{starring}$ as an illustrative example), we extract a set of relation-aware textual tokens $\{x_j\}_{j=1}^{K}$ from the candidate tail entity's description (e.g., *Geoffrey*, *Rush*, *actor*, *stage*, *screen*, *AcademyAward*), and select a set of informative image regions $\{p_i\}_{i=1}^{M}$ (e.g., *Face*, *Eyes*, *Torso*, *Trophy*, *Background*). The model produces normalized mass distributions on both sides, i.e., token mass $\boldsymbol{\rho} \in \Delta^K$ and region mass $\boldsymbol{\pi} \in \Delta^M$. We then compute an OT coupling matrix $\mathbf{T} \in \mathbb{R}_+^{M\times K}$ satisfying the marginal constraints $\sum_j T_{ij} = \pi_i$ and $\sum_i T_{ij} = \rho_j$. Each entry T_{ij} can be interpreted as the strength of transporting evidence mass from region p_i to token x_j, thus directly characterizing the fine-grained cross-modal alignment pattern. For readability, we further extract the Top-L largest entries in $\mathbf{T}$ as *key alignment pairs* and visualize them with a bar chart as Figure 5.

Observations. The OT coupling and the Top-L alignment pairs reveal consistent, relation-aligned cross-modal evidence flow.

(1) Identity cues align to facial regions. Tokens carrying identity information (e.g., *Geoffrey*, *Rush*) receive higher coupling mass from *Face* (and partially *Eyes*), indicating that the model grounds entity identity primarily in portrait-related visual evidence rather than in awards or background textures. This behavior matches the basic requirement of reasoning about `starring`, where recognizing the actor/actress is essential.

(2) Relation-relevant context concentrates on torso/holistic regions. Tokens more directly related to $r = \texttt{starring}$ (e.g., *actor*, *stage*, *screen*, *eccentric*) exhibit stronger couplings with the *Torso* region. This suggests that the model leverages holistic human appearance/contextual cues and aligns them with profession/performance semantics in text, which is crucial for relation-specific inference. Such "context tokens $\rightarrow$ torso/holistic region" patterns also dominate the Top-L pairs, supporting the contribution of fine-grained alignment to relation reasoning.

(3) Awards tokens match trophy regions without hijacking the main reasoning path. Award-related tokens (e.g., *AcademyAward*, *Emmy*, *Tony*, *BAFTA*, *GoldenGlobe*) mainly couple to the *Trophy* region, demonstrating that the model can map award semantics to plausible visual evidence when present. Importantly, due to our relation-aware evidence allocation (reflected by $\boldsymbol{\rho}$, $\boldsymbol{\pi}$ and the OT constraints), these tokens do not improperly dominate the evidence pathway for `starring`, preventing visually salient but relation-irrelevant cues from overwhelming the inference.

(4) Background is weak and diffuse, indicating noise suppression. The *Background* region distributes only small and more diffuse coupling mass across tokens, implying that the model does not rely on background textures as primary evidence. This provides empirical support that our alignment mechanism is robust to irrelevant visual content and reduces the impact of noisy multimodal signals.

Takeaway. Overall, the visualization offers direct evidence that our model performs fine-grained, relation-consistent cross-modal grounding: relation-relevant textual cues (e.g., *actor*/*stage*/*screen*) are aligned to semantically appropriate visual regions, auxiliary award information is confined to the corresponding visual evidence, and background noise is suppressed. This interpretable evidence flow helps explain the stable performance gains observed in our MMKGC experiments.

D.5.2 Difficulty–Risk–Weight (DRW) Mechanism Visualization

To make the effect of our *difficulty–risk–weight* (DRW) reweighting intuitive and verifiable, we visualize how candidate negatives are scored and weighted on a **real** DB15K test triple, $\langle h, r, t\rangle$=¡*Ocean's Eleven*, *producer*, *Jerry Weintraub*¿. For each negative tail candidate $\tilde{t}$, we esti-

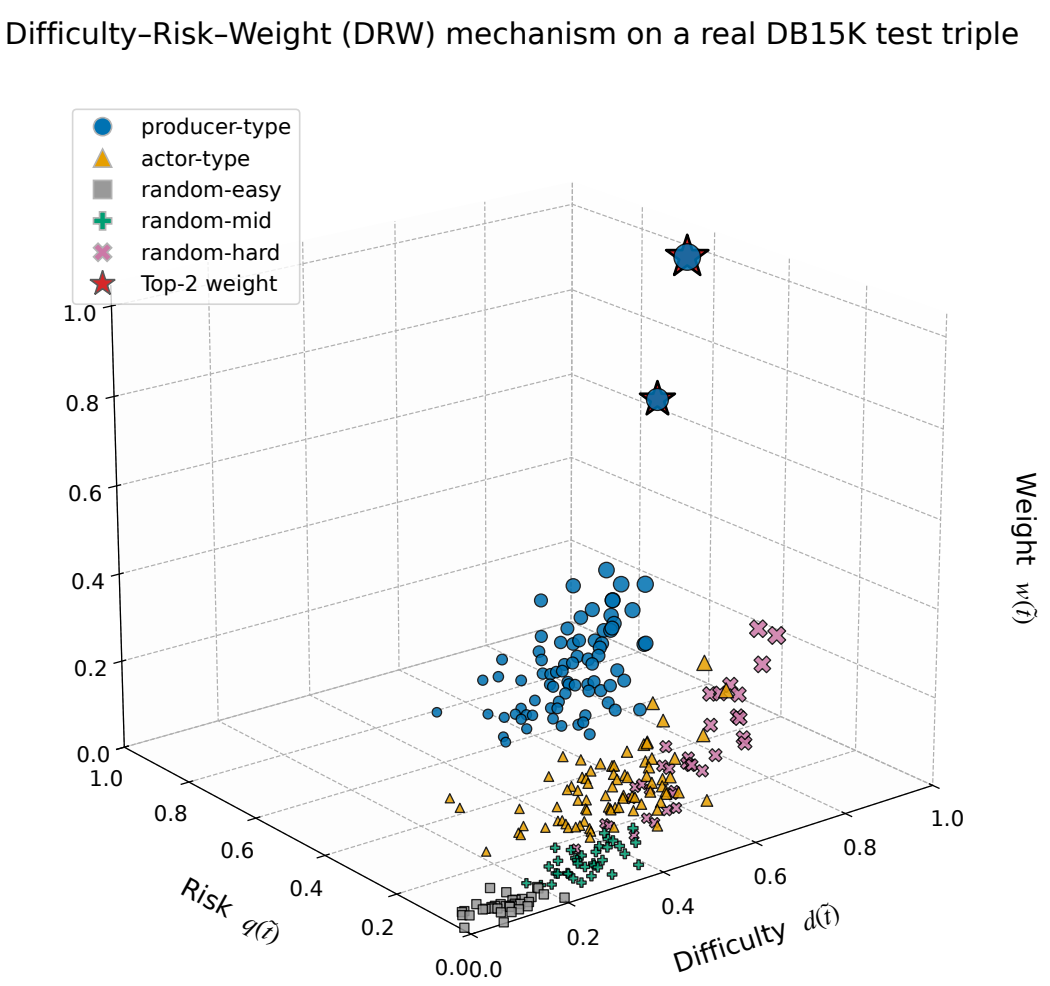


Figure 6: **DRW (Difficulty–Risk–Weight) visualization on a real DB15K test triple.** Each point corresponds to a negative tail candidate $\tilde{t}$ for $\langle h, r, t\rangle$. The x-axis is difficulty $d(\tilde{t})$, the y-axis is risk $q(\tilde{t})$, and the z-axis is the normalized training weight $w(\tilde{t})$ defined in Eq. equation 40. Point size is proportional to $w(\tilde{t})$. Colors/markers denote different negative sources (e.g., *producer-type*, *actor-type*, and random negatives with three difficulty tiers). The **Top-2** highest-weight negatives are emphasized with **red stars**, illustrating that DRW prioritizes *hard-but-safe* negatives while suppressing high-risk candidates.

mate (i) a *difficulty* score $d(\tilde{t}) \in [0, 1]$ that reflects how confusing the negative is to the current model (harder negatives yield larger d), and (ii) a *risk* score $q(\tilde{t}) \in [0, 1]$ that measures the probability of $\tilde{t}$ being a *false negative* (higher q indicates a larger chance that $\langle h, r, \tilde{t}\rangle$ is actually plausible/true). DRW then assigns a training weight:

$$w(\tilde{t}) \;\propto\; \exp\Big(\frac{d(\tilde{t})}{\tau}\Big) \cdot \exp\big(-\lambda\, q(\tilde{t})\big), \tag{40}$$

where $\tau > 0$ controls the amplification of hard negatives and $\lambda > 0$ controls the penalty on high-risk (potentially false) negatives. In the visualization, the x-, y-, and z-axes correspond to $d(\tilde{t})$, $q(\tilde{t})$, and the normalized weight $w(\tilde{t})$, respectively; point size is also proportional to $w(\tilde{t})$. We further distinguish negatives by their semantic/type source (e.g., *producer-type*, *actor-type*, and three tiers of random negatives), and mark the **Top-2** highest-weight negatives with **red stars** to highlight the most influential training signals under DRW.

Observations. Fig. 6 reveals three consistent patterns. **(1) Hard-and-safe negatives are strongly upweighted.** Negatives with large $d(\tilde{t})$ but moderate/low $q(\tilde{t})$ form prominent high-$w(\tilde{t})$ points near the top of the plot; the red-star points fall into this region, indicating that DRW prioritizes negatives that are both informative and relatively safe. **(2) Risky negatives are suppressed even when they are hard.** When $q(\tilde{t})$ is high, the penalty term $\exp(-\lambda q(\tilde{t}))$ dominates, leading to noticeably smaller $w(\tilde{t})$ despite non-trivial difficulty, thus reducing gradient contamination from potential false negatives. **(3) Negatives exhibit structured distributions by source/type.** Type-compatible negatives (e.g., *producer-type*) and context-relevant negatives (e.g., *actor-type*) tend to occupy higher-difficulty regions, whereas random negatives are more dispersed and generally receive lower weights, matching the intuition that semantically closer negatives provide stronger supervision. Overall, this visualization corroborates that DRW simultaneously *amplifies difficult yet reliable negatives* and *attenuates high-risk candidates*, improving training stability under noisy multimodal evidence.