EVIDENCE LATE, RELATION EARLY: ON WHEN TO AGGREGATE AND CONDITION IN MULTIMODAL KNOWLEDGE GRAPH COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal knowledge graph completion (MMKGC) leverages multimodal information to infer missing facts in incomplete knowledge graphs. Most existing methods construct entity representations before triple scoring, leaving two fundamental design questions underexplored: *when to read out multimodal evidence* and *at what stage should relational guidance interact with multimodal evidence*. We revisit MMKGC from this perspective and identify two complementary principles: **evidence-late**, which defers the compression of fine-grained evidence to the scoring stage, and **relation-early**, which incorporates relational information during representation learning to emphasize relation-relevant evidence. We first instantiate the evidence-late principle in **Origin**, enabling a controlled comparison with early evidence compression and showing that preserving fine-grained evidence until scoring yields better performance than premature compression into a fixed vector. Building on this observation, we propose **REEL**, a lightweight framework that operationalizes both principles in a unified design. Experiments on three MMKGC benchmarks show that REEL achieves strong completion performance, while ablation studies and case analyses further validate the effectiveness of both design principles. Our code is available at `https://anonymous.4open.science/r/reel-1955`.

1 INTRODUCTION

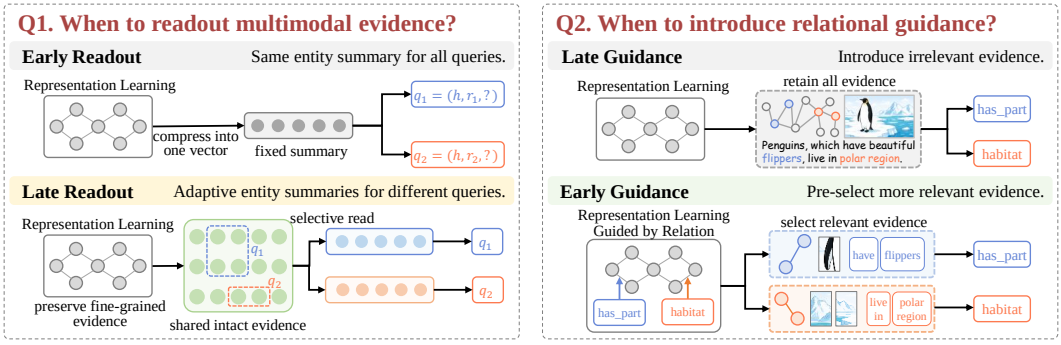
Knowledge graph completion (KGC) (Bordes et al., 2013; Nickel et al., 2016) aims to infer missing facts in incomplete KGs (Hogan et al., 2021) and supports applications such as question answering (Karpukhin et al., 2020; Lewis et al., 2020; Zhuo et al., 2026), recommendation (Koren et al., 2009), and information retrieval (Robertson & Zaragoza, 2009; Khattab & Zaharia, 2020). Multimodal KGC (MMKGC) (Pezeshkpour et al., 2018) extends KGC by integrating structural and multimodal entity information, thereby providing a richer evidence space for factual reasoning.

Existing MMKGC methods (Shang et al., 2024; Zhang et al., 2025) typically compress multimodal evidence into a fixed summary before triple scoring, which is subsequently shared across different queries. Consequently, fine-grained evidence is aggregated before its relevance to the current query is known, while entity features are often learned with limited interaction between relational semantics and multimodal evidence formation. Two more fundamental architectural questions deserve further investigation: *when should multimodal evidence be read out*, and *at what stage should relational guidance interact with multimodal evidence*, leading to two complementary design principles.

First, fine-grained multimodal evidence should remain accessible throughout representation learning rather than being prematurely compressed before scoring. As illustrated in Figure 1a, early-readout collapses the learned evidence into a single fixed summary during representation learning, preventing the scorer from directly accessing the fine-grained evidence tokens. This motivates our first design principle, **evidence-late**: defer evidence readout to the scoring stage, allowing the scorer to directly access and selectively exploit fine-grained evidence for the current query.

Second, rather than deferring relational information to the scoring stage, it should be incorporated earlier to guide multimodal evidence representation. As illustrated in Figure 1b, different relations may rely on distinct subsets of evidence, making relation-aware evidence selection essential for

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.



(a) (b)

Figure 1: (a) **Evidence-Late**: Preserve fine-grained evidence until readout, avoiding premature compression into a fixed summary. (b) **Relation-Early**: Introduce the target relation before evidence selection, so that relevant evidence can be identified while irrelevant observations are excluded.

forming query-relevant entity features. This motivates our second design principle, **relation-early**: incorporate relational information during entity representation learning so that the resulting features emphasize evidence relevant to the queried relation.

Together, these principles avoid prematurely constructing a query-independent entity representation. Evidence-late preserves fine-grained evidence for adaptive selection, while relation-early provides the relational signal to guide it. We first examine the evidence-late principle in isolation by constructing **Origin**, a full-token variant using FacetE as the scorer, and comparing it with a [CLS]-based counterpart under otherwise comparable settings. Building on this validation, we further propose **REEL**, a lightweight MMKGC approach that integrates both principles. Empirical results support the benefit of injecting relational information early in representation learning, and REEL achieves strong completion performance on three MMKGC benchmarks. Ablation studies and case analyses further examine the contributions and practical effects of the proposed design choices.

The remainder of this paper is organized as follows. Section 2 introduces the MMKGC task and summarizes the common pipeline of existing MMKGC methods. In Section 3, we introduce Origin and conduct controlled experiments to systematically evaluate the evidence-late paradigm. Section 4 presents REEL, a method developed following the two design principles, together with further empirical analysis of relation-early. Section 5 evaluates REEL on three benchmark datasets and provides ablation studies and case analyses. Section 6 discusses related work.

2 Preliminaries

2.1 Task Formulation

A MMKG is denoted by $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{X})$, where $\mathcal{E}$ and $\mathcal{R}$ denote the sets of entities and relations, respectively, and $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of observed relational facts. The collection $\mathcal{X}$ contains the multimodal observations associated with the entities. Let $\mathcal{M}$ denote the set of available modalities, including the structural modality, and let $m \in \mathcal{M}$ index a specific modality. For each entity $e \in \mathcal{E}$, $\mathbf{x}_e^m$ denotes its modality-specific information under modality m. Features for non-structural modalities are extracted using modality-specific encoders (Dosovitskiy et al., 2021; Devlin et al., 2019), whereas structural representations are learned from the graph topology.

Given an observed MMKG, MMKGC aims to infer missing entities from multimodal evidence through tail and head prediction queries, $(h, r, ?)$ and $(?, r, t)$, respectively. Candidate entities $e \in \mathcal{E}$ are ranked by plausibility, with the objective of placing the ground-truth entity above corrupted candidates. To provide contrastive supervision during training, for each observed triple $(h, r, t) \in \mathcal{T}$, negative triples are constructed by corrupting either the head or tail entity:

$$\mathcal{T}^-(h, r, t) = \{(h', r, t) \mid h' \in \mathcal{E} - \{h\}\} \cup \{(h, r, t') \mid t' \in \mathcal{E} - \{t\}\}, \quad (1)$$

excluding corrupted triples that also belong to $\mathcal{T}$. In practice, a subset of these valid negative candidates may be sampled for optimization.

2.2 General Formulation of MMKGC

Despite substantial architectural differences, existing MMKGC methods can be described by two stages: multimodal representation learning and triple scoring. We formulate the first stage as

$$\mathbf{z}_e = \mathcal{A}_\theta\left(\left\{\mathbf{P}_m \begin{cases} f_{\mathrm{str}}(e;\mathcal{G}), & m = \mathrm{str}, \\ f_m(\mathbf{x}_e^m), & m \neq \mathrm{str} \end{cases}\right\}_{m\in\mathcal{M}}; \mathbf{r}\right), \tag{2}$$

where $f_{(\cdot)}$ denotes modality-specific feature extractors, and $\mathbf{P}_m$ projects the resulting heterogeneous features into compatible latent spaces. The operator $\mathcal{A}_\theta$ learns the multimodal entity representation $\mathbf{z}_e$ from the aligned features of all modalities. The triple-scoring stage can be expressed under the following general formulation:

$$\mathbf{s}_\theta(h, r, \cdot) = \mathcal{S}_\theta\big(\underbrace{\mathcal{Q}_\theta(\mathbf{z}_h, \boldsymbol{r})}_{\mathbf{q}_{h,r}}, \mathbf{C}\big), \tag{3}$$

where $\mathcal{Q}_\theta$ combines $\mathbf{z}_h$ with the relation $\mathbf{r}$ to construct the query representation $\mathbf{q}_{h,r}$. The scorer $\mathcal{S}_\theta$ then ranks the candidates in $\mathbf{C}$. In practice, $\mathbf{C}$ typically stacks the structural embeddings of all candidate entities, and the scores are computed as $\mathbf{s}_\theta(h, r, \cdot) = \mathbf{q}_{h,r}\mathbf{C}^\top$. This formulation captures the common pipeline of existing methods while allowing different designs for multimodal fusion and reasoning, thereby providing a unified framework for characterizing and comparing architectures.

3 Evidence Late

Most existing MMKGC methods adopt an early-readout paradigm, in which multimodal representations are aggregated into a fixed summary vector during representation learning. However, this query-agnostic summary is reused across queries, limiting the model's ability to prioritize query-relevant evidence. In this section, we empirically investigate this limitation through controlled experiments on readout strategies.

3.1 Controlled Comparison Setup: Origin

Early-readout compresses evidence into an entity-level summary before scoring, whereas late-readout retains token-level evidence until scoring, yielding representations in $\mathbb{R}^{1\times d}$ and $\mathbb{R}^{n\times d}$, respectively. This granularity difference is inherent to the readout strategies and cannot be removed without altering the mechanisms being compared. We therefore keep the representation learning process matched across the two variants while preserving their native readout granularities. We instantiate this controlled setting as *Origin*, and introduce a token-level scorer for late-readout so that its representations can be evaluated without additional pre-scoring compression.

Matched Representation Learning To control the comparison as closely as possible, we keep the representation learning procedure identical across both variants, using a shared Transformer (Vaswani et al., 2017) encoder and varying only the readout strategy:

$$\left(\mathbf{z}_e^{\mathrm{early}}, \mathbf{z}_e^{\mathrm{late}}\right) = \mathrm{Transformer}(\mathrm{Concat}[\mathbf{c}, \mathbf{X}_e]), \quad \mathbf{X}_e = \{\mathbf{x}_e^m\}_{m\in\mathcal{M}}, \tag{4}$$

where $\mathbf{c}$ is a learnable `[CLS]`-token. Early-readout takes the final `[CLS]`-token as the global summary, following common practice in multimodal representation learning and MMKGC (Chen et al., 2022; Wang et al., 2026), whereas late-readout directly exposes the token-level outputs to the scorer. We further exclude relational information from the encoder input, preventing relation-specific interactions before readout and thereby reducing confounding factors in the comparison.

Scoring Function: FacetE Conventional scorers assume a single vector per entity and therefore cannot directly operate on token-level representations by late-readout. To support token-level entity evidence, we introduce FacetE, a lightweight scorer that adaptively aggregates token representations conditioned on the query. Specifically, FacetE estimates the relevance of each evidence token to the relation and normalizes the relevance scores into token-wise weights:

$$\mathbf{q}_{e,r}^{\mathrm{late}} = \underbrace{\mathrm{softmax}\left(\frac{(\mathbf{r}\mathbf{W}^Q)(\mathbf{z}_e^{\mathrm{late}}\mathbf{W}^K)^\top}{\sqrt{d_k}}\right)}_{\boldsymbol{\alpha}}(\mathbf{z}_e^{\mathrm{late}}\mathbf{W}^V)\mathbf{W}^O, \tag{5}$$

where α denotes the relevance weights assigned to entity evidence tokens conditioned on the relation. By deferring aggregation to the scoring stage, FacetE dynamically adjusts the contribution of evidence according to the relation, yielding an entity summary tailored to each query.

3.2 Evaluation Results

Under the *Origin* setup described above, we empirically compare the two readout strategies on standard MMKGC benchmarks. Implementation details are provided in Section 5.1. For early-readout, we evaluate five widely used KGC scorers: DistMult (Yang et al., 2015), ConvE (Dettmers et al., 2018), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), and TuckER (Balazevic et al., 2019). These scorers cover representative bilinear, complex-valued, rotational, convolutional, and tensor-factorization formulations, providing broad coverage of common KGC scoring paradigms. As shown in Table 1, the strongest early-readout scorer differs across datasets: ComplEx performs best on DB15K, whereas TuckER achieves the best early-readout results on MKG-W and MKG-Y. FacetE nevertheless outperforms the best early-readout scorer on each of the three benchmarks. Although the per-dataset gains are modest, FacetE achieves an average MRR of 37.53%, compared with 36.86% for TuckER, the strongest early-readout scorer on average. These results show that deferring evidence readout to the scoring stage is feasible and enables more effective evidence utilization across heterogeneous link prediction queries.

Table 1: Early- vs. late-readout performance, evaluated by MRR (%).

Scorer	DB15K	MKG-W	MKG-Y	Avg.
Early readout				
DistMult	35.48±0.12	32.28±0.09	35.87±0.15	34.54
ConvE	34.98±0.08	35.42±0.14	38.19±0.11	36.20
ComplEx	37.57±0.18	34.05±0.10	36.46±0.13	36.03
RotatE	37.23±0.16	33.99±0.07	36.21±0.12	35.81
TuckER	35.72±0.14	36.10±0.09	38.70±0.17	36.86
Late readout				
FacetE	37.61±0.11	36.22±0.08	38.77±0.10	37.53

3.3 Modality Masking Experiment

To examine whether the query-dependent modality weights learned under late-readout meaningfully reflect evidence importance, we conduct a modality masking experiment, with results reported in Figure 2. For each query, we mask either the modality assigned the largest weight or the modality assigned the smallest weight in α and measure the resulting performance degradation. Masking either modality reduces performance, indicating that both contribute useful evidence. More importantly, masking the highest-weighted modality consistently causes a larger degradation than masking the lowest-weighted one, suggesting that α captures the relative utility of each modality for individual queries. This distinction is more pronounced on DB15K and MKG-W than on MKG-Y, potentially reflecting dataset-specific differences in cross-modal evidence redundancy. These results provide complementary evidence that the query-dependent weighting mechanism in late-readout identifies modalities that are more consequential for the current prediction.

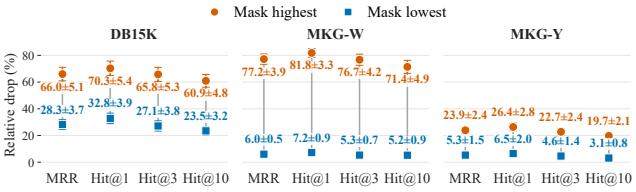


Figure 2: **Modality masking analysis**. Relative performance drops caused by masking the highest- or lowest-weighted modality during scoring across three datasets.

4 Relation Early, Evidence Late

In this section, we investigate how introducing relation information earlier into representation learning can facilitate relation-aware multimodal evidence selection.

4.1 Relation Early

For a given query, different pieces of multimodal evidence may vary substantially in their relevance to the queried relation, as entity observations often contain diverse semantic cues that are not equally informative for relational reasoning. Representations learned without considering the query relation

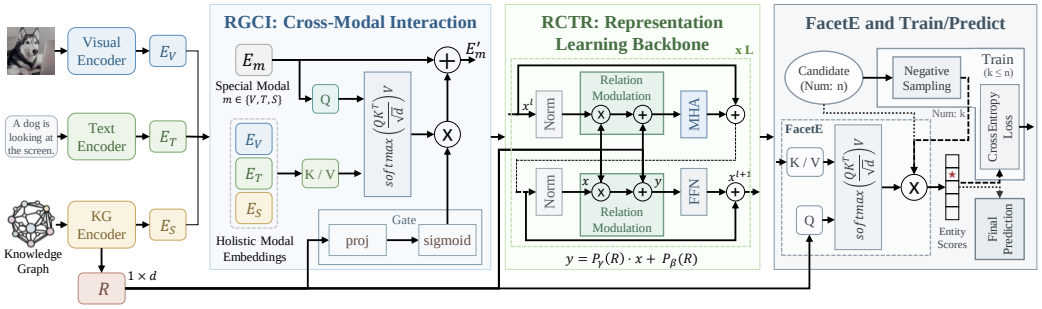


Figure 3: **Overall framework of REEL**. Given multimodal tokens of the source entity and the relation embedding, RGCI first enables cross-modal interaction, RCTR then performs token refinement, and FacetE finally conducts late evidence readout for prediction.

may thus retain distracting information while obscuring the evidence that is most useful for resolving the current relation. For example, for the query (`Penguin`, `has_part`, ?) in Figure 1b, visual features should emphasize anatomical regions of the penguin rather than background content, while textual features should focus on phrases describing constituent parts.

This motivates relation-early: introducing relation information during representation learning to selectively emphasize relevant fine-grained evidence and suppress irrelevant cues. In this way, the learned features are better aligned with the semantic requirements of the current relation before subsequent evidence aggregation and prediction.

4.2 METHOD: REEL

Following the principles of **R**elation **E**arly and **E**vidence **L**ate, we propose REEL. Figure 3 illustrates the overall architecture.

Relation-Gated Cross-Modal Interaction Different modalities characterize the same entity from heterogeneous yet complementary perspectives. Processing them independently limits the ability to capture cross-modal dependencies, whereas unrestricted interaction may introduce evidence irrelevant to the queried relation and obscure predictive signals. We therefore propose a ***R**elation-**G**ated **C**ross-modal **I**nteraction* (RGCI) module. For each modality, RGCI retrieves complementary evidence from the joint multimodal context while retaining the original representation as a semantic anchor. A relation-conditioned gate then controls the contribution of the retrieved evidence:

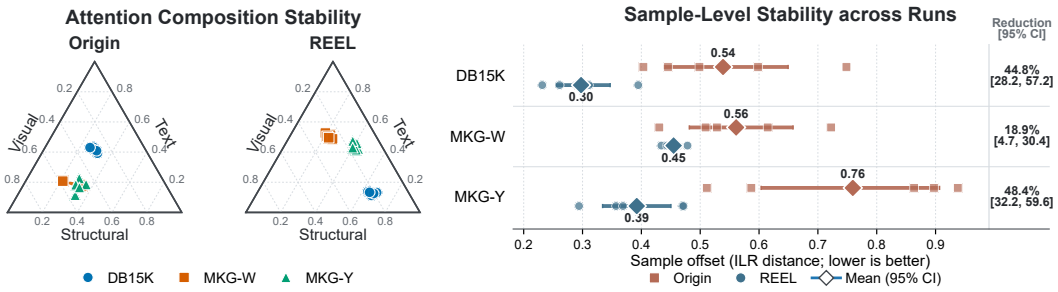
$$\widehat{\mathbf{x}}_e^m = \mathbf{x}_e^m + \alpha_r^m \,\mathrm{MHA}(\mathbf{x}_e^m, \mathbf{X}_e, \mathbf{X}_e)\,, \quad \alpha_r^m = \sigma(g^m(\mathbf{r}))\,, \tag{6}$$

where $g^m(\cdot)$ denotes a modality-specific projection function and $\sigma(\cdot)$ denotes the sigmoid function. The attention operator uses $\mathbf{x}_e^m$ as query and the joint multimodal context $\mathbf{X}_e$ as key and value, enabling the retrieval of complementary evidence from other modalities. The relation-conditioned gate is broadcast across tokens to modulate the retrieved evidence. This design supports relation-specific information exchange while suppressing cross-modal signals irrelevant to the query.

Relation-Conditioned Token Refinement While RGCI controls information exchange across modalities, relation-specific prediction also requires the features to adapt to the queried relation. Each token may encode multiple latent attributes, only a subset of which is relevant to a given relation. We therefore introduce a ***R**elation-**C**onditioned **T**oken **R**efinement* (RCTR) module that incorporates relational information into feature transformations, thereby adapting representations to the prediction target. Given the multimodal sequence $\widehat{\mathbf{X}}_e$, RCTR applies relational scaling and shifting to its normalized features:

$$\widetilde{\mathbf{X}}_e = \boldsymbol{\gamma}(\mathbf{r}) \odot \mathrm{LN}\left(\widehat{\mathbf{X}}_e\right) + \boldsymbol{\beta}(\mathbf{r}), \tag{7}$$

where $\boldsymbol{\gamma}(\mathbf{r})$ and $\boldsymbol{\beta}(\mathbf{r})$ are generated from the relation and serve as feature-wise scaling and shifting factors, respectively. By conditioning features on the queried relation before subsequent transformations, RCTR aligns the multimodal representation with the prediction target and promotes the attributes most informative for relation-specific reasoning.



(a) (b)

Figure 4: (a) Dataset-level mean attention composition across modalities for independent runs. (b) Sample-level modality offsets across runs, measured by ILR distance; lower values indicate greater reproducibility. Diamonds and error bars denote the mean and 95% confidence interval.

Training and Efficiency RGCI and RCTR jointly produce relation-aware multimodal representations, which FacetE reads out only at the final scoring stage. We train REEL with negative sampling and a cross-entropy objective. Despite the Transformer backbone in RCTR, REEL remains lightweight, benefiting from the compact design of the RGCI and FacetE, together with the efficient relation modulation in RCTR. Further efficiency analysis is provided in Appendix D.

4.3 Evidence for Relation Early

To quantify the effect of relational information on multimodal representation learning, we conduct independently initialized runs of Origin and REEL. We assess whether relational guidance consistently shifts modality composition and whether these patterns support the relation-early principle.

As shown in Figure 4a, independent runs form compact clusters under both methods, indicating that the aggregate modality compositions are reproducible. Under Origin, MKG-W and MKG-Y have nearly overlapping mean compositions, whereas REEL yields clearly separated cluster centers. Given the similar construction pipelines of the two datasets (Xu et al., 2022), their strong overlap under Origin suggests that, without relational guidance, coarse evidence allocation is largely governed by shared dataset-level statistics rather than relation-specific structure. Moreover, visual weight consistently decreases across datasets, while textual and structural evidence are redistributed differently. This trend suggests that relational guidance enables the model to better distinguish relevant evidence from less informative visual cues, while adaptively reallocating the remaining evidence across modalities.

At the sample level, Figure 4b shows that REEL consistently reduces cross-run discrepancy in compositional attention, lowering the average *Isometric Log-Ratio transform* (ILR) distance (Egozcue et al., 2003) from 0.54 to 0.30 on DB15K, from 0.56 to 0.45 on MKG-W, and from 0.76 to 0.39 on MKG-Y. This suggests that relational information constrains evidence selection and yields more consistent compositions across independent runs. Overall, relational guidance shift representation learning from generic feature aggregation toward relation-aware evidence selection. The modality redistribution and improved cross-run consistency support the relation-early principle as an effective design for MMKGC.

5 Experiments and Results

5.1 Datasets and Experiments Settings

Datasets We conduct experiments on three widely used MMKGC benchmarks: DB15K (Liu et al., 2019), MKG-W (Xu et al., 2022), and MKG-Y (Xu et al., 2022). These benchmarks provide a testbed for comparison with strong baselines across diverse multimodal settings. Detailed dataset descriptions are provided in Appendix A.

Baselines. We compare REEL with six representative MMKGC baselines: OTKGE (Cao et al., 2022), NativE (Zhang et al., 2024a), MoMoK (Zhang et al., 2024b), MyGO (Zhang et al., 2025),

Table 2: **Experimental results on three MMKGC benchmark datasets.** The best and second-best mean performance are highlighted in **bold** and underlined, respectively.

Model	DB15K				MKG-W				MKG-Y			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
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
NativE	34.30	25.08	39.48	51.35	36.84	29.94	40.06	49.39	39.21	35.03	41.21	46.25
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
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
MCKGC	39.79	31.92	43.80	54.66	36.88	31.32	38.92	47.43	38.92	35.49	40.57	45.21
APKGC	36.40	28.20	41.30	52.70	37.40	30.60	40.40	**50.10**	38.41	34.52	40.27	45.02
REEL	**39.92**	32.25	**43.92**	**55.81**	**38.07**	**32.42**	**40.58**	50.01	**39.64**	**36.12**	**41.38**	**46.75**
± Std.	0.12	0.16	0.21	0.09	0.08	0.11	0.13	0.17	0.18	0.06	0.13	0.21

MCKGC (Gao et al., 2025) and APKGC (Jian et al., 2025). These methods cover recent advances in MMKGC, including multimodal entity representation learning, modality-aware fusion, robust multimodal learning, and missing-modality handling. All compared methods are non-LLM-based, enabling a fair comparison under the conventional MMKGC setting.

Metrics. We evaluate link prediction performance using widely adopted rank-based metrics, including *Mean Reciprocal Rank* (MRR) (Sun et al., 2019) and Hits@K with $K = 1, 3$, and 10 (Bordes et al., 2013). MRR captures the ranking quality by averaging the reciprocal ranks of correct entities, whereas Hits@K measures the proportion of correct entities that appear among the top-K predictions. To comprehensively assess model effectiveness, these metrics are reported for both head prediction $(h, r, ?)$ and tail prediction $(?, r, t)$. Please refer to Appendix B for details.

Implementation Details. All experiments were implemented in PyTorch 2.4.1 and conducted on a server with an NVIDIA A100-SXM4 40GB GPU. We optimized all models using AdamW (Loshchilov & Hutter, 2019) for 1000 epochs, with a learning rate schedule consisting of linear warmup followed by cosine annealing. EMA (Salimans & Kingma, 2016) was adopted to improve training stability. The embedding dimension was fixed to 256, and the RCTR module used 8 layers. More detailed experimental configurations are provided in Appendix C.

5.2 Main Results

Table 2 compares REEL with representative MMKGC methods on three benchmarks. On DB15K, it obtains the best results on MRR, Hit@3 and Hit@10 and remains very close to the strongest baselines on Hit@1. On MKG-W, it leads on the main ranking metrics and is only slightly behind the best Hit@10 result. On MKG-Y, it achieves the best performance on all metrics. Overall, these results demonstrate that REEL provides consistent improvements in both ranking quality and top-ranked prediction accuracy across diverse MMKGC benchmarks.

5.3 Ablation Study

Component Ablation We first examine the two components in the representation learning stage, i.e., RGCI and RCTR. As shown in Table 3, removing either component generally degrades performance, confirming the necessity of representation learning in REEL. In particular, removing RCTR leads to a more pronounced degradation across datasets, which is consistent with its role as the main backbone for feature refinement, while RGCI primarily facilitates cross-modal interaction. We further remove relation information from RGCI, RCTR, or both to examine the relation-early principle. These variants generally underperform the full model, indicating that incorporating relation information during representation learning provides useful guidance for learning task-relevant features. Together, these results support both the effectiveness of the proposed representation learning modules and the motivation for relation-aware early interaction.

Table 3: Ablation study on model components by MRR (%).

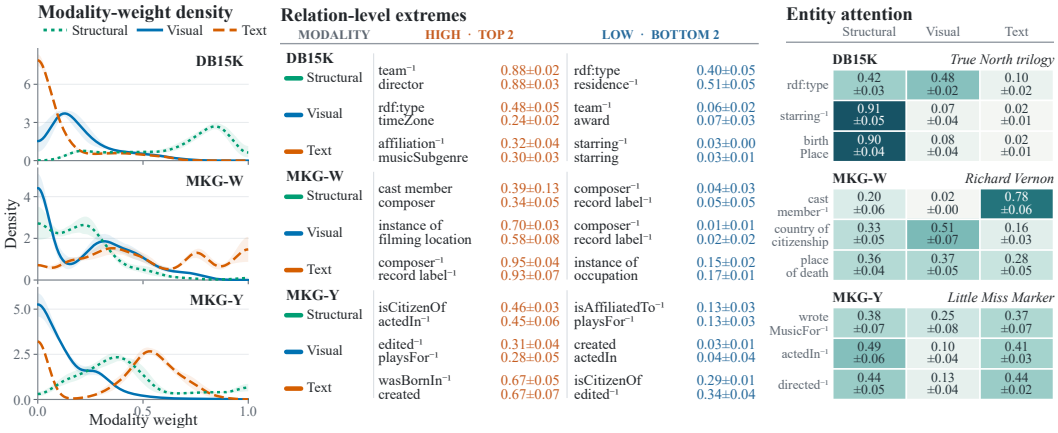
Setting	DB15K	MKG-W	MKG-Y	Avg.
w/o RGCI	37.44±0.14	36.79±0.21	39.54±0.09	37.92
w/o RCTR	35.38±0.18	35.27±0.11	38.48±0.16	36.38
RGCI w/o Rel.	37.54±0.07	38.01±0.19	39.35±0.13	38.30
RCTR w/o Rel.	37.27±0.22	37.76±0.08	39.22±0.15	38.08
w/o Rel.	36.74±0.17	37.46±0.10	39.39±0.23	37.86

Scoring Function Ablation As part of the ablation study, we replace FacetE in REEL with the same alternative scoring functions evaluated under Origin. FacetE consistently achieves the best performance across all three datasets, confirming that its advantage remains under relation-aware representation learning. Meanwhile, REEL improves the performance of the alternative scorers over their corresponding Origin variants, showing that relational guidance provides consistent gains beyond the choice of scoring function. It therefore provides further empirical evidence that the relation-early and evidence-late principles are complementary: the former improves relation-aware evidence representation, while the latter enables more effective evidence aggregation during scoring.

Table 4: Ablation study on scoring functions by MRR (%).

Scorer	DB15K	MKG-W	MKG-Y	Avg.
DistMult	36.99±0.17	33.83±0.24	35.84±0.12	35.55
ConvE	35.65±0.21	35.69±0.14	38.36±0.26	36.57
ComplEx	34.32±0.11	34.19±0.19	36.43±0.15	34.98
RotatE	35.32±0.23	35.55±0.09	37.12±0.18	36.00
TuckER	35.05±0.16	36.10±0.28	38.79±0.13	36.65

5.4 Case Study: Modality Weight Analysis



(a) (b) (c)

Figure 5: (a) Distribution of modality attention weights, showing mean boundary-corrected kernel density estimates with shaded 95% bootstrap confidence intervals across runs. (b) Relations supported by at least 20 queries with the highest and lowest attention weights for each modality. (c) Modality attention weights across different relations for a representative entity selected from each dataset. In (b) and (c), r^{-1} denotes the inverse direction of relation r.

Figure 5 shows that REEL learns distinct modality preferences across datasets and relations. In Figure 5a, DB15K relies more heavily on structural information, while MKG-W and MKG-Y exhibit broader distributions across structure, vision, and text. This suggests that a fixed fusion strategy is suboptimal, since the informative modality varies substantially across queries.

Figure 5b further illustrates how modality preferences vary across datasets and relation directions. DB15K consistently favors structural information, with structural weights remaining at least 0.40 even for *rdf:type*, the relation with the lowest structural weight. In contrast, MKG-W exhibits a preference for textual information, whose weight remains at least 0.29 for *isCitizenOf*, the relation with the lowest textual weight. MKG-Y further shows pronounced directional asymmetry: *composer* and *composer*$^{-1}$ exhibit markedly different modality-weight distributions. This asymmetry suggests that reversing a relation can change which modalities provide useful evidence, potentially reflecting differences in entity roles, candidate spaces, or relation cardinalities across the two directions.

Figure 5c confirms that modality selection is also entity relation dependent. For the same DB15K entity *True North trilogy*, *rdf:type* favors visual information, while *starring*$^{-1}$ and *birthPlace* rely almost entirely on structure. Similarly, for *Richard Vernon* in MKG-W, *cast member*$^{-1}$ is text dominant, whereas *country of citizenship* assigns the largest weight to vision. These results indicate that REEL does not learn a static modality preference for each entity. Instead, it dynamically selects the most informative modality according to the current relation, providing an interpretable mechanism for adaptive multimodal fusion.

5.5 Robustness Analysis

Figure 6 shows that REEL remains the most robust method on MKG-W under both missing modalities and Gaussian noise. Under missing modalities, all methods degrade as the perturbation becomes stronger, highlighting the challenge of reasoning with incomplete multi-modal inputs. Nevertheless, REEL consistently achieves the highest MRR and maintains a clear advantage over MyGO and MoMoK, suggesting that it can better exploit the remaining evidence. Under Gaussian noise, REEL also exhibits relatively small performance variations while consistently outperforming the baselines, demonstrating stronger robustness to corrupted multimodal features. Overall, these results indicate that REEL is less sensitive to both missing and noisy evidence, enabling more reliable reasoning under realistic perturbations.

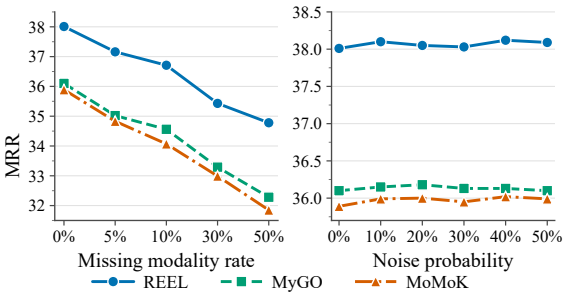


Figure 6: Experiments on MKG-W under missing modalities and Gaussian noise, evaluated by MRR and compared with MyGO and MoMoK.

6 Related Work

Multimodal fusion MMKGC extends conventional KGC by incorporating auxiliary evidence, such as images and textual descriptions, into entity and relation modeling. Early methods mainly fuse multimodal features with structural representations. IKRL (Xie et al., 2017) aggregates entity images into visual embeddings, while RSME (Wang et al., 2021) and OTKGE (Cao et al., 2022) further explore multimodal representation learning and cross-modal alignment. MKGformer (Chen et al., 2022) strengthens cross-modal interaction through Transformer-based architectures.

Adaptive multimodal modeling Subsequent work explores more adaptive mechanisms for exploiting multimodal information under different prediction contexts. MoSE (Zhao et al., 2022) learns modality-specific representations to capture complementary signals, while MM-RNS (Xu et al., 2022) enhances cross-modal interaction for more informative negative sampling. LAFA (Shang et al., 2024) introduces link-aware visual fusion to adapt visual information to individual predictions, and NativE (Zhang et al., 2024a) dynamically combines available modalities to better handle heterogeneous and incomplete multimodal evidence.

Fine-grained evidence modeling Recent methods further move toward finer-grained multimodal representations. VISTA (Lee et al., 2023) jointly models visual and textual information with Transformer-based architectures, while MyGO (Zhang et al., 2025) represents multimodal evidence as fine-grained discrete tokens and models their interactions. Despite these advances, multimodal evidence is typically aggregated into entity representations before triple scoring, so the scorer has limited direct access to the underlying fine-grained evidence.

7 Conclusion and Future Work

In this work, we revisit the conventional MMKGC pipeline that constructs query-independent entity representations before triple scoring. We introduce two complementary principles, evidence late and relation early, which respectively preserve fine-grained multimodal evidence for query-adaptive aggregation and incorporate relational information during representation learning. Through Origin, we validate the benefit of delaying evidence compression, and further develop REEL to integrate both principles. Experiments on three benchmarks, together with ablation studies and case analyses, demonstrate the effectiveness of these design choices.

Future work may explore more efficient mechanisms for preserving fine-grained evidence, stronger relation-conditioned evidence selection, and extensions to noisy or incomplete multimodal settings. It would also be interesting to investigate whether these principles generalize to other multimodal relational reasoning tasks.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks requiring disclosure, and the remaining required disclosure tasks are not applicable to this work. We used generative AI tools only to edit the manuscript for grammar and readability. All AI-assisted edits were reviewed by the authors to ensure that they did not alter the technical content, scientific claims, or intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We confirm that our work has been conducted in accordance with the ICLR Code of Ethics (`https://iclr.cc/public/CodeofEthics`). The study does not involve human subjects, sensitive personal data, or experiments that may cause harm to individuals or groups. The datasets used are publicly available and no personally identifiable information is included. Our methodology and findings are intended for academic purposes and do not pose foreseeable risks of misuse. We have carefully considered issues of fairness, bias, and privacy, and to the best of our knowledge, our research maintains integrity and complies with all applicable ethical standards.

REPRODUCIBILITY STATEMENT

We confirm that our study is reproducible. Specifically, we have submitted our anonymized project repository to an anonymous GitHub repository (`https://anonymous.4open.science/r/reel-1955`). In addition, we provide specific details of the experimental setup in the main text, including dataset details (A) and hyperparameter settings (5.1).

REFERENCES

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 5184–5193. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1522.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2787–2795, 2013. URL `https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html`.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Otkge: Multi-modal knowledge graph embeddings via optimal transport. In *Advances in Neural Information Processing Systems*, volume 35, pp. 39090–39102. Curran Associates, Inc., 2022. doi: 10.52202/068431-2833.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pp. 904–915. ACM, 2022. doi: 10.1145/3477495.3531992.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1811–1818. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11573.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3):279–300, April 2003. ISSN 1573-8868. doi: 10.1023/A:1023818214614.

Yuxiao Gao, Fuwei Zhang, Zhao Zhang, Xiaoshuang Min, and Fuzhen Zhuang. Mixed-curvature multi-modal knowledge graph completion. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 11699–11707. AAAI Press, 2025. doi: 10.1609/AAAI.V39I11.33273.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021. ISBN 978-3-031-00790-3. doi: 10.2200/S01125ED1V01Y202109DSK022.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.550.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075.

Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. VISTA: visual-textual knowledge graph representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pp. 7314–7328. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.488.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6 (2):167–195, 2015. doi: 10.3233/SW-140134.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. MMKG: multi-modal knowledge graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, Lecture Notes in Computer Science, pp. 459–474. Springer, 2019. doi: 10.1007/978-3-030-21348-0_30.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proc. IEEE*, 104(1):11–33, 2016. doi: 10.1109/JPROC.2015.2483592.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3208–3218. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1359.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 901, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html.

Bin Shang, Yinliang Zhao, Jun Liu, and Di Wang. LAFA: multimodal knowledge graph completion with link aware fusion and aggregation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 8957–8965. AAAI Press, 2024. doi: 10.1609/AAAI.V38I8.28744.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proc. VLDB Endow.*, 13(11):2326–2340, 2020. URL http://www.vldb.org/pvldb/vol13/p2326-sun.pdf.

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

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, JMLR Workshop and Conference Proceedings, pp. 2071–2080. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/trouillon16.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp.

5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. Is visual context really helpful for knowledge graph? A representation learning perspective. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 2735–2743. ACM, 2021. doi: 10.1145/3474085.3475470.

Ziyao Wang, Chen Chen, Jingtao Li, Weiming Zhuang, Jiabo Huang, Ang Li, and Lingjuan Lyu. Unicompress: Token compression for unified vision-language understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24663–24674, June 2026.

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, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 3140–3146. ijcai.org, 2017. doi: 10.24963/IJCAI.2017/438.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 3857–3866. ACM, 2022. doi: 10.1145/3503161.3548388.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6575.

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, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 91–101. ACM, 2024a. doi: 10.1145/3626772.3657800.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Mixture of modality knowledge experts for robust multi-modal knowledge graph completion. *CoRR*, abs/2405.16869, 2024b. doi: 10.48550/ARXIV.2405.16869.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, and Wen Zhang and-Huajun Chen. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 13322–13330. AAAI Press, 2025. doi: 10.1609/AAAI.V39I12.33454.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 10527–10536. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.719.

Xingrui Zhuo, Jiapu Wang, Gongqing Wu, Zhongyuan Wang, Jichen Zhang, Shirui Pan, and Xindong Wu. Knowledge reasoning language model: Unifying knowledge and language for inductive knowledge graph reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2g8EmFwNTB.

Table 5: Statistical information of the three MMKGs in our experiments. The image and text modality features are provided by the original datasets and kept the same for all baselines.

Dataset	Entity	Relation	Train	Valid	Test	Image		Text	
						Num	Dim	Num	Dim
DB15K	12842	279	79222	9902	9904	12818	4096	9078	768
MKG-W	15000	169	34196	4276	4274	14463	383	14123	384
MKG-Y	15000	28	21310	2665	2663	14244	383	12305	384

A Dataset

We provide additional details of the three datasets used in our experiments: DB15K (Liu et al., 2019), MKG-W (Xu et al., 2022), and MKG-Y (Xu et al., 2022). These benchmarks are widely used in MMKGC and related MMKG tasks. DB15K is a real-world MMKG derived from DBpedia (Lehmann et al., 2015) under the BSD-3-Clause License. It contains structural relational triples and is further enriched with image and textual modalities for entities, making it a standard benchmark for evaluating multi-modal reasoning and representation learning methods. MKG-W and MKG-Y are real-world MMKG benchmarks derived from Wikidata Vrandecic & Krötzsch (2014) and YAGO Sun et al. (2020), respectively, and augmented with visual and textual information. Overall, all three datasets contain structural triples together with image and text modalities, and they have been widely adopted in recent studies. Their use allows fair and consistent comparison with strong baselines in the literature. Summary statistics of the datasets are listed in Table 5.

B Evaluation Metrics

We evaluate MMKGC models on the link prediction task, where the goal is to rank the correct entity among all candidates for an incomplete triple $(h, r, ?)$ or $(?, r, t)$. Following standard practice in KGC (Bordes et al., 2013), we consider both head prediction and tail prediction, and adopt the filtered evaluation protocol, where candidate entities that form true triples already observed in the training, validation, or test sets are excluded from ranking. We report four widely used ranking-based metrics, including mean reciprocal rank (MRR) (Sun et al., 2019) and Hit@K with $K \in \{1, 3, 10\}$ (Bordes et al., 2013).

Formally, let Q denote the set of test queries, and for each query $q_i \in Q$, let rank_i be the rank of the ground-truth entity in the model's sorted candidate list. The evaluation metrics are defined as follows:

- Mean reciprocal rank (MRR) is computed as

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

- Hit@K measures the proportion of queries for which the ground-truth entity is ranked within the top-K positions:

$$\text{Hit@}K = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbb{I}(\text{rank}_i \leq K),$$

  where $\mathbb{I}(\cdot)$ is the indicator function.

Higher values of MRR and Hit@K indicate better predictive performance. Together, these metrics reflect both the quality of the overall ranking and the accuracy of top-ranked predictions.

In addition to ranking-based link prediction metrics, we use the isometric log-ratio (ILR) distance to quantify the stability of multimodal attention compositions across different random seeds and samples. Since the attention weights over the visual, textual, and structural modalities form a compositional vector whose elements are non-negative and sum to one, directly measuring their differences

in the original simplex with standard Euclidean distance may lead to an inappropriate characterization of compositional variation. We therefore map each attention composition to an unconstrained Euclidean space using the ILR transformation. Let $\mathbf{a} = (a_v, a_t, a_s)$ denote the normalized attention composition over the visual, textual, and structural modalities, respectively, where $a_v + a_t + a_s = 1$. We define its two-dimensional ILR representation as

$$\mathrm{ilr}(\mathbf{a}) = \begin{bmatrix} \frac{1}{\sqrt{2}} \log \frac{a_v}{a_t} \\ \frac{1}{\sqrt{6}} \log \frac{a_v a_t}{a_s^2} \end{bmatrix}. \tag{8}$$

Given two attention compositions $\mathbf{a}$ and $\mathbf{b}$, their ILR distance is computed as the Euclidean distance between the corresponding ILR coordinates:

$$d_{\mathrm{ILR}}(\mathbf{a}, \mathbf{b}) = \|\mathrm{ilr}(\mathbf{a}) - \mathrm{ilr}(\mathbf{b})\|_2 . \tag{9}$$

A smaller ILR distance indicates that two attention compositions exhibit more similar relative modality proportions, whereas a larger value indicates a greater shift in the balance among visual, textual, and structural information. In our stability analysis, lower ILR distances therefore correspond to more consistent attention behavior across runs or samples.

C Implementation Details

All experiments were implemented in PyTorch 2.4.1 and conducted on a server running Ubuntu 20.04, equipped with an Intel Xeon Platinum 8375C CPU and an NVIDIA A100-SXM4 40GB GPU. We optimized all models using AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 5×10^{-4} and a weight decay of 1×10^{-3}. The batch size was selected from $\{512, 1024, 2048\}$ based on validation performance, and 1024 was used in all experiments. Each model was trained for 1000 epochs. We adopted a learning rate schedule with a linear warmup over the first 30 epochs, followed by cosine annealing for the remaining epochs. Exponential moving average (EMA) (Salimans & Kingma, 2016) was further employed to improve training stability and accuracy. The RCTR module was instantiated with 8 layers. The embedding dimension for entities, modalities, and relations was fixed at 256. For embedding-level regularization, the dropout rates were initialized to 0.5 for structural embeddings and 0.1 for both visual and textual embeddings. No additional dropout was applied in other parts of the model. Instead, DropPath was used within the RCTR module with a rate of 0.1. For negative sampling, 64 negative samples were drawn for each positive sample.

D Efficiency Analysis

As shown in Table 6, REELachieves a favorable trade-off between effectiveness and efficiency. Although retaining token-level evidence until scoring could in principle increase computational and memory costs, the practical overhead remains small. Unless otherwise specified, the efficiency results in Table 6 are reported without negative sampling, so they reflect the base architectural cost of each method. In the full training setting for multi-modal knowledge graph completion, negative sampling incurs additional computation and memory overhead for all methods. These results suggest that preserving holistic multimodal evidence for late readout is practical and introduces only modest intrinsic overhead.

Table 6: Training efficiency on DB15K.

Method	Time/epoch	GPU memory
MoMoK	8.3 s	4366 MB
MyGO	22.9 s	15576 MB
OTKGE	65.8 s	2016 MB
REEL	**5.8 s**	**1758 MB**

E Scoring Functions

Given a partial triple $(h, r, ?)$, the model predicts the missing tail entity by assigning a score to each candidate entity $t \in \mathcal{E}$. In general, the score of a triple is written as

$$S(h, r, t) = \mu(\mathbf{h}, \mathbf{r}, \mathbf{t}),$$

where $\mathbf{h}, \mathbf{r}, \mathbf{t}$ denote the embeddings of the head entity, relation, and candidate tail entity, respectively, and $\mu(\cdot)$ is a model-specific scoring function.

Many knowledge graph embedding models first construct a query representation from the head entity and relation,

$$\mathbf{q}_{h,r} = \phi(\mathbf{h}, \mathbf{r}),$$

and then score each candidate tail by matching $\mathbf{q}_{h,r}$ with the tail embedding:

$$S(h, r, t) = \mathbf{q}_{h,r}^\top \mathbf{t}.$$

Different choices of $\phi(\cdot)$ define different interaction patterns between the head entity and the relation.

DistMult DistMult constructs the query representation through element-wise multiplication:

$$\mathbf{q}_{h,r} = \mathbf{h} \odot \mathbf{r}.$$

The score is then computed as

$$S(h, r, t) = (\mathbf{h} \odot \mathbf{r})^\top \mathbf{t}.$$

Equivalently, DistMult uses a tri-linear dot product among the head, relation, and tail embeddings.

TuckER TuckER models the interaction between the head entity and relation through a relation-specific linear transformation. Given a relation-specific matrix $\mathbf{W}_r$, generated from the relation embedding and a shared core tensor, the query representation can be written as

$$\mathbf{q}_{h,r} = \mathbf{W}_r^\top \mathbf{h}.$$

The score is computed by matching this query representation with the candidate tail embedding:

$$S(h, r, t) = \mathbf{q}_{h,r}^\top \mathbf{t}.$$

ConvE ConvE constructs the query representation by applying convolutional operations to reshaped head and relation embeddings. Specifically, $\mathbf{h}$ and $\mathbf{r}$ are first reshaped into two-dimensional feature maps $\overline{\mathbf{h}}$ and $\overline{\mathbf{r}}$, which are then concatenated and passed through a convolutional network:

$$\mathbf{q}_{h,r} = \sigma\big(\mathbf{W}\,\mathrm{vec}\left(\mathrm{Conv}\left([\overline{\mathbf{h}}; \overline{\mathbf{r}}]\right)\right)\big),$$

where $\mathrm{Conv}(\cdot)$ denotes a 2-dimensional convolution operation, $\mathrm{vec}(\cdot)$ denotes vectorization, $\mathbf{W}$ is a learnable projection matrix, and $\sigma(\cdot)$ is a nonlinear activation function. The score is then given by

$$S(h, r, t) = \mathbf{q}_{h,r}^\top \mathbf{t}.$$

ComplEx ComplEx extends bilinear scoring to the complex domain. Let $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$ denote complex-valued embeddings of the head entity, relation, and tail entity. The score is defined as

$$S(h, r, t) = \mathrm{Re}\left(\langle \mathbf{h}, \mathbf{r}, \overline{\mathbf{t}} \rangle\right) = \mathrm{Re}\left(\sum_{i=1}^{d} h_i r_i \overline{t_i}\right),$$

where $\overline{\mathbf{t}}$ denotes the complex conjugate of $\mathbf{t}$, $\langle \cdot, \cdot, \cdot \rangle$ denotes the tri-linear dot product, and $\mathrm{Re}(\cdot)$ extracts the real part. Equivalently, the complex query representation can be viewed as

$$\mathbf{q}_{h,r} = \mathbf{h} \odot \mathbf{r},$$

which is matched with the conjugated tail embedding in the complex space. The use of complex-valued embeddings allows ComplEx to model asymmetric relations.

RotatE RotatE also represents entities and relations in the complex space, but it uses a distance-based scoring function rather than an inner-product matching score. Let $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$, where each relation embedding $\mathbf{r}$ is constrained to have unit modulus. The relation is interpreted as an element-wise rotation, and the transformed head embedding is

$$\mathbf{q}_{h,r} = \mathbf{h} \odot \mathbf{r}.$$

The score is defined by the negative distance between the rotated head embedding and the tail embedding:

$$S(h, r, t) = -\left\|\mathbf{h} \odot \mathbf{r} - \mathbf{t}\right\|,$$

where $\|\cdot\|$ usually denotes the L_1 or L_2 norm. A higher score indicates that the candidate tail is closer to the rotated head embedding and is therefore more likely to form a valid triple.

For a given query $(h, r, ?)$, the model ranks all candidate entities $t \in \mathcal{E}$ according to $S(h, r, t)$, and the entity with the highest score is selected as the predicted tail.

F NEGATIVE SAMPLING

Since the MMKG only provides observed factual triples in $\mathcal{T}$, negative sampling is used during training to construct corrupted triples for contrastive learning. For a factual triple $(h, r, t) \in \mathcal{T}$, negative triples can be generated by randomly replacing the head or tail entity with another entity from $\mathcal{E}$. The corresponding corrupted set is written as

$$\mathcal{T}^- = \{(h', r, t) \mid (h, r, t) \in \mathcal{T},\ h' \in \mathcal{E} \setminus \{h\}\} \cup \{(h, r, t') \mid (h, r, t) \in \mathcal{T},\ t' \in \mathcal{E} \setminus \{t\}\}. \tag{10}$$

In practice, we mainly construct negatives by corrupting the target entity, for example, the tail entity for $(h, r, ?)$ prediction. To reduce false negatives, sampled candidates that correspond to known factual triples in $\mathcal{T}$ are filtered out. For tail corruption, we define the set of valid tails for a given (h, r) as

$$\mathcal{E}^+_{h,r} = \{\hat{t} \in \mathcal{E} \mid (h, r, \hat{t}) \in \mathcal{T}\},$$

and a sampled tail entity t' is accepted as a negative sample only if

$$t' \in \mathcal{E} \setminus \mathcal{E}^+_{h,r}.$$

Given a predefined number of negative samples k, the negative tail set for (h, r) is denoted as

$$\mathcal{N}(h, r) = \{t'_1, \ldots, t'_k\}, \qquad t'_i \in \mathcal{E} \setminus \mathcal{E}^+_{h,r}.$$

These negative samples are then used together with factual triples to optimize the scoring function $S(h, r, t)$, improving training efficiency while avoiding incorrect supervision from existing true triples. The same principle can be naturally extended to head corruption or bidirectional corruption.

G CASE STUDY DETAILS

Modality weight distribution Figure 7 shows the distributions of modality-specific weights on DB15K, MKG-W, and MKG-Y. The overall trends are consistent with the mean weights in Figure 4a. Compared with Origin, REEL exhibits more concentrated distributions and smaller variations, indicating more stable modality weighting.

Relation-specific modality weights Figure 8 and Figure 9 shows the modality weights of different relations on DB15K, MKG-W, and MKG-Y, obtained from $\boldsymbol{\alpha}$ in the FacetE module. Different relations exhibit distinct modality preferences, demonstrating that FacetE can adaptively capture relation-specific multimodal information.

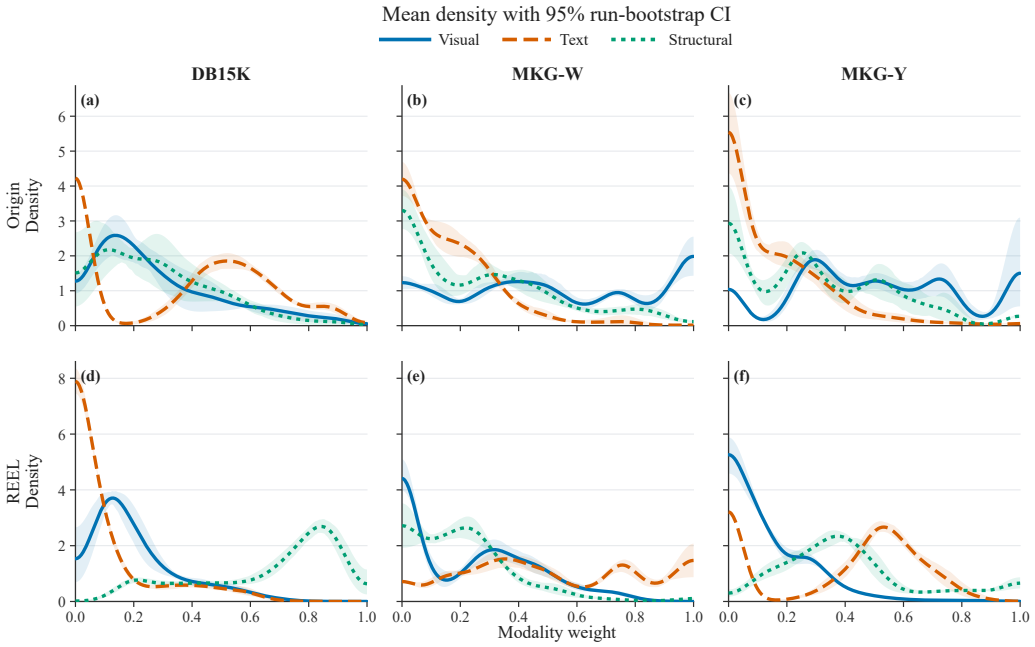


Figure 7: Distribution densities of modality-specific weights for all samples obtained by Origin and REEL on the DB15K, MKG-W, and MKG-Y datasets. The curves represent mean boundary-corrected kernel density estimates across multiple runs, with shaded regions indicating 95% bootstrap confidence intervals.

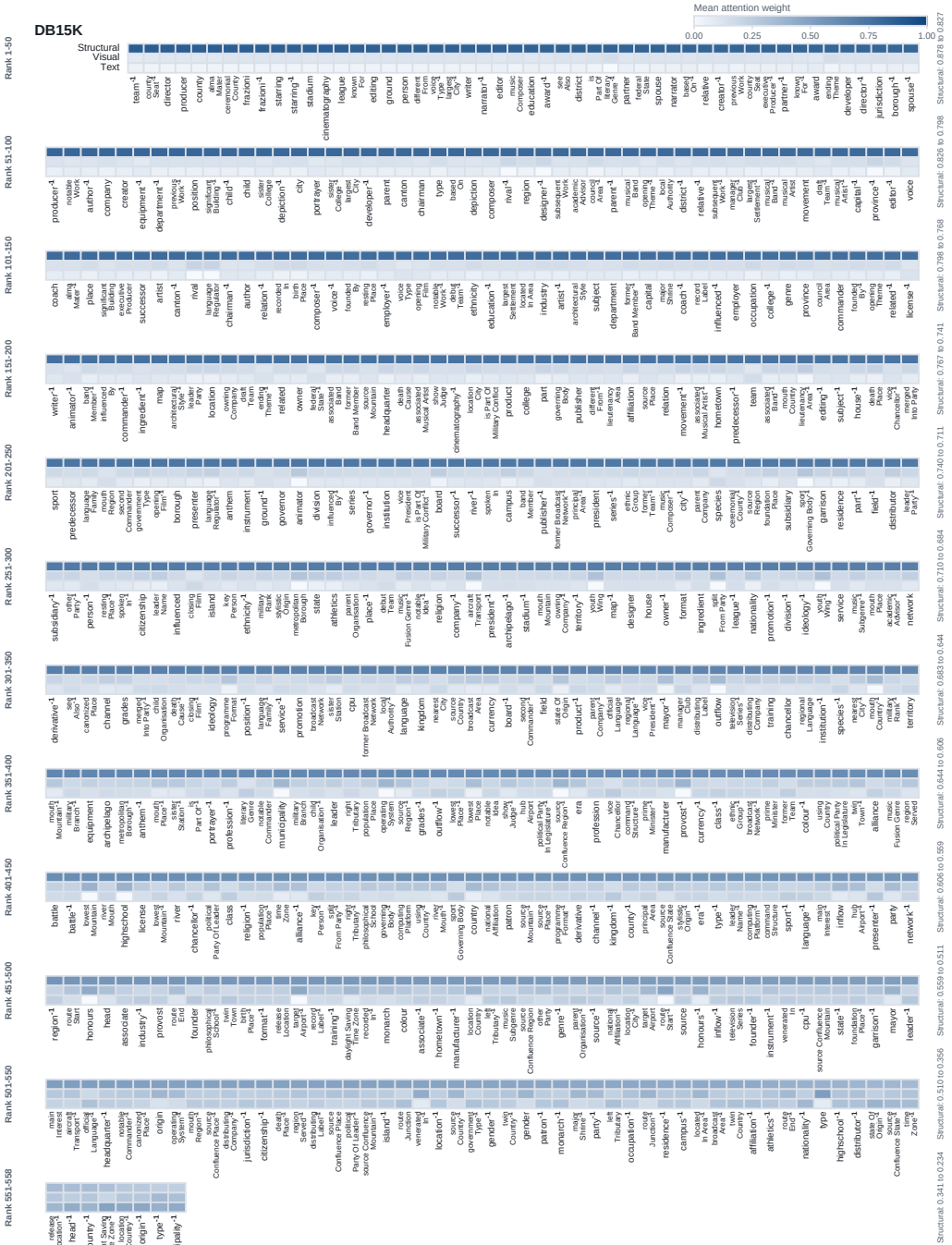


Figure 8: Modality-specific weight allocation for each relation on the DB15K dataset. Results are averaged over multiple runs, and all relations are sorted by their structural weights. r^{-1} denotes the inverse direction of relation r.

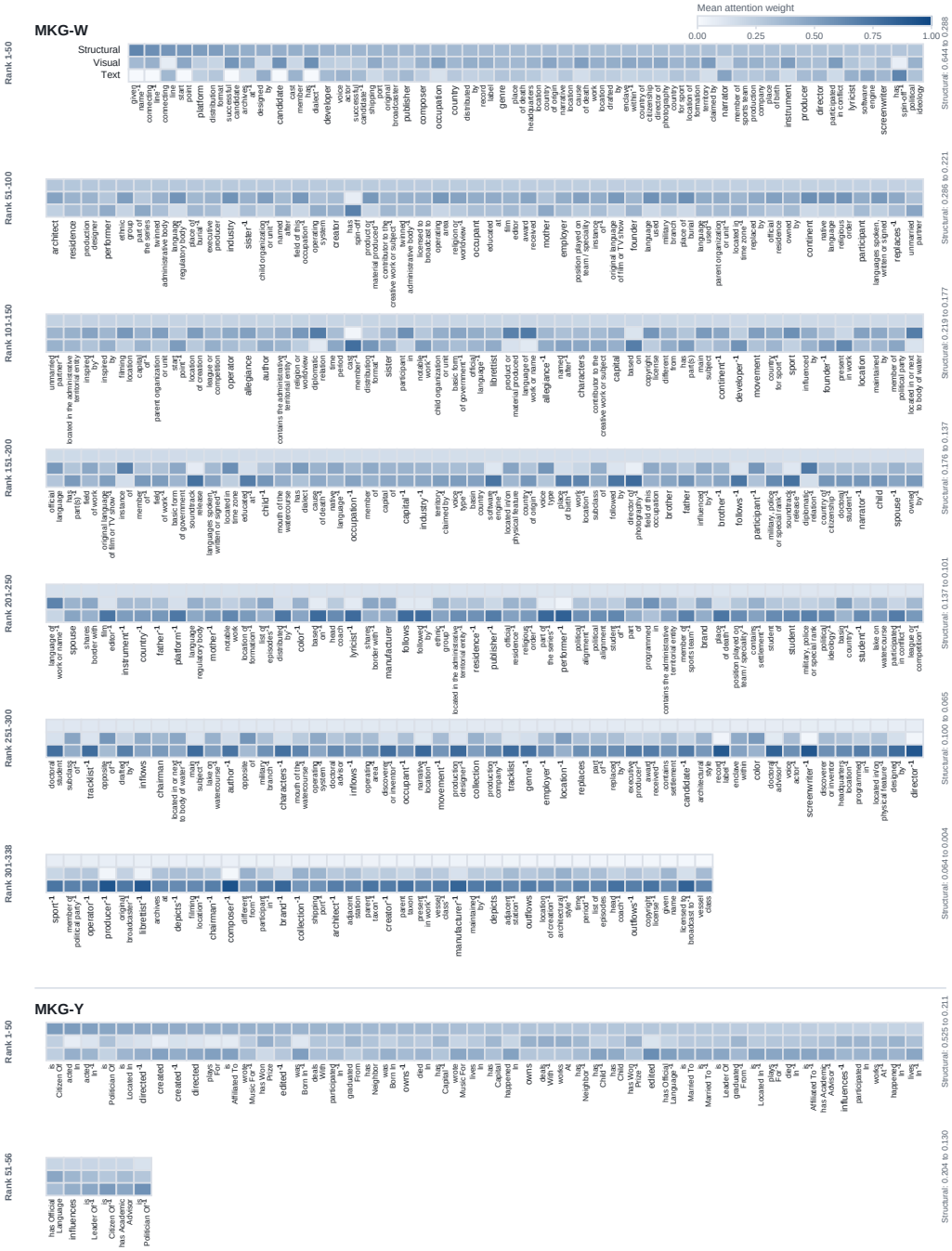


Figure 9: Modality-specific weight allocation for each relation on the MKG-W and MKG-Y dataset. Results are averaged over multiple runs, and all relations are sorted by their structural weights. r^{-1} denotes the inverse direction of relation r.