

ARTKGC: Leakage-Controlled Multimodal Knowledge Graph Completion for Artwork Explanation

Xuan Zhou¹, Xuefeng Wei¹, Yuta Kato², Shintaro Ozaki¹, Kazuki Hayashi¹,
Yusuke Sakai¹, Hidetaka Kamigaito¹, Katsuhiko Hayashi², Taro Watanabe¹

¹Nara Institute of Science and Technology (NAIST), Japan

²The University of Tokyo, Japan

{xuan.zhou.xc7,xuefeng.wei.yb1,ozaki.shintaro.ou6}@naist.ac.jp

{hayashi.kazuki.hl4,sakai.yusuke.sr9,kamigaito.h,taro}@is.naist.jp

{ukato6209,katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp

Abstract

Explaining an artwork requires knowledge beyond visual recognition, including facts about its creator, subject, material, style, collection, and historical context. Knowledge graphs can provide such evidence, but artwork KGs are incomplete, especially for newly encountered artworks. We introduce ARTKGC, a leakage-controlled multimodal benchmark and evaluation protocol for artwork knowledge graph completion. ARTKGC links artworks to Wikidata facts, Wikipedia descriptions, and visual evidence while controlling ambiguous mappings, graph leakage, and descriptions that directly reveal held-out answers. We show that graph-only models are weak for unseen artworks, raw text can inflate evaluation through answer leakage, and visual evidence is strongly relation-dependent. As an interpretable diagnostic baseline, we use a validation-calibrated reranker that combines structural, sentence-redacted textual, and visual evidence. Experiments show improved completion in the primary inductive-artwork setting and modest downstream evidence gains for artwork QA and explanation diagnostics.

1 Introduction

Explaining an artwork is a knowledge-intensive language task. A useful explanation may need to identify the creator, depicted subject, material, style or movement, collection, and historical context of the work. Recent artwork explanation benchmarks show that large vision-language models still struggle to connect images with the background knowledge needed for rich explanations (Hayashi et al., 2024). Knowledge graphs provide explicit structured evidence for knowledge-grounded language tasks, while retrieval-augmented generation demonstrates the value of external evidence for knowledge-intensive generation (Lewis et al., 2020). However, artwork KGs are often incomplete. This is especially problematic for newly

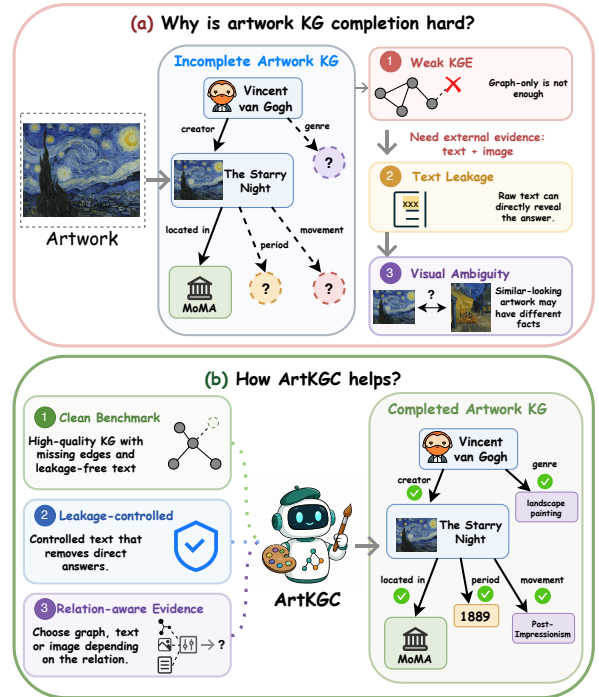


Figure 1: Motivation and overview. Artwork KG completion is challenging because graph-only models are weak for unseen artworks, raw descriptions can leak held-out facts, and visual evidence is relation-dependent. ARTKGC addresses these issues with a clean benchmark, leakage control, and relation-aware multimodal evidence selection.

encountered artworks whose graph neighborhoods are sparse or unseen during training.

Artwork KG completion is not a standard link prediction problem. Graph-only knowledge graph embedding (KGE) models such as TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), and RotatE (Sun et al., 2019) rely mainly on observed graph structure, which can be weak when completing held-out artwork entities. Text and images are natural external evidence sources, but both introduce risks.

Wikipedia descriptions may state the very facts that a model is asked to predict, causing evaluation leakage and echoing broader concerns about memorization in text-enhanced KGC (Sakai et al., 2024). Visual representations from vision-language models such as CLIP (Radford et al., 2021) provide a natural source of external image evidence, but similar-looking artworks may differ in creator, subject, collection, or historical context. Thus, artwork KGC requires clean evaluation, leakage-controlled text, and relation-aware multimodal evidence.

We introduce ARTKGC,¹ a leakage-controlled benchmark and evaluation protocol for artwork knowledge graph completion. ARTKGC is built from ExpArt artwork examples (Hayashi et al., 2024), Wikidata facts (Vrandečić and Krötzsch, 2014), Wikipedia descriptions, and artwork/entity images, and focuses on 16 explanation-oriented relations such as *creator*, *depicts*, *made from material*, *genre*, *movement*, and *collection*. The benchmark contains 40,938 clean triples over 13,335 entities, with both transductive and inductive-artwork splits. We show that standard KGE models are weak when completing held-out artwork entities, while a simple relation-frequency prior is a strong baseline. Sentence-redacted text improves reranking over this prior, and our calibrated multimodal reranker further incorporates visual evidence through validation-tuned, relation-aware fusion. Finally, we evaluate completed triples in a controlled artwork QA / KG-RAG diagnostic, showing modest evidence-use gains while leaving a large oracle gap. Our contributions are:

- We introduce ARTKGC, a leakage-controlled artwork KGC benchmark and evaluation protocol with complementary transductive-clean and inductive-artwork splits.
- We systematically quantify direct-answer text leakage and study structural, sentence-redacted textual, and visual evidence using a validation-calibrated, interpretable multimodal reranker as a diagnostic baseline.
- We characterize when multimodal evidence helps through relation-, frequency-, and coverage-conditioned analyses, and evaluate completed triples in a controlled artwork QA / KG-RAG evidence-use diagnostic.

2 Related Work

Artwork vision-language understanding. Artwork understanding requires connecting visual content with knowledge about style, authorship, subject matter, and cultural context. Prior datasets study this connection through artwork comments and metadata (Garcia and Vogiatzis, 2018), visual and contextual sentences for art-domain retrieval (Stefanini et al., 2019), affective descriptions of visual art (Achlioptas et al., 2021), and automatic art-style recognition (Lecoutre et al., 2017). Broader reviews survey AI methods for art analysis and creation (Cetinic and She, 2022). Recent work also studies cross-lingual artwork explanation and reports weaker LLM performance in non-English languages (Ozaki et al., 2025). These works show that artwork understanding goes beyond object recognition, motivating knowledge-rich evaluation for artwork-centered language tasks. Public cultural resources can also exhibit substantial representational skews: prior work documents geographic and institutional imbalance in aggregated digitized art collections and examines the robustness of bias measurements in Wikidata (Kizhner et al., 2021; Das et al., 2023).

Textual and multimodal knowledge graph completion. Knowledge graph completion has been extended with textual descriptions, pretrained language models, and visual features. Description-based KGC methods such as DKRL encode entity descriptions to improve entity representations (Xie et al., 2016). KG-BERT formulates triples as text sequences for language-model-based scoring (Yao et al., 2019), while SimKGC uses contrastive learning and hard negatives for text-based KGC (Wang et al., 2022). Multimodal KGC further incorporates visual evidence. IKRL introduces image-embodied knowledge representation learning (Xie et al., 2017); MoSE separates and ensembles modality-specific signals (Zhao et al., 2022); AdaMKGC models adaptive modality interactions (Jian et al., 2025); and NativE studies multimodal KGC under noisy and incomplete real-world modality conditions (Zhang et al., 2024). These works motivate the use of text and images for KG completion, while also showing that modality availability and modality usefulness vary across entities and relations.

Evaluation leakage and evidence use. KGC evaluation is sensitive to dataset construction, fil-

¹<https://github.com/xuanZhou111/ARTKGC>

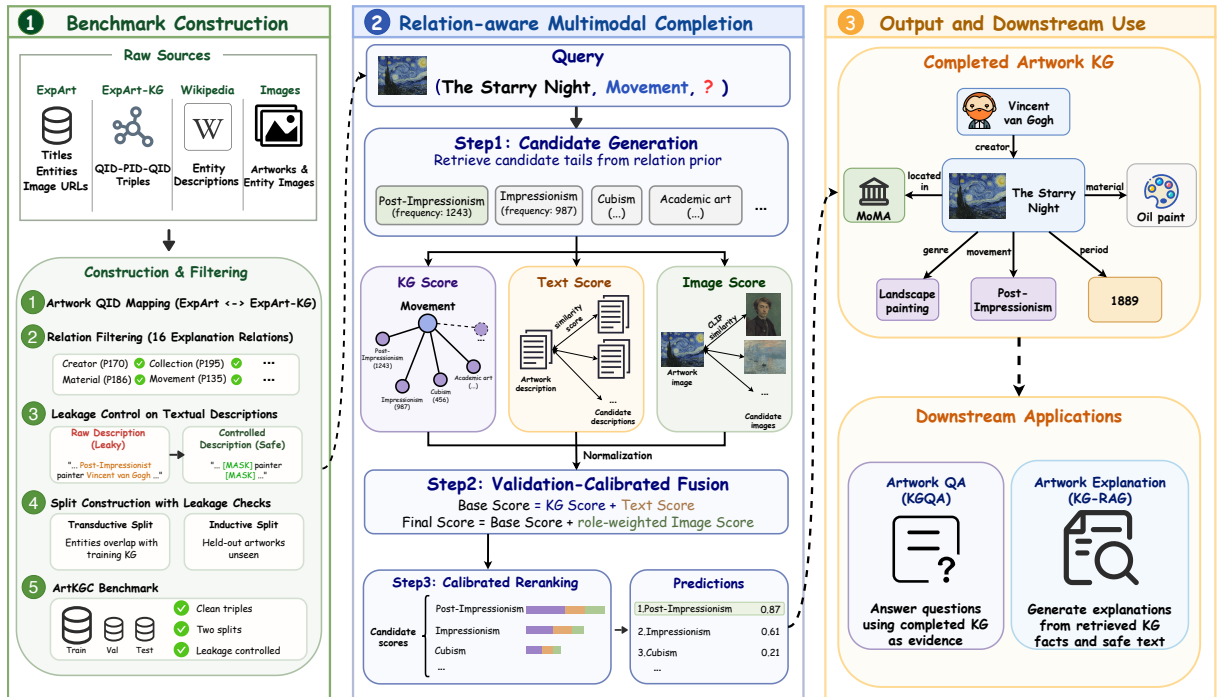


Figure 2: Overview of the ARTKGC pipeline. We construct a clean benchmark from ExpArt, ExpArt-KG, Wikipedia descriptions, and cached visual evidence through QID mapping, relation filtering, leakage control, and split construction. Given an artwork-relation query, ARTKGC retrieves candidate tails from a relation prior, scores them using KG statistics, sentence-redacted text, and image evidence, and applies validation-calibrated two-stage fusion before reranking. The resulting completed triples are evaluated as controlled evidence for downstream artwork QA and KG-RAG explanation diagnostics.

tering protocols, and artifacts in train/test splits. Prior work has shown that KGC conclusions are sensitive to evaluation protocols as well as training and hyperparameter choices (Sun et al., 2020; Ruffinelli et al., 2020). Text-enhanced KGC raises an additional concern: pretrained language models can recall factual knowledge from pretraining rather than infer unseen links, and textual entity information can provide strong shortcuts (Petroni et al., 2019; Sakai et al., 2024). In our setting, descriptions can additionally state held-out answers directly. Beyond link prediction, retrieval-based QA and evidence aggregation methods study how retrieved passages or facts are supplied to readers or generators (Guu et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021). This line of work motivates evaluating not only whether a model ranks a held-out entity highly, but also whether the predicted evidence can support downstream question answering or explanation under controlled conditions.

Inductive KGC and benchmark construction. Inductive KGC evaluates generalization to entities that are unseen during training. Recent bench-

mark analysis further shows that common inductive benchmark construction procedures can introduce structural shortcuts, motivating careful auditing of inductive evaluation protocols (Shomer et al., 2025). ARTKGC shares the unseen-entity motivation but additionally controls inverse and near-duplicate graph leakage, direct-answer textual leakage, and the external evidence supplied to multimodal methods. These controls are complementary to prior work on inductive KGC and are designed around newly encountered artwork entities.

3 ARTKGC: Task and Benchmark

The previous sections motivate a need for artwork knowledge completion under controlled evidence: graph structure alone can be insufficient, raw descriptions may leak answers, and visual evidence is relation-dependent. Before introducing our reranking model, we first define the evaluation setting that exposes these issues. ARTKGC is a clean completion benchmark built from observed artwork facts, not a fully complete artwork knowledge graph. A system receives an incomplete artwork KG and

is evaluated on whether it can recover held-out explanation-oriented triples without using leaked answers.

3.1 Task Definition

We formulate artwork knowledge graph completion as a tail prediction. Each fact is represented as a triple (h, r, t) , where h is a head entity, r is a relation, and t is a tail entity. In our main setting, h is an artwork entity and r is an explanation-oriented relation, such as *creator*, *depicts*, *made from material*, *genre*, *movement*, or *collection*. Given a query $(h, r, ?)$, the model ranks candidate tail entities and is evaluated by whether the gold tail is ranked highly.

This formulation matches our motivation. An artwork explanation system may need to know who created an artwork, what it depicts, what material it uses, or where it is held. These facts can later be verbalized by a QA or explanation system, but we first evaluate them at the structured triple level.

3.2 Evidence Sources

We use the artwork KG from ExpArt-KG (Kato et al., 2026) as the starting KG source, and convert it into a leakage-controlled KGC benchmark with train/validation/test splits, sentence-redacted descriptions, and artwork/entity images. ExpArt provides artwork titles, linked or mentioned entities, and artwork image URLs (Hayashi et al., 2024). These strings are mapped to Wikidata QIDs when possible. Wikidata provides QID–PID–QID triples and relation labels (Vrandečić and Krötzsch, 2014). Wikipedia provides entity descriptions and sitelinks. No machine translation is applied in the evaluated pipeline; textual evidence uses the available public entity labels and descriptions. We also build an image cache that contains artwork images from ExpArt and available entity images from Wikidata or Wikipedia. Thus, artwork images describe the head artwork, while entity images provide visual evidence for candidate or related entities.

3.3 Benchmark Construction

We construct the benchmark in three stages. First, we map artwork titles and linked entities to Wikidata QIDs. To prioritize precision, ambiguous artwork-title mappings and invalid namespace-like rows are excluded from the clean benchmark. Second, we filter Wikidata triples to 16 explanation-oriented relations. We remove relations that are

Item	Value
Triples	40,938
Entities	13,335
Relations	16
Splits	Transductive-clean; inductive-artwork
Transductive train/valid/test	32,752 / 4,093 / 4,093
Inductive train/valid/test	32,706±247 / 4,079±144 / 4,153±134
Seeds	13, 21, 42
Text	Sentence-redacted
Leakage checks	0 inverse; 0 near-duplicate
Inductive rule	designated held-out artwork entities unseen in train

Table 1: ARTKGC-clean benchmark statistics. Transductive split sizes are identical across seeds; inductive-artwork sizes are mean $\pm$ standard deviation over seeds 13, 21, and 42. The inductive rule refers to designated held-out artwork entities.

mostly metadata, external identifiers, or otherwise not useful for artwork explanation. Third, we attach textual and visual evidence to entities when available. These evidence sources are not used to change the gold triples; they are used only as model inputs and downstream evidence.

The resulting clean benchmark contains 40,938 triples over 13,335 entities and 16 relations. Table 1 summarizes the main statistics.

3.4 Leakage-Controlled Text

Text descriptions are useful but can invalidate evaluation. For example, if the held-out triple is $(h, \textit{movement}, \textit{Post-Impressionism})$, a raw Wikipedia description may directly state that the artwork is Post-Impressionist. A model using that sentence would not complete the graph; it would simply read the answer.

We therefore define leakage-controlled descriptions. Our main setting uses a fixed sentence-redacted resource constructed before retrieval and reranking. For each entity, we collect the labels and aliases of counterpart entities appearing with it in the validation and test triples and remove any sentence containing one of these mentions. The resulting per-entity description is fixed and reused across all text-based and multimodal methods rather than being redacted dynamically for each query. Across the 5,340 descriptions with raw text, sentence segmentation yields 40,404 sentence instances; sentence redaction removes 7,730 instances (19.13%) and affects 2,331 descriptions (43.65%). The redaction process records 5,799 unique removed description sentences; further redaction statistics are pro-

vided in Appendix A.7. Raw descriptions and strict redaction are retained as diagnostic settings. Strict redaction first applies the same sentence-level removal as the main protocol and then replaces any remaining counterpart label or alias mentions with a redaction token. All main text-based and multimodal results use sentence-redacted descriptions.

3.5 Transductive and Inductive Splits

ARTKGC provides two split families, each with its own train, validation, and test sets. In the *transductive-clean* split, triples are partitioned into train, validation, and test. Entities may appear across splits, but the target test triples are held out. This corresponds to a standard KGC setting.

In the *inductive-artwork* split, we construct validation and test sets around designated held-out artwork entities. These held-out artwork entities have zero overlap with training entities. Because some retained relations can connect multiple artwork entities, a small number of evaluation triples may still have heads that also occur as training artwork heads; the formal inductive guarantee applies to the designated held-out artwork entity lists. This setting reflects our intended use case: completing facts involving artworks that are newly encountered or sparsely connected.

For both split families, we remove inverse and near-duplicate leakage. For example, if a held-out triple uses a part-whole relation, its inverse should not appear in training. We also prevent near-duplicate relations, such as closely related subject relations, from leaking the same head-tail pair across splits. For each split and random seed, hyperparameters and calibration weights are tuned on the corresponding validation set and then fixed for its test evaluation, and final results are reported on the corresponding test sets. Table 1 reports the train/validation/test sizes for both split families; seed-level split counts and overlap diagnostics are provided in Appendix A.6.

4 Relation-Aware Multimodal Completion

Having defined the benchmark and its leakage-controlled evidence, we now describe the completion model used to rank candidate tails. Our goal is not to build a large end-to-end multimodal model, but to test whether graph, text, and image evidence can be combined in a controlled and interpretable way. The method follows the pipeline in Figure 2:

it first retrieves plausible candidate tails from a structural prior, then scores each candidate using three evidence sources, and finally reranks candidates with validation-calibrated two-stage fusion.

4.1 Candidate Generation

For a query $q = (h, r, ?)$, directly ranking all entities is expensive and noisy. We therefore first retrieve a small candidate set using a relation-conditioned frequency prior computed only from the training KG. Let $\mathcal{G}_{\text{train}}$ denote the training triples of the current split family, and let (h_i, r_i, t_i) range over these triples. For each relation r and candidate tail t , we compute

$$c_r(t) = \sum_{(h_i, r_i, t_i) \in \mathcal{G}_{\text{train}}} \mathbf{1}[r_i = r \wedge t_i = t], \quad (1)$$

where $\mathbf{1}[\cdot]$ is an indicator function. Thus, $c_r(t)$ counts how often t appears as the tail of relation r in the training KG.

The top- K entities under this count form the candidate set C_q for query q . This prior is intentionally simple and does not use the head artwork h ; it captures relation-specific tail-frequency biases in the training graph. Such biases are common in artwork relations, where certain materials, genres, movements, or institutions appear frequently. We use the same relation prior both as a baseline and as the candidate generator for reranking. Head-specific evidence is introduced later through redacted text and image scores.

4.2 Evidence Scores

For each query $q = (h, r, ?)$ and candidate tail $t \in C_q$, we compute three evidence scores: a KG prior score, a text score, and an image score.

KG prior. The KG prior is derived from the training frequency $c_r(t)$ defined above. Standalone relation-prior ranking uses the raw count, while the calibrated reranker uses a log-scaled prior:

$$a_{\text{kg}}(q, t) = \log(1 + c_r(t)). \quad (2)$$

Text score. Let T_q be the artwork-side query text, constructed from the artwork label, relation label, and sentence-redacted artwork description. Let T_t be the candidate-side text, constructed from the tail label and its description. We encode both texts with a sentence encoder and L2-normalize the resulting vectors:

$$\begin{aligned} \mathbf{u}_q &= \text{norm}(E_{\text{txt}}(T_q)), \\ \mathbf{u}_t &= \text{norm}(E_{\text{txt}}(T_t)). \end{aligned} \quad (3)$$

The raw text score is:

$$a_{\text{txt}}(q, t) = \mathbf{u}_q^\top \mathbf{u}_t. \quad (4)$$

Because the vectors are L2-normalized, this is cosine similarity. This component performs embedding-based matching only; it does not translate the input or generate free-form textual explanations.

Image score. Let I_h be the artwork image and I_t be the candidate entity image when available. We encode both images and L2-normalize the resulting vectors:

$$\begin{aligned} \mathbf{v}_h &= \text{norm}(E_{\text{img}}(I_h)), \\ \mathbf{v}_t &= \text{norm}(E_{\text{img}}(I_t)). \end{aligned} \quad (5)$$

The raw image score is:

$$a_{\text{img}}(q, t) = \mathbf{v}_h^\top \mathbf{v}_t. \quad (6)$$

This is equivalent to CLIP cosine similarity. If image evidence is unavailable, the reranker uses the available KG and text scores.

4.3 Score Normalization

The raw scores have different scales: the KG prior is count-based, while text and image scores are cosine similarities. We therefore normalize scores within the candidate set C_q of each query.

For the KG prior, we divide the log-count by the largest log-count in the candidate set:

$$p(q, t) = \frac{a_{\text{kg}}(q, t)}{\max_{u \in C_q} a_{\text{kg}}(q, u) + \epsilon}. \quad (7)$$

where ϵ is a small constant for numerical stability.

For text and image scores, we apply candidate-set min-max normalization:

$$x(q, t) = \text{MM}_q(a_{\text{txt}}, t), \quad (8)$$

$$y(q, t) = \text{MM}_q(a_{\text{img}}, t). \quad (9)$$

Here MM_q denotes min-max normalization over the candidates in C_q . Thus, $p(q, t)$, $x(q, t)$, and $y(q, t)$ are the normalized KG, text, and image scores used for fusion. The explicit definition of MM_q is given in Appendix A.1.

4.4 Two-Stage Validation-Calibrated Fusion

The reranker combines the normalized scores in two stages. First, it mixes the KG prior and dense text score:

$$b(q, t) = (1 - \alpha_{\text{txt}}) p(q, t) + \alpha_{\text{txt}} x(q, t). \quad (10)$$

The text mixing weight α_{txt} and the candidate-pool size K are selected on the validation set.

Second, the reranker incorporates image evidence. Since image usefulness differs across relation types, we use a relation-role-specific image weight β_r . The final score is:

$$s(q, t) = (1 - \beta_r) b(q, t) + \beta_r y(q, t). \quad (11)$$

For relations where image evidence is unreliable, β_r is set to zero. When image evidence is unavailable, we fall back to the base score. Candidates are reranked by $s(q, t)$, and the top-ranked candidate is used as the predicted tail. For each split and random seed, all candidate-pool and fusion settings are selected using the corresponding validation set and then fixed for test evaluation. Appendix A.1 provides additional normalization and calibration details, including the relation-role mapping.

5 Artwork KGC Experiments

We evaluate whether the benchmark exposes three central phenomena: graph-only link prediction is insufficient for held-out artworks, raw descriptions can inflate text-enhanced evaluation through direct answer leakage, and the utility of redacted textual and visual evidence depends on calibration, relation type, and evidence availability. We organize the experiments around these questions and use the calibrated reranker as an interpretable diagnostic baseline.

5.1 Experimental Setup

We evaluate tail prediction on the two split families defined in Section 3. The inductive-artwork split is our primary setting because it is constructed around designated artwork entities held out from the training graph. The transductive-clean split provides a standard KGC comparison in which entities may appear across splits but target triples are held out. For each split and random seed, candidate size and calibration parameters are selected on the corresponding validation set; all reported numbers are test results.

Following standard KGC and multimodal KGC evaluation (Bordes et al., 2013; Sun et al., 2020; Zhao et al., 2022; Zhang et al., 2024), we report mean reciprocal rank (MRR) and Hits@10. MRR averages the reciprocal rank of the gold tail across test queries, so higher values indicate that the correct entity is ranked earlier. Hits@10 is the fraction of queries for which the gold tail appears in

the top 10 ranked candidates. Since ARTKGC is formulated as $(h, r, ?)$ completion, all metrics are reported for tail prediction.

Table 2 groups methods by family. Non-neural priors test simple frequency and degree signals. KGE baselines test standard graph-only embedding models. MKGC diagnostics test adapted multimodal KGC baselines. Text methods test whether sentence-redacted descriptions help as standalone retrieval or reranking evidence. Additional adapted MKGC diagnostics indicate that MoSE and AdaMKGC-adapted are near random under our clean adaptation, while NativeE is the strongest adapted MKGC baseline; we therefore include NativeE tail completion as a diagnostic row in the main table. Further baseline details are provided in Appendix A.2. The final row is our calibrated multimodal reranker. Implementation details, hyperparameter grids, model sizes, and compute information are provided in Appendix A.5.

5.2 Structure-Only Models Are Not Enough

Table 2 shows that standard graph embedding models are weak in the primary inductive-artwork setting. Among TransE, DistMult, ComplEx, and RotatE, the strongest inductive KGE model is TransE, with only 0.043 MRR and 0.110 Hits@10. In contrast, the simple relation prior reaches 0.235 MRR and 0.393 Hits@10. This means the task cannot be evaluated only against random or KGE baselines: a strong relation-frequency prior already captures many easy cases and must be treated as a serious baseline. The strength of this prior is relation-dependent and follows the concentration of training tails. For *material*, the most frequent tail accounts for about 61% of training triples and prior Recall@10 is about 88-90%; for *genre*, the five most frequent tails account for about 77%. In contrast, *creator* has a much more diffuse tail distribution, with top-1 mass below 2%, top-5 mass below 8%, and prior Recall@10 of only 13-18%. Thus, the head-agnostic prior is competitive mainly for relations with strongly concentrated tail distributions. Further details on relation-frequency concentration are provided in Appendix A.11.

The adapted multimodal KGC diagnostics further clarify the setting. Adapted feature-DistMult uses text and image features, but remains below relation prior on both splits. NativeE tail diagnostic is strong in the transductive-clean setting, where entities are observed during training. However, in the primary inductive-artwork setting, our cali-

brated multimodal reranker obtains the best MRR and Hits@10. We therefore do not claim universal dominance across all KGC settings; our main claim is leakage-controlled, relation-aware completion for held-out artwork entities.

5.3 Redacted Text Helps as Reranking Evidence

To validate the leakage-control protocol, we compare raw, sentence-redacted, and strict-redacted descriptions while holding the reranker configuration fixed across text settings. Raw text increases reranker MRR from 0.2699 to 0.2768 on inductive-artwork and from 0.3028 to 0.3111 on transductive-clean, whereas strict redaction is nearly unchanged from sentence redaction. These results show that direct-answer mentions measurably inflate evaluation, while sentence-level removal captures most of this effect. Further details of the redaction ablation are provided in Appendix A.8. Text alone is weak after leakage control. BM25 and dense retrieval perform poorly as standalone methods, with MRR below 0.04 on both splits. This is expected: once direct answer mentions are removed, descriptions no longer provide an easy lookup signal. However, the Text Evidence Reranker improves over relation prior on both splits, from 0.235 to 0.257 MRR in inductive-artwork and from 0.255 to 0.286 MRR in transductive-clean. This shows that sentence-redacted text is still useful, but mainly as a reranking signal over structurally plausible candidates.

5.4 Calibration Matters

The calibrated multimodal reranker improves over both relation prior and the text evidence reranker in MRR. Table 3 isolates the multimodal part of the method. CLIP image reranking alone is weaker than the calibrated model, and fixed fusion is also below validation-calibrated fusion. This supports the role of calibration rather than simply adding images. The multimodal gain is not confined to richly described artworks. The calibrated multimodal reranker remains above both the relation prior and the Text Evidence Reranker in every description-availability bucket, including queries with unavailable or empty artwork-side descriptions. Further details on description-availability buckets and query counts are provided in Appendix A.9. Image availability shows a complementary pattern. Visual gains are concentrated among queries with a cached head image. When no cached head image is available, the image weight is inactive and the

Family	Method	Inductive-artwork		Transductive-clean	
		MRR	H@10	MRR	H@10
Prior	Random	0.001	0.001	0.001	0.001
Prior	Degree prior	0.124	0.226	0.132	0.244
Prior	Relation prior	0.235	0.393	0.255	0.429
KGE	TransE	0.043	0.110	0.066	0.174
KGE	DistMult	0.000	0.000	0.138	0.230
KGE	ComplEx	0.001	0.001	0.002	0.003
KGE	RotatE	0.000	0.000	0.114	0.183
MKGC	Adapted feature-DistMult	0.097	0.222	0.126	0.244
MKGC	NativeE tail diagnostic [†]	0.236	0.314	0.396	0.598
Text	BM25 text	0.038	0.086	0.024	0.061
Text	Dense text	0.025	0.068	0.015	0.046
Rerank	Text Evidence Reranker	0.257	0.414	0.286	0.463
Ours	Calibrated multimodal reranker (ours)	0.270	0.417	0.303	0.469

Table 2: Main artwork KGC results. The inductive-artwork split is our primary setting because it is constructed around designated artwork entities held out from training. All text-based methods use sentence-redacted descriptions. [†]NativeE tail diagnostic evaluates side-specific tail completion; it is strongest in the transductive setting, while our calibrated reranker is strongest in the primary inductive-artwork setting.

Variant	Ind. MRR	Trans. MRR
Text Evidence Reranker	0.257	0.286
CLIP image rerank	0.255	0.277
Fixed image-text fusion	0.264	0.290
Calibrated multimodal rerank	0.270	0.303

Table 3: Compact multimodal ablation. Calibration improves over text-only reranking, image-only reranking, and fixed image-text fusion.

model falls back to its validation-selected structural and textual score. Further details on image availability are provided in Appendix A.10.

The main improvement is in MRR. We use paired bootstrap resampling following Koehn (2004). The calibrated reranker improves over the text reranker by +0.013 MRR on inductive-artwork and +0.020 MRR on transductive-clean, with 95% confidence intervals that do not cross zero on either split. It also yields significant MRR improvements over the relation prior on both splits. Hits@10 is more mixed: the transductive Hits@10 gain over text reranking is significant, while the inductive Hits@10 gain is small and not statistically supported. We therefore frame the main KGC improvement as an MRR gain. Further statistical details are provided in Appendix A.2.

5.5 Which Relations Benefit?

Average performance hides an important pattern: multimodal evidence is not uniformly useful. Figure 3 shows the MRR gain of the calibrated mul-

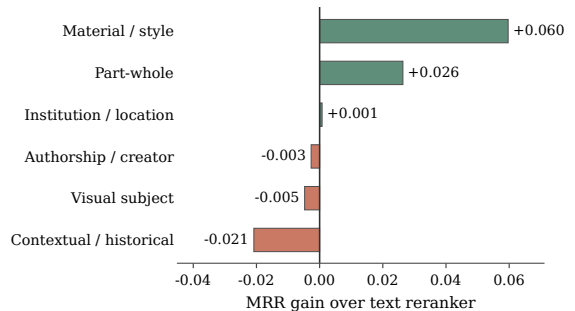


Figure 3: Relation-group MRR gains of the calibrated multimodal reranker over the Text Evidence Reranker. Multimodal evidence is most useful for material/style and part-whole relations, while visual-subject and contextual-historical relations remain unstable.

timodal reranker over the text evidence reranker by relation group. The largest gains are for material/style and part-whole relations. Institution/location is slightly positive but small. Authorship, visual-subject, and contextual-historical relations are mixed or negative.

This analysis explains why relation-aware calibration is needed. Visual cues can help with coarse material, style, or part-whole evidence, but they can also retrieve related yet incorrect candidates for visually ambiguous relations such as *depicts* or *main subject*. At the group level, material/style gains +0.060 MRR and part-whole gains +0.026 MRR, while visual-subject and contextual-historical groups decrease relative to text reranking. Further relation-group details are provided in Ap-

pendix A.3. Thus, we avoid claiming that image evidence universally improves artwork KGC; the evidence is useful, but relation-dependent. Aligned query-level cases illustrate both sides of this behavior. For *The Death of Marat*, visual fusion moves the gold *drapery* tail for *depicts* from rank 64 to 9. In contrast, for *The Cliff Walk at Pourville*, the gold *Art Institute of Chicago* tail for *collection* falls from rank 8 to 56, consistent with collection membership being provenance-dependent rather than visually recoverable. Further qualitative examples are provided in Appendix A.12.

The KGC experiments show that leakage-controlled text and calibrated visual evidence improve completion over strong structural baselines in the main inductive setting. The next section asks whether the predicted triples are useful as evidence for downstream artwork QA and explanation.

6 Downstream Evidence-Use Diagnostic

The KGC experiments evaluate link prediction; we next test whether predicted triples are useful as evidence for artwork QA and KG-RAG-style explanation. This is a controlled diagnostic, not a claim that production retrieval is solved. The evaluation asks for structured tail entities through fixed relation-specific question templates; it does not evaluate free-form artwork appreciation or long-form explanation generation.

6.1 Setup

We convert each held-out triple (h, r, t) into an artwork QA question asking for the missing tail t . Each method supplies evidence under the same budget. We compare no KG evidence, observed training KG, random completion, relation-prior completion, calibrated multimodal completion, and oracle evidence; non-oracle settings never use the held-out gold triple.

We report evidence recall, Copy EM, Reader EM, and unknown rate. Evidence recall checks whether the supplied evidence contains the gold tail. Copy EM evaluates a constrained evidence-copy reader, Reader EM averages normalized exact match across downstream readers, and unknown rate measures abstention. Appendix A.4 gives construction details and split-level results.

6.2 Completed KG Improves Evidence Use

Table 4 shows that the observed training KG provides little help for held-out questions, and random

Evidence	Rec.	Copy	Reader	Unk.
No KG	0.000	0.000	0.056	0.778
Train KG	0.000	0.000	0.075	0.570
Random	0.000	0.000	0.040	0.440
Relation prior	0.226	0.153	0.153	0.277
Multimodal	0.274	0.207	0.183	0.253
Oracle	1.000	1.000	0.783	0.087

Table 4: Controlled downstream evidence-use diagnostic. Rec. is evidence recall, Copy is constrained evidence-copy exact match, Reader is multi-reader exact match, and Unk. is unknown rate. Bold marks the best non-oracle result.

completion also fails. Relation-prior completion gives the main downstream jump, raising evidence recall to 0.226 and Copy EM to 0.153. Calibrated multimodal completion provides a further gain, increasing evidence recall from 0.226 to 0.274, Copy EM from 0.153 to 0.207, and Reader EM from 0.153 to 0.183, while reducing the unknown rate from 0.277 to 0.253. The same improvement is observed on both split families. Further split-level details are provided in Appendix A.4.

6.3 The Oracle Gap Remains Large

The oracle setting shows that the downstream task remains far from solved: supplying the held-out gold triple yields 1.000 evidence recall and Copy EM. To characterize the remaining gap, we decompose non-oracle outcomes into correct answers, cases where the gold evidence is present but is not copied correctly, and cases where the gold tail is absent from the supplied evidence. With the final multimodal completions, the correct-answer rate increases from 14.52% to 19.56% on inductive-artwork and from 16.09% to 21.92% on transductive-clean relative to the relation prior. The gold-absent rate decreases from 78.79% to 74.14% and from 75.98% to 71.09%, respectively. Thus, the dominant remaining source of error is that the gold tail is not included in the evidence supplied under the fixed evidence budget.

7 Conclusion

We introduced ARTKGC, a leakage-controlled benchmark and evaluation protocol for artwork KGC. Results show that graph-only KGE is weak for held-out artworks, raw descriptions inflate evaluation, and calibrated multimodal evidence yields relation- and coverage-dependent gains, with useful but limited downstream improvements.

Limitations

ARTKGC is not a complete artwork knowledge graph. ARTKGC is a benchmark built from observed artwork facts, not an exhaustive representation of artwork knowledge. Its triples are derived from available sources such as Wikidata, Wikipedia, and ExpArt-linked artworks. As a result, missing facts in ARTKGC should not be interpreted as false facts, and predicted triples should be treated as candidate completions rather than verified art-historical truth.

Coverage is limited by public resources. The benchmark inherits the coverage of publicly documented artworks and entities. Well-known artworks, Western collections, and entities with rich Wikidata or Wikipedia coverage are more likely to have triples, descriptions, and images. Less digitized collections, non-Western artworks, anonymous works, and culturally specific concepts may be underrepresented. This affects both benchmark coverage and model behavior, and means that ARTKGC should be interpreted as an artwork-domain benchmark rather than a general-purpose KGC benchmark. No machine translation is applied in the evaluated pipeline, but public descriptions may still provide uneven coverage of multilingual and culturally specific terminology. Language- and source-stratified evaluation remains future work.

The relation set is explanation-oriented but not exhaustive. We focus on 16 relations that are useful for artwork explanation, such as creator, depicts, material, genre, movement, and collection. This makes the benchmark focused and interpretable, but it excludes many other forms of artwork knowledge, including provenance details, conservation history, exhibition history, ownership chains, iconographic interpretation, and expert-curated contextual knowledge.

Leakage control reduces but does not eliminate all textual clues. Our sentence-redacted descriptions remove direct mentions of held-out entities, but they cannot remove every indirect clue. A description may still contain temporal, stylistic, geographic, or contextual information that helps infer the answer. Conversely, redaction can also remove useful context and make text evidence less informative than it would be in realistic downstream use. Therefore, the redacted setting should be understood as a controlled evaluation protocol rather

than a perfect simulation of real-world evidence access.

The multimodal reranker is deliberately simple. Our method uses validation-calibrated two-stage fusion rather than an end-to-end trained multimodal architecture. This design makes the role of structural, textual, and visual evidence easier to interpret, but it may limit performance compared with larger or more expressive models. Our goal is to establish a clean benchmark and analyze relation-dependent evidence utility, not to claim that this reranker is the final architecture for artwork KGC.

Visual evidence is relation-dependent and sometimes unstable. The results show that image evidence helps some relation groups, especially material/style and part-whole relations, but it is not uniformly beneficial. Visual-subject and contextual-historical relations remain challenging because visually similar artworks may correspond to different subjects, events, collections, or cultural contexts. This limits claims about visual evidence: images can help, but they should not be treated as universally reliable.

The downstream evaluation is diagnostic, not production KG-RAG. Our downstream QA / KG-RAG-style experiment is a controlled evidence-use diagnostic. It tests whether completed triples can help readers under fixed evidence budgets, but it does not solve open retrieval, evidence verification, long-form explanation generation, or real deployment. The large oracle gap shows that current completion methods still miss many useful facts and that downstream systems may fail even when strong evidence is available.

Domain specificity. ARTKGC focuses on artworks rather than general-domain KGC. This is intentional: artwork explanation requires a mixture of visual, textual, and structured knowledge, making it a useful testbed for leakage-controlled multimodal completion. However, our empirical conclusions are specific to artwork-centered relations and should not be assumed to transfer directly to other domains. Extending the benchmark construction and relation-aware evidence analysis to broader cultural heritage or general-domain KGs is an important direction for future work.

Ethical Considerations

Cultural and institutional bias. ARTKGC relies on publicly available resources such as Wikidata, Wikipedia, ExpArt-linked artworks, and public image sources. These resources may reflect cultural, linguistic, geographic, and institutional biases. Artworks from well-documented museums, Western art history, and widely represented languages are likely to be better covered than artworks from underrepresented cultures or less digitized collections. Models trained or evaluated on ARTKGC may reproduce these imbalances. A direct cultural audit would require curated cultural, geographic, linguistic, and institutional metadata that are not currently available in ARTKGC. Coverage-stratified evaluation, provenance-aware weighting, source-aware confidence calibration, and culturally grounded multi-evidence explanation therefore remain important directions for future work.

Predicted triples are not verified facts. Knowledge graph completion produces candidate facts, not authoritative art-historical statements. Incorrect predictions about creators, depicted subjects, cultural affiliations, historical context, or collections could mislead users if presented without uncertainty. We recommend that completed triples be displayed with provenance, confidence scores, and clear indications that they are model predictions. In educational, museum, or cultural heritage settings, predicted facts should be reviewed by domain experts before being presented as factual claims.

Risk of hallucinated or oversimplified explanations. If predicted triples are used in downstream explanation systems, erroneous or overly generic triples may lead to hallucinated or misleading explanations. For example, a model may assign a high-frequency artist, movement, or subject to an artwork simply because it is common in the training data. Systems using ARTKGC outputs should distinguish between observed KG facts, model-completed facts, and generated explanatory text.

Resource release, licensing, and attribution. ARTKGC is built from public resources, including ExpArt, Wikidata, Wikipedia, and image sources. To support reproducibility, we publicly release the code and ARTKGC benchmark. For source materials with license or attribution requirements, we provide source identifiers and provenance information and redistribute original content only when permitted by the corresponding licenses.

Use of AI assistants. AI assistants were used primarily for language editing and manuscript proof-reading during manuscript preparation, e.g., the camera-ready phase. All scientific content and final revisions were reviewed and approved by the authors.

Appropriate use. ARTKGC is intended for research on leakage-controlled multimodal KGC and knowledge-grounded artwork explanation. It should not be used as an authoritative database of artwork facts, nor as a replacement for curators, art historians, or cultural heritage professionals. Its most appropriate use is as a benchmark and diagnostic tool for studying how structured, textual, and visual evidence can support more reliable artwork-related language systems.

Personal and sensitive information. ARTKGC is built from public artwork and entity records rather than newly collected private personal data. Some entities may refer to artists, collectors, institutions, or historical figures already present in Wikidata or Wikipedia. We do not infer private attributes, and downstream systems should avoid presenting model-completed triples about living persons or culturally sensitive subjects without verification.

Acknowledgments

This research was supported by the China Scholarship Council (CSC) under grant numbers 202308070070 (Xuan Zhou) and 202308070072 (Xuefeng Wei), and was partly supported by JSPS Kakenhi Grant Number JP24K02993 and JP26H02537.

References

- Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11569–11579.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13*, page 2787–2795, Red Hook, NY, USA. Curran Associates Inc.

Eva Cetinic and James She. 2022. [Understanding and](#)

- creating art with ai: Review and outlook. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(2).
- Paramita Das, Sai Keerthana Karnam, Anirban Panda, Bhanu Prakash Reddy Guda, Soumya Sarkar, and Animesh Mukherjee. 2023. [Diversity matters: Robustness of bias measurements in wikidata](#). In *Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23*, page 208–218, New York, NY, USA. Association for Computing Machinery.
- Noa Garcia and George Vogiatzis. 2018. [How to read paintings: Semantic art understanding with multi-modal retrieval](#). In *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part II*, page 676–691, Berlin, Heidelberg. Springer-Verlag.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML'20*. JMLR.org.
- Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. [Towards artwork explanation in large-scale vision language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Yue Jian, Miao Zhang, Ziyue Qin, Chuyuan Xie, Kui Xiao, Yan Zhang, and Zhifei Li. 2025. [Adaptive modality interaction transformer for multimodal knowledge graph completion](#). *ACM Trans. Knowl. Discov. Data*, 19(8).
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Yuta Kato, Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2026. [ExpArt-KG: Artwork Image Description Generation through Iterative Exploration of Knowledge Graphs](#). ResearchGate. DOI: 10.13140/RG.2.2.29734.08003.
- Inna Kizhner, Melissa Terras, Maxim Rumyantsev, Valentina Khokhlova, Elisaveta Demeshkova, Ivan Rudov, and Julia Afanasieva. 2021. [Digital cultural colonialism: measuring bias in aggregated digitized content held in google arts and culture](#). *Digital Scholarship in the Humanities*, 36(3):607–640.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Adrian Lecoutre, Benjamin Negrevergne, and Florian Yger. 2017. [Recognizing art style automatically in painting with deep learning](#). In *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 327–342, Yonsei University, Seoul, Republic of Korea. PMLR.
- 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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. [Towards cross-lingual explanation of artwork in large-scale vision language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3773–3809, Albuquerque, New Mexico. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. [You can teach an old dog new tricks! on training knowledge graph embeddings](#). In *International Conference on Learning Representations*.
- Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. [Does pre-trained language model actually infer unseen links in knowledge graph completion?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8091–8106, Mexico City, Mexico. Association for Computational Linguistics.
- Harry Shomer, Jay Revolinsky, and Jiliang Tang. 2025. [Towards better benchmark datasets for inductive knowledge graph completion](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5777–5788, New York, NY, USA. Association for Computing Machinery.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. [Art-pedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain](#). In *Image Analysis and Processing – ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II*, page 729–740, Berlin, Heidelberg. Springer-Verlag.
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. [Rotate: Knowledge graph embedding by relational rotation in complex space](#). In *International Conference on Learning Representations*.
- Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. 2020. [A re-evaluation of knowledge graph completion methods](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5516–5522, Online. Association for Computational Linguistics.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. [Complex embeddings for simple link prediction](#). In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. [SimKGC: Simple contrastive knowledge graph completion with pre-trained language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294, Dublin, Ireland. Association for Computational Linguistics.
- Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2659–2665. AAAI Press.
- Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3140–3146. AAAI Press.
- Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. [Embedding entities and relations for learning and inference in knowledge bases](#). *Preprint*, arXiv:1412.6575.
- Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. [Kgbert: Bert for knowledge graph completion](#). *Preprint*, arXiv:1909.03193.
- Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2024. [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 '24, page 91–101, New York, NY, USA. Association for Computing Machinery.
- Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022. [MoSE: Modality split and ensemble for multimodal knowledge graph completion](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10527–10536, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Method Details

Candidate generation. For each split family, relation priors are computed only from the corresponding training triples. For each relation r , we count how often each tail entity t appears with r in the training KG. Candidate tails are retrieved by sorting this raw count. Validation and test triples are not used to compute these counts.

KG prior score. The standalone relation-prior baseline ranks candidates by raw training count $c_r(t)$. In the fusion model, we use the log-scaled score $\log(1 + c_r(t))$ and divide it by the maximum log-scaled count inside the candidate pool.

Text score. The main text score uses dense similarity. The artwork-side text concatenates the artwork label, relation label, and sentence-redacted artwork description. The candidate-side text concatenates the candidate label and description. Both texts are encoded with a sentence encoder, L2-normalized, and scored by dot product. Since the vectors are normalized, the dot product is cosine similarity. BM25 is used as a separate sparse-text baseline in the experiments, not as the main fusion score.

Image score. The image score uses CLIP image features. Artwork images describe the head artwork, while entity images describe candidate

tail entities when available. Features are L2-normalized, and the image score is computed by dot product, equivalent to CLIP cosine similarity. If image features are unavailable, reranking falls back to the available KG and text evidence.

Candidate-set min-max normalization. For a raw score function a , let $a_{\min}$ and $a_{\max}$ be the minimum and maximum of $a(q, u)$ over candidates $u \in C_q$. We define:

$$\text{MM}_q(a, t) = \frac{a(q, t) - a_{\min}}{a_{\max} - a_{\min} + \epsilon}.$$

We use this normalization for dense text and image scores. The KG prior instead uses log-count divided by the maximum log-count in the same candidate set.

Validation-calibrated fusion. Fusion is performed in two stages. First, a validation-tuned α_{txt} combines the KG prior and dense text score. Calibration is performed independently for each split and random seed. Second, image evidence is added with a relation-role-specific weight. Formally, $\beta_r = \alpha_{\text{img}}(\text{role}(r))$, where $\text{role}(r)$ is the visual-evidence role assigned to relation r . Relations with unreliable image evidence use zero image weight. All calibration parameters are selected on validation data and fixed for test evaluation. The visual-evidence role assigned to each retained relation is listed in Table 5. Role assignments are fixed across splits and seeds; only the validation-selected image weight associated with each role can vary across split-seed runs.

Candidate-pool fallback. The relation prior is used both to retrieve the candidate pool and as part of the final fusion score. If a gold tail is outside the candidate pool during evaluation, its rank falls back to the relation-prior rank. Otherwise, the reranker computes fused scores for candidates inside the pool.

A.2 Additional KGC Diagnostics

Adapted MKGC baselines. Table 6 summarizes additional adapted MKGC baselines. Feature-DistMult is non-trivial but below the relation prior and our calibrated reranker. MoSE and AdaMKGC-adapted are near random under our clean adaptation. NativE is the strongest adapted MKGC baseline: its official averaged MRR is competitive, and its side-specific tail diagnostic is especially strong in the transductive split.

PID	Relation	Visual-evidence role
P1028	donated by	low
P1071	location of creation	low
P127	owned by	low
P135	movement	low
P136	genre	medium
P144	based on	low
P170	creator	low
P180	depicts	high
P186	made from material	high
P195	collection	medium
P276	location	medium
P361	part of	mixed
P527	has part(s)	mixed
P547	commemorates	low
P921	main subject	medium
P941	inspired by	low

Table 5: Visual-evidence roles for the 16 retained ARTKGC relations. The role assignment is fixed across splits and seeds and determines which validation-selected image-weight family is used during multimodal reranking. Relations assigned to the low-evidence role use zero image weight.

Baseline	Ind. MRR	Trans. MRR
Feature-DistMult	0.097	0.126
MoSE-CompExMDR	0.002	0.002
NativE-AdvRelRotatE	0.119	0.250
AdaMKGC-adapted	0.001	0.001
NativE tail diagnostic	0.236	0.396

Table 6: Adapted MKGC baselines. The first four rows report official averaged MRR; the last row reports side-specific tail-completion MRR.

Seed-level feature-DistMult. Table 7 reports the split-seed results for the adapted feature-DistMult baseline. The model uses sentence-redacted text embeddings and CLIP image features. The average values are reported in Table 2.

Paired bootstrap. Table 8 reports paired bootstrap deltas for the main MRR comparisons. The calibrated multimodal reranker significantly improves MRR over both the text evidence reranker and relation prior on both splits. The inductive Hits@10 gain over the text evidence reranker is directional but not statistically supported, so the main claim focuses on MRR.

A.3 Relation-Group Details

Table 9 expands Figure 3. The positive average gain is concentrated in material/style and part-whole relations, while visual-subject and contextual-historical relations remain unstable.

Split	Seed	MRR	H@10
Inductive	13	0.122	0.230
Inductive	21	0.081	0.213
Inductive	42	0.090	0.223
Transductive	13	0.130	0.238
Transductive	21	0.116	0.247
Transductive	42	0.132	0.249

Table 7: Seed-level adapted feature-DistMult results.

Split	Comparison	MRR Δ	95% CI
Inductive	Cal. MM – Text	+0.013	[0.009, 0.017]
Transductive	Cal. MM – Text	+0.020	[0.016, 0.024]
Inductive	Cal. MM – Prior	+0.035	[0.029, 0.041]
Transductive	Cal. MM – Prior	+0.048	[0.042, 0.054]

Table 8: Paired bootstrap deltas for the main MRR comparisons. Cal. MM denotes the calibrated multimodal reranker, Text denotes the Text Evidence Reranker, and Prior denotes the relation prior. All confidence intervals exclude zero, supporting the main MRR gains.

A.4 Downstream Diagnostic Details

Query construction. For each held-out triple (h, r, t) , we create an artwork QA-style question asking for the missing tail entity t . Relation-specific templates verbalize the relation, such as asking for the creator, material, movement, collection, or depicted subject of the artwork. The gold answer is the label of the held-out tail entity. Aliases are used only for evaluation normalization.

Evidence settings and controls. All non-oracle evidence settings use the same evidence budget. The no-KG setting provides no structured triples. The observed-train-KG setting provides only triples available in the training graph. Random completion adds randomly selected predicted triples. Relation-prior completion adds triples predicted by the relation-frequency prior. Calibrated multimodal completion adds triples predicted by the calibrated reranker. Oracle evidence adds the held-out gold triple and is used only as an upper bound. Non-oracle evidence construction does not use the held-out gold triple, the gold answer QID, or oracle relation paths.

Split-level results. Table 10 breaks down the main downstream comparison by split. Using the final multimodal completions improves both evidence recall and Copy EM over relation-prior completion on inductive-artwork and transductive-clean.

Oracle-gap decomposition. We further decompose the non-oracle outcomes into three mutually exclusive categories: correct answers, cases where the gold evidence is present in the supplied evidence but is not copied correctly, and cases where the gold tail is absent from the supplied evidence. Table 11 reports the split-level results.

A.5 Implementation and Compute Details

Encoders and retrieval. For dense text scoring, we use all-MiniLM-L6-v2. Text embeddings are L2-normalized and compared by dot product, equivalent to cosine similarity. BM25 is used only as a sparse-text baseline with $k_1 = 1.5$ and $b = 0.75$. For image scoring, we use openai/clip-vit-base-patch32; CLIP image features are 512-dimensional, L2-normalized, and compared by dot product. Relation priors are computed only from training triples. Standalone relation prior ranks candidates by raw training counts. In fusion, the KG prior uses $\log(1 + c_r(t))$ divided by the maximum log-count in the candidate pool. Dense text and image scores are normalized within each candidate pool. For each split and random seed, candidate size, text mixing weight, and image mixing weights are selected on the corresponding validation set and fixed for test evaluation.

Hyperparameter grids. For text reranking, we search $\text{top-}K \in \{50, 100, 200, 500\}$ and $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$. For calibrated multimodal reranking, we search $\text{top-}K \in \{100, 200, 500\}$, $\alpha_{\text{txt}} \in \{0.1, 0.3, 0.5\}$, and image weights in $\{0.0, 0.05, 0.1, 0.2, 0.3\}$. Validation MRR is used for selection.

KGE and MKGC baselines. KGE baselines use PyKEEN 1.11.1 with embedding dimension 200, 200 epochs, 64 negative samples per positive, and batch size 1024. Optimizer, loss, and learning rate are not explicitly overridden and follow PyKEEN defaults. Feature-DistMult uses 40 epochs, dimension 128, 16 negatives, batch size 1024, AdamW with learning rate 10^{-3} and weight decay 10^{-4} . MoSE-CompLexMDR uses rank 200, up to 100 epochs, batch size 512, early stopping patience 20, and SGD with learning rate 10^{-4} . Native-AdvRelRotatE uses 200 epochs, batch size 512, 64 negatives, dimension 250, margin 12, and learning rate 10^{-4} . AdaMKGC-adapted uses bert-base-uncased, up to 10 epochs, batch size 2, and evaluation batch size 4.

Group	Relation prior	Text rerank	Multimodal	Delta
Authorship/creator	0.061	0.071	0.068	-0.003
Visual subject	0.160	0.223	0.218	-0.005
Material/style	0.536	0.567	0.626	+0.060
Institution/location	0.117	0.126	0.127	+0.001
Part-whole	0.083	0.091	0.117	+0.026
Contextual/historical	0.100	0.155	0.134	-0.021

Table 9: Relation-group MRR analysis. Delta is multimodal reranker minus text evidence reranker. The gains are concentrated in material/style and part-whole relations.

Split	Evidence	Rec.	Copy	Reader
Inductive	Relation prior	0.212	0.145	0.129
Inductive	Multimodal	0.259	0.196	0.160
Transductive	Relation prior	0.240	0.161	0.177
Transductive	Multimodal	0.289	0.219	0.206

Table 10: Split-level downstream evidence-use diagnostic. Multimodal completion improves evidence recall, constrained copy exact match, and Reader EM over relation-prior completion on both splits.

Downstream diagnostic. Downstream evidence uses at most 8 triples per query, with at most 6 original triples, 2 completion triples, and 3 triples per relation. Reader decoding uses temperature 0.0 and a maximum of 32 new tokens.

Compute. Exact end-to-end GPU-hour accounting was not logged because experiments were run on a shared Slurm cluster with retries. Partial Slurm accounting recovered approximately 0.77 GPU-hours for the KGE full run, 1.5 GPU-hours for Native, and 41 GPU-hours for AdaMKGC-adapted. Observed GPU types include NVIDIA A100, RTX A6000, V100S, and Quadro RTX 6000.

Model sizes. We use fixed pretrained encoders rather than training large foundation models from scratch. The dense text encoder all-MiniLM-L6-v2 has 22,713,728 parameters, counted from the local model.safetensors file. The image encoder openai/clip-vit-base-patch32 has 151,277,440 parameters for the full local CLIP checkpoint; the image tower alone was not separately counted. Downstream reader scales include Qwen2.5 0.5B, 1.5B, 3B, 7B, and 14B, Llama-3.1 8B, Mistral 7B, and Phi-3.5 mini. AdaMKGC-adapted uses bert-base-uncased; the exact local parameter count was not verified. Exact trainable parameter counts for custom KGE/MKGC baselines were not logged.

A.6 Split Size Details

Table 12 reports seed-level split sizes and overlap diagnostics. For the inductive-artwork split, the key validity check is held-out overlap: designated validation/test held-out artwork entities have zero overlap with training entities for all seeds. Eval-head overlap is reported separately because some retained relations can connect multiple artwork entities, so a small number of evaluation triples may have heads that also occur as training heads.

A.7 Sentence Redaction Statistics

Table 13 reports corpus-level statistics from the sentence-redaction process, including the numbers of affected descriptions and removed sentence instances.

A.8 Redaction Ablation

Table 14 compares raw, sentence-redacted, and strict-redacted text for both standalone BM25 retrieval and the multimodal reranker. Sentence redaction is the main protocol used throughout the paper, while raw and strict text are included only as diagnostic settings. For the multimodal reranker, the candidate-pool size and validation-selected fusion weights are held fixed across the three text settings so that the comparison isolates the effect of redaction.

Raw descriptions consistently yield higher MRR than sentence-redacted descriptions on both split families. In the multimodal reranker, MRR decreases from 0.276780 to 0.269887 on inductive-artwork and from 0.311088 to 0.302775 on transductive-clean after sentence redaction. In contrast, strict redaction produces nearly the same performance as sentence redaction. This indicates that direct-answer mentions in raw descriptions measurably inflate evaluation, while sentence-level removal captures most of this measurable leakage effect.

Split	Completion	Correct	Evidence present, wrong copy	Gold absent
Inductive	Relation prior	14.52%	6.69%	78.79%
Inductive	Multimodal	19.56%	6.30%	74.14%
Transductive	Relation prior	16.09%	7.93%	75.98%
Transductive	Multimodal	21.92%	6.99%	71.09%

Table 11: Downstream error decomposition under the fixed evidence budget. Correct corresponds to Copy EM; evidence-present error denotes cases where the gold tail is available in the supplied evidence but is not copied correctly; gold absent denotes cases where the gold tail is not included in the supplied evidence. “Gold absent” does not mean that the gold tail is absent from the full KGC ranking.

Split family	Seed	Train	Valid	Test	Total	Eval-head overlap	Held-out overlap
Transductive-clean	13	32,752	4,093	4,093	40,938	3,907	–
Transductive-clean	21	32,752	4,093	4,093	40,938	3,886	–
Transductive-clean	42	32,752	4,093	4,093	40,938	3,883	–
Inductive-artwork	13	32,431	4,203	4,304	40,938	19	0
Inductive-artwork	21	32,777	4,114	4,047	40,938	14	0
Inductive-artwork	42	32,909	3,921	4,108	40,938	15	0

Table 12: Seed-level split sizes and overlap diagnostics. Eval-head overlap counts broad validation/test artwork heads that also appear as training heads. Held-out overlap counts overlap between designated held-out artwork entities and training entities; this value is zero for all inductive-artwork seeds.

Item	Count
Description rows	13,335
Descriptions with raw text	5,340
Raw sentence instances	40,404
Remaining after redaction	32,674
Removed sentence instances	7,730
Removed sentence rate	19.13%
Unique removed description sentences	5,799
Affected descriptions	2,331
Affected description rate	43.65%
Empty after redaction	447
Empty description rate	8.37%

Table 13: Corpus-level sentence-redaction statistics. Counts are computed from the redaction process itself, not by reparsing the serialized redacted JSONL files.

A.9 Description Availability and Richness

Table 15 evaluates sensitivity to the availability and length of the artwork-side description after sentence redaction. We distinguish *unavailable*, where no raw artwork-side description exists, from *empty*, where a raw description exists but becomes empty after sentence redaction. Artwork and relation labels, as well as candidate-side text, remain available in both cases. The remaining queries are grouped by the number of words in the post-redaction artwork description. Results are pooled over test queries from the three random seeds.

A.10 Image Availability

Table 16 evaluates performance according to whether a cached head-artwork image is avail-

Method	Text	Ind. MRR	Trans. MRR
BM25	Raw	0.045055	0.030931
BM25	Sentence	0.038242	0.023911
BM25	Strict	0.038243	0.023908
MM reranker	Raw	0.276780	0.311088
MM reranker	Sentence	0.269887	0.302775
MM reranker	Strict	0.269926	0.302771

Table 14: Raw, sentence-redacted, and strict-redacted text diagnostics. For the multimodal (MM) reranker, the reranking configuration is held fixed across text settings. Raw descriptions yield higher MRR, while strict redaction is nearly unchanged from sentence redaction.

able. The distinction is based on the image actually available to the evaluation pipeline, rather than merely on the presence of an image URL. When no cached head image is available, the relation-role image weight is inactive and the reranker uses its validation-selected structural and textual score.

The multimodal reranker improves over the separately tuned Text Evidence Reranker when a cached head image is available, by 0.013897 MRR on inductive-artwork and 0.018158 on transductive-clean. Without a cached head image, the image weight is inactive; the small remaining differences arise because the structural/textual base configuration is independently selected on validation data. Thus, the aggregate visual gain should be interpreted as a coverage-dependent effect rather than as a uniform improvement for all queries.

Split	Post-redaction description	Queries	Relation prior	Text reranker	Multimodal
Inductive	Unavailable	7,695	0.237376	0.258430	0.271326
Inductive	Empty	1,397	0.235312	0.257776	0.274004
Inductive	1-25 words	2,914	0.228325	0.254866	0.267219
Inductive	26-75 words	453	0.229017	0.246631	0.249569
Transductive	Unavailable	7,297	0.255827	0.282494	0.304833
Transductive	Empty	1,048	0.247991	0.276286	0.291176
Transductive	1-25 words	3,333	0.254618	0.297089	0.303623
Transductive	26-75 words	601	0.255859	0.290863	0.293313

Table 15: Sensitivity to post-redaction artwork-description availability and length. MRR is reported for the relation prior, Text Evidence Reranker, and calibrated multimodal reranker. Results pool test queries across the three seeds. *Unavailable* denotes queries without a raw artwork-side description, whereas *empty* denotes descriptions that become empty after sentence redaction.

Split	Head image	Queries	Text reranker	Multimodal	Δ MRR
Inductive	Cached	11,580	0.261062	0.274959	+0.013897
Inductive	No cached image	879	0.204822	0.202887	-0.001934
Transductive	Cached	11,354	0.291115	0.309273	+0.018158
Transductive	No cached image	925	0.227675	0.223014	-0.004661

Table 16: Sensitivity to head-image availability. Multimodal gains are concentrated among queries with cached artwork images. “No cached image” denotes queries for which no local head-image feature is available to the evaluation pipeline, regardless of whether an image URL is present.

A.11 Relation-Frequency Prior Analysis

To understand why the head-agnostic relation prior is competitive, Table 17 measures the concentration of relation-specific tails in the training graph. Top-1 and top-5 mass denote the fractions of training triples for a relation accounted for by its most frequent one and five tails, respectively. Prior Recall@10 is the fraction of test queries whose gold tail appears among the ten most frequent training tails for that relation.

The contrast is consistent across both split families. Material is dominated by a single frequent tail, while genre is concentrated among a small set of common tails. Creator is substantially more diffuse. Since the relation prior ignores the head artwork entirely, its strong performance therefore reflects relation-specific tail concentration rather than successful artwork-specific inference.

A.12 Qualitative Multimodal Cases

Table 18 reports aligned examples from the final Text Evidence Reranker and calibrated multimodal reranker on identical test queries. The examples illustrate both successful and harmful uses of visual evidence.

Split	Relation	Top-1 mass	Top-5 mass	Prior Recall@10
Inductive	Genre (P136)	0.229958	0.767642	0.887184
Inductive	Material (P186)	0.612984	0.814887	0.879257
Inductive	Creator (P170)	0.018623	0.073089	0.132638
Transductive	Genre (P136)	0.232999	0.769974	0.893096
Transductive	Material (P186)	0.606053	0.808722	0.901042
Transductive	Creator (P170)	0.015425	0.064208	0.178571

Table 17: Training-tail concentration and relation-prior Recall@10 for representative relations. Statistics are aggregated over the three random seeds. High tail concentration explains the strength of the relation-frequency prior for genre and material, whereas creator has a substantially more diffuse tail distribution.

Split	Relation	Artwork	Gold tail	Text rank	MM rank
Transductive	depicts	The Death of Marat	drapery	64	9
Transductive	depicts	Recanati Polyptych	Virgin Mary	54	3
Transductive	depicts	The Burial of the Count of Orgaz	angel	48	6
Inductive	collection	The Cliff Walk at Pourville	Art Institute of Chicago	8	56
Inductive	collection	Plain near Auvers	Bayerische Staatsgemäldesammlungen	19	90

Table 18: Aligned query-level examples comparing the final Text Evidence Reranker with the calibrated multimodal reranker. The first three examples illustrate visually grounded improvements, whereas the last two show harmful visual fusion for collection membership. These are instance-level examples and do not imply a positive average gain for the visual-subject relation group.