UBER: UNCERTAINTY-AWARE BRIDGE-STRUCTURED EVIDENCE ROUTING FOR MULTI-KNOWLEDGE GRAPH JOINT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-knowledge graph (multi-KG) joint reasoning aims to answer queries or complete missing facts by exploiting complementary evidence distributed across multiple knowledge graphs. Existing methods mainly focus on cross-KG alignment, global graph integration, or federated representation learning, but often overlook two key challenges: how to route query-specific evidence across different source KGs, and how to reason robustly when cross-KG bridge links are incomplete or unreliable. To address these challenges, we propose **UBER**, an **U**ncertainty-aware **B**ridge-structured **E**vidence **R**outing framework for multi-KG joint reasoning. UBER consists of three tightly coupled modules: a Query-aware Source Router (QSR) that dynamically allocates evidence budgets across KGs, an Uncertainty-aware Bridge Constructor (UBC) that models reliable entity- and relation-level bridges, and a Reliability-Gated Fusion Reasoner (RFR) that performs query-conditioned reasoning over compact fused evidence graphs. Instead of globally merging all KGs, UBER constructs a query-specific evidence graph and controls cross-KG propagation according to bridge reliability and relation compatibility. Extensive experiments on three constructed multi-KG completion benchmarks demonstrate that UBER achieves the strongest overall performance over representative non-federated multi-KG, cross-KG, and federated baselines.

1 INTRODUCTION

Knowledge graphs (KGs) provide structured representations of entities and relations and support a wide range of applications, including question answering Lu et al. (2025), recommendation Wang et al. (2025), and reasoning-intensive decision making Men et al. (2025). In real-world scenarios, however, relevant knowledge is often distributed across multiple KGs constructed by different institutions, domains, languages, or modalities Li & Huang (2026b). These KGs are typically complementary: domain-specific KGs provide precise factual knowledge, commonsense KGs offer general semantic priors, and multimedia-enriched KGs contain textual or visual evidence Liu et al. (2025). Effective inference may therefore require jointly exploiting evidence scattered across multiple sources.

This setting has motivated increasing interest in multi-KG joint reasoning, which improves inference by transferring and integrating complementary knowledge across graphs Xie et al. (2025). Existing studies have advanced cross-KG alignment, multilingual KG completion, and multi-source representation learning Akhtar et al. (2025); Shen et al. (2026). Most methods first establish cross-KG correspondences or learn a shared representation space, and then perform reasoning over the aligned graphs. However, this alignment-first paradigm remains insufficient for query-time reasoning over heterogeneous and imperfectly connected KGs Li & Huang (2026a). As illustrated in Figure 1, two key limitations remain.

First, **existing methods provide limited support for query-specific evidence routing.** After different KGs are aligned or embedded into a shared space, their information is often aggregated through global representations or largely query-independent strategies. Yet, the usefulness of each source KG varies across queries: a graph that is informative for one relation may be irrelevant or misleading for another. A multi-KG reasoner should therefore determine which KGs to access, how much

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

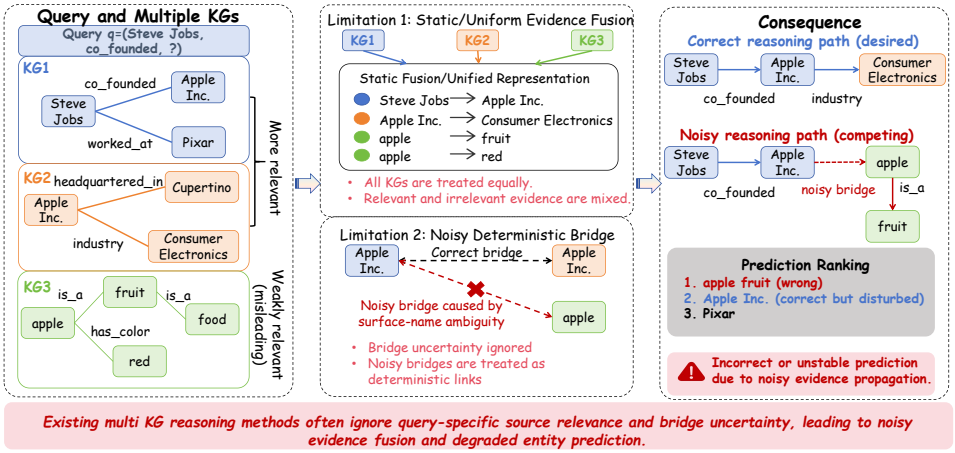

Figure 1: Motivation of UBER. Existing multi-KG reasoning methods may uniformly aggregate sources with different query relevance and propagate evidence through uncertain cross-KG bridges, resulting in irrelevant evidence fusion and unstable predictions.

evidence to retrieve from each source, and which local regions to retain under a limited inference budget. Without such query-aware control, relevant and distracting evidence may be mixed before their utility is established.

Second, **cross-KG reasoning is highly sensitive to bridge reliability.** Existing pipelines mainly rely on entity correspondences, while relation-level compatibility and bridge uncertainty remain comparatively underexplored. In practice, entity bridges may be incomplete or ambiguous because of name collisions, sparse descriptions, and heterogeneous schemas. Likewise, semantically related relations may use different labels or exhibit different structural patterns across KGs. Treating these correspondences as deterministic links can propagate noisy messages and corrupt otherwise valid reasoning paths. Robust reasoning therefore requires uncertainty-aware entity bridges, relation-level compatibility modeling, and explicit control over cross-KG propagation.

Motivated by these observations, we formulate multi-KG joint reasoning as an **uncertainty-aware bridge-structured evidence routing** problem. Instead of globally merging all source KGs, the model constructs a compact evidence graph tailored to each query. It dynamically selects useful KGs, retrieves budgeted local evidence, establishes reliable entity- and relation-level bridges, and regulates propagation according to bridge uncertainty and relation compatibility. This formulation shifts multi-KG reasoning from static graph integration to query-specific and reliability-controlled evidence composition.

Based on this formulation, we propose **UBER**, an **U**ncertainty-aware **B**ridge-structured **E**vidence **R**outing framework for multi-KG joint reasoning. UBER contains three tightly coupled modules. The **Query-aware Source Router (QSR)** estimates source relevance, while its **Dynamic Budget Allocator (DBA)** assigns source-specific retrieval budgets and extracts compact evidence subgraphs. The **Uncertainty-aware Bridge Constructor (UBC)** builds reliability-weighted entity bridges, and its **Relation Bridge Matcher (RBM)** captures semantic and structural compatibility between cross-KG relations. Finally, the **Reliability-Gated Fusion Reasoner (RFR)** performs query-conditioned message passing over the fused evidence graph, strengthening informative reasoning paths while suppressing unreliable or incompatible propagation. Thus, UBER enables compact, on-demand, and reliability-controlled reasoning without indiscriminately merging all KGs. Our main contributions are summarized as follows:

- We formulate multi-KG joint reasoning as an **uncertainty-aware bridge-structured evidence routing** problem, explicitly considering query-dependent source relevance and unreliable cross-KG connectivity.
- We propose UBER, which integrates query-aware source routing and dynamic evidence budgeting with uncertainty-aware entity bridges, relation-level bridge matching, and reliability-gated fusion reasoning.

- Extensive experiments on three fragmented multi-KG completion benchmarks demonstrate consistent improvements over non-federated and federated baselines. Further ablation, robustness, scalability, sensitivity, efficiency, and interpretability analyses verify the effectiveness of the proposed components.

2 Related Work

2.1 Non-Federated Multi-KG Completion and Cross-KG Reasoning

Non-federated multi-KG completion mainly studies knowledge transfer across related KGs, particularly in multilingual settings. KEnS Chen et al. (2020) learns a shared space for language-specific KGs, while AlignKGC Singh et al. (2021) jointly models KG completion and entity/relation alignment. SS-AGA Huang et al. (2022) and JMAC Tong et al. (2022) exploit alignment links for relation-aware cross-KG propagation, whereas LSMGA Tang et al. (2023) introduces language-sensitive multi-graph attention to mitigate cross-KG inconsistency. Recent methods further leverage pretrained language models, mixture-of-experts, information-theoretic learning, and diffusion: GC-PLM Song et al. (2023) incorporates global and local KG constraints into pretrained models; KL-GMoE Mao et al. (2025) exploits multilingual shared knowledge through grouped experts; IMKGC Sheng et al. (2026b) learns minimal sufficient representations to preserve complementary knowledge while suppressing redundancy; and DMKGC Sheng et al. (2026a) transfers domain-general knowledge via conditional diffusion. Despite their effectiveness, these methods largely rely on static alignment, shared representations, or global aggregation, without explicitly deciding which KGs to access for each query, how to allocate evidence budgets, or how bridge uncertainty should regulate propagation. In contrast, UBER constructs compact query-specific evidence graphs and performs reliability-controlled reasoning over uncertain entity- and relation-level bridges.

2.2 Federated Multi-KG Completion and Reasoning

Federated KGC enables collaborative completion over distributed KGs without sharing raw triples. FedE Chen et al. (2021) and FedR Zhang et al. (2022) aggregate locally trained entity and relation embeddings, respectively, while FedEC Chen et al. (2022) uses contrastive learning to mitigate representation drift and client heterogeneity, and FedLU Zhu et al. (2023) supports heterogeneous federated learning and unlearning through mutual knowledge distillation. Recent methods further relax alignment and efficiency constraints: UniFE Chen et al. (2024) removes the need for explicit aligned entity sets, PFLU Zhu et al. (2025) introduces parameter-efficient federated learning and unlearning, and FKGE Peng et al. (2021) extends federated reasoning to complex query answering over distributed neural graph databases. Despite their relevance, these methods primarily focus on training-time collaboration or decentralized privacy-preserving reasoning. UBER instead targets centralized inference-time multi-KG reasoning, dynamically routing query-relevant evidence, constructing compact subgraphs, estimating bridge reliability, and performing query-conditioned fusion over uncertain cross-KG bridges.

3 Methodology

3.1 Problem Formulation

We consider M source knowledge graphs $\{\mathcal{G}^{(m)}\}_{m=1}^{M}$, where $\mathcal{G}^{(m)} = (\mathcal{E}^{(m)}, \mathcal{R}^{(m)}, \mathcal{T}^{(m)})$ contains the entity, relation, and triple sets of source m. Given a completion query $q = (h, r_q, ?)$, the goal is to rank candidate answers from $\mathcal{E}^{\text{all}} = \bigcup_{m=1}^{M} \mathcal{E}^{(m)}$. Each entity v is represented by observable semantic, relation-profile, and local structural information, yielding an initial representation $\mathbf{u}_v = \text{LN}(\mathbf{W}_x \mathbf{x}_v + \mathbf{b}_x)$. The query is encoded as $\mathbf{q} = \text{LN}(\mathbf{W}_h \mathbf{u}_h + \mathbf{W}_r \mathbf{e}_{r_q})$, where $\mathbf{e}_{r_q}$ denotes the query-relation representation. The construction of the observable entity and relation features is specified in Appendix B. Instead of globally merging all source KGs, UBER constructs a compact query-specific fused evidence graph $\mathcal{G}_q^{\text{fus}}$. This turns multi-KG reasoning into three coupled decisions: **which sources should be accessed**, **which cross-KG connections should be trusted**, and **how trusted evidence should influence prediction**.

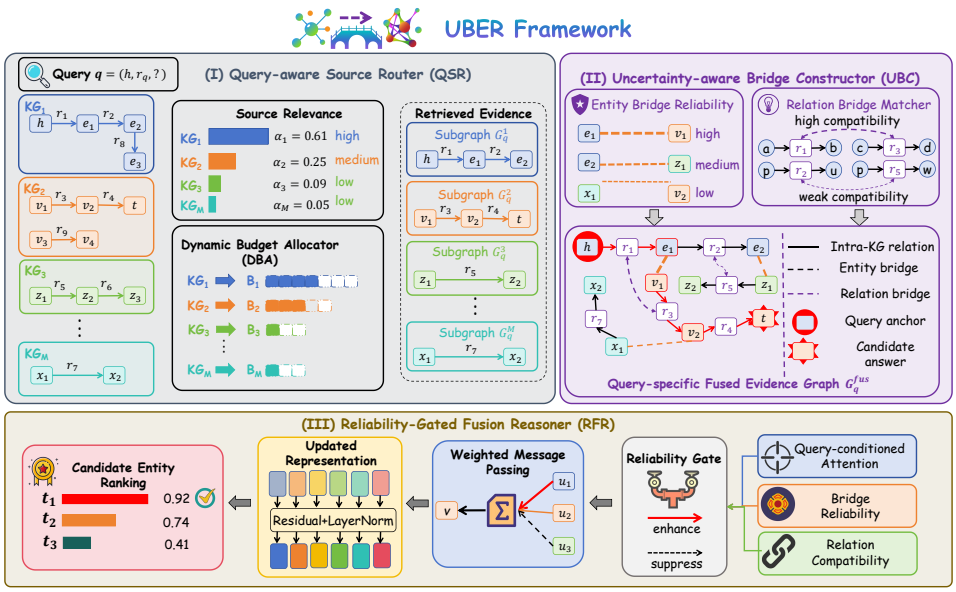

Figure 2: Overall framework of UBER. QSR determines where and how much evidence to retrieve; UBC estimates which cross-KG connections are trustworthy; RFR performs reliability-controlled reasoning over the resulting query-specific fused evidence graph.

3.2 Framework Overview

As illustrated in Figure 2, UBER follows a three-stage route–connect–reason pipeline. First, the **Query-aware Source Router (QSR)** determines **where** and **how much** evidence should be retrieved. Its Dynamic Budget Allocator (DBA) distributes a finite evidence budget according to query-specific source relevance and extracts compact local evidence subgraphs. Second, the **Uncertainty-aware Bridge Constructor (UBC)** determines **which cross-KG connections are trustworthy**. It evaluates candidate entity correspondences from multiple observable views, while the Relation Bridge Matcher (RBM) models semantic and structural compatibility between relations from different source KGs. Finally, the **Reliability-Gated Fusion Reasoner (RFR)** determines **how the retrieved evidence should affect prediction**. It performs query-conditioned message passing while explicitly suppressing propagation through uncertain entity bridges and incompatible relations. Together, these stages replace static multi-KG integration with compact, query-specific, and reliability-controlled evidence composition.

3.3 Query-aware Source Router

QSR first estimates source relevance, then distributes a finite evidence budget, and finally retrieves query-compatible local evidence.

Source relevance. We first construct a high-recall semantic pre-candidate pool using edge-independent textual representations and then rerank these candidates with observable graph-dependent features. During training, all graph-dependent reranking features are computed from the target-masked support graphs. For source KG $\mathcal{G}^{(m)}$, the resulting cross-KG counterparts of the query entity form the anchor set $\mathcal{A}^{(m)}(q)$. Pooling their representations gives the source summary $\mathbf{g}_m$. The query-specific relevance of source m is

$$a_m = \mathrm{softmax}_m \left[\mathbf{w}_a^\top \tanh \left(\mathbf{W}_q \mathbf{q} + \mathbf{W}_g \mathbf{g}_m \right) \right], \tag{1}$$

where $\sum_{m=1}^{M} a_m = 1$. Unlike static source weighting, a_m varies with the query, allowing different source KGs to specialize in different reasoning contexts.

Dynamic budget allocation. Given a total evidence budget B_{tot} and a minimum exploration budget $B_{\min}$, DBA computes

$$\widetilde{B}_m = B_{\min} + (B_{\text{tot}} - MB_{\min})a_m. \tag{2}$$

The continuous budgets are deterministically rounded into integer budgets $\{B_m\}_{m=1}^{M}$ while preserving $\sum_m B_m = B_{\text{tot}}$. Thus, informative sources receive more retrieval capacity without completely discarding weakly scored sources.

Query-aware evidence retrieval. Within source KG m, outgoing edges are prioritized according to their compatibility with the current query:

$$P_q^{(m)}(u, r, v) = \frac{\exp\left(\mathbf{q}^\top \mathbf{W}_\omega \mathbf{e}_r / \tau_q\right)}{\sum_{(u,r',v')\in\mathcal{T}^{(m)}(u)} \exp\left(\mathbf{q}^\top \mathbf{W}_\omega \mathbf{e}_{r'} / \tau_q\right)}, \tag{3}$$

where τ_q is a temperature parameter and $\mathcal{T}^{(m)}(u)$ denotes the outgoing triples of entity u. Starting from $\mathcal{A}^{(m)}(q)$, QSR expands high-probability paths and retains at most B_m entities, yielding $\mathcal{G}_q^{(m)} = (\mathcal{V}_q^{(m)}, \mathcal{F}_q^{(m)})$.

Learning discrete evidence selection. The integer budget and hard top-B_m retrieval are discrete operations; we therefore do not differentiate through the selection operators themselves. Instead, for each training query, auxiliary source- and path-level targets are constructed from the masked support graphs defined above.

Source-level targets identify KGs containing useful reasoning evidence, while path-level targets supervise the relation-guided transition model. Both signals are derived exclusively from training data and observable cross-KG candidates. Because the positive target edge has already been removed before the forward pass, neither retrieval supervision nor downstream message passing can directly access the target fact.

The selection objective is

$$\mathcal{L}_{\text{sel}} = \mathcal{L}_{\text{route}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}}. \tag{4}$$

This explicitly trains the routing and retrieval scoring functions without requiring gradients through hard evidence selection. where $\mathcal{L}_{\text{route}}$ trains source relevance and $\mathcal{L}_{\text{ret}}$ trains relation-guided local expansion.

3.4 Uncertainty-aware Bridge Constructor

Although the retrieved subgraphs are individually relevant to the query, they remain disconnected across sources. UBC connects them by estimating whether candidate cross-KG correspondences are both plausible and reliable.

Multi-view entity correspondence. For candidate entities (u, v) from different KGs, UBC combines semantic, relation-neighborhood, and structural evidence:

$$\boldsymbol{\phi}_{uv} = \left[\cos(\mathbf{z}_u, \mathbf{z}_v); \cos(\mathbf{p}_u, \mathbf{p}_v); \text{sim}_{\text{str}}(u, v)\right], \tag{5}$$

where $\mathbf{z}_v$ represents entity semantics, $\mathbf{p}_v$ summarizes its direction-aware relation neighborhood, and sim_{str} compares local structural patterns. Two prediction heads estimate matching confidence and uncertainty:

$$\mu_{uv} = \sigma\left(\mathbf{w}_\mu^\top \boldsymbol{\phi}_{uv} + b_\mu\right), \qquad \nu_{uv} = \text{softplus}\left(\mathbf{w}_\nu^\top \boldsymbol{\phi}_{uv} + b_\nu\right). \tag{6}$$

Their combination defines bridge reliability:

$$\rho_{uv} = \mu_{uv} \exp(-\lambda_u \nu_{uv}), \tag{7}$$

where λ_u controls uncertainty suppression. Hence, high matching confidence alone is insufficient: a correspondence is down-weighted when the available evidence remains uncertain. Only candidate pairs satisfying $\rho_{uv} \geq \tau_e$ are retained. The confidence and uncertainty heads are calibrated from training-observable reciprocal matching signals rather than latent canonical entity mappings.

Relation Bridge Matcher. Reliable entity correspondence does not guarantee that two source KGs describe evidence using compatible relations. We therefore represent relation r by

$$\mathbf{c}_r = \mathrm{MLP}\left([\mathbf{e}_r; \mathbf{s}_r]\right), \tag{8}$$

where $\mathbf{s}_r$ summarizes the source-local head/tail distributions, frequency, functionality, and inverse functionality of relation r. For relations r and r' from different source KGs, RBM computes

$$\kappa_{rr'} = \sigma\left(\mathbf{c}_r^\top \mathbf{W}_c \mathbf{c}_{r'}\right). \tag{9}$$

Pairs satisfying $\kappa_{rr'} \geq \tau_r$ are retained. Unlike entity bridges, relation bridges are used as compatibility priors rather than ordinary graph edges. The retained entity bridges are inserted between the retrieved local subgraphs, producing the query-specific fused evidence graph $\mathcal{G}_q^{\mathrm{fus}}$.

3.5 Reliability-Gated Fusion Reasoner

RFR performs multi-hop reasoning over the fused evidence graph. Its central principle is that evidence should contribute according to both its relevance to the current query and the reliability of the cross-KG path through which it is propagated.

Query-conditioned relation compatibility. Let $\bar{\kappa}_r$ denote the maximum retained cross-KG compatibility of relation r, with $\bar{\kappa}_r = 0$ when no compatible relation exists. We define

$$\eta_r(q) = \sigma\left(\mathbf{q}^\top \mathbf{W}_\eta \mathbf{c}_r\right)(1 + \bar{\kappa}_r), \tag{10}$$

where the first term measures query–relation relevance and $\bar{\kappa}_r$ provides a cross-KG consistency prior. For the special entity-bridge relation r_b, we set $\eta_{r_b}(q) = 1$.

Reliability-gated propagation. Each entity v is initialized through a source-specific projection. At reasoning layer ℓ, the attention logit of edge (u, r, v) is

$$e_{uv}^{(\ell)} = \frac{\left(\mathbf{W}_Q[\mathbf{h}_v^{(\ell)}; \mathbf{q}]\right)^\top \left(\mathbf{W}_K[\mathbf{h}_u^{(\ell)}; \mathbf{e}_r]\right)}{\sqrt{d}} + \lambda_b \log(\widetilde{\rho}_{uv} + \epsilon) + \lambda_c \log(\eta_r(q) + \epsilon), \tag{11}$$

where $\widetilde{\rho}_{uv} = 1$ for intra-KG edges and $\widetilde{\rho}_{uv} = \rho_{uv}$ for entity bridges. The three terms respectively capture query-conditioned message relevance, entity-bridge reliability, and relation-level compatibility. After normalizing the logits into $\alpha_{uv}^{(\ell)}$, the representation is updated by

$$\mathbf{h}_v^{(\ell+1)} = \mathrm{LN}\left(\mathbf{h}_v^{(\ell)} + \delta\left(\sum_{u \in \mathcal{N}(v)} \alpha_{uv}^{(\ell)} \mathbf{W}_V[\mathbf{h}_u^{(\ell)}; \mathbf{e}_r]\right)\right). \tag{12}$$

The residual connection preserves source-local information while integrating reliability-controlled evidence over multiple hops.

Candidate scoring. After L reasoning layers, candidate entity t is scored as

$$s(q, t) = \left(\mathbf{h}_{\mathrm{src}}^{(L)}\right)^\top \mathrm{Diag}(\mathbf{W}_s \mathbf{q}) \overline{\mathbf{h}}_t, \tag{13}$$

where $\overline{\mathbf{h}}_t = \mathbf{h}_t^{(L)}$ for retrieved entities and is the projected base representation otherwise. Thus, message passing remains restricted to the compact evidence graph, while final ranking still covers the complete candidate space.

3.6 Optimization Objective

Target-edge masking during training. To prevent direct target leakage, each training query is evaluated on a query-specific support graph in which its positive training edge is temporarily removed. For a training query $q = (h, r_q, t)$ whose positive triple belongs to source KG s, we define

$$\mathcal{T}_{-q}^{(s)} = \mathcal{T}_{\mathrm{tr}}^{(s)} \setminus \{(h, r_q, t)\}, \tag{14}$$

while $\mathcal{T}_{-q}^{(m)} = \mathcal{T}_{\mathrm{tr}}^{(m)}$ for $m \neq s$. If inverse relations are explicitly materialized, the corresponding inverse edge (t, r_q^{-1}, h) is removed simultaneously.

All query-dependent graph operations, including candidate reranking, source summarization, evidence retrieval, supervision-path construction, bridge construction from graph-dependent features, and reliability-gated message passing, are performed on the masked support graphs $\{\mathcal{G}_{-q}^{(m)}\}_{m=1}^{M}$. The removed triple is used only as the supervision label in $\mathcal{L}_{\mathrm{pred}}$ and is never exposed to the forward reasoning procedure. Given ground-truth answer t and negative candidate set $\mathcal{N}^-$, the prediction objective is

$$\mathcal{L}_{\mathrm{pred}} = -\log \sigma(s(q,t)) - \sum_{t^- \in \mathcal{N}^-} \log \sigma(-s(q,t^-)). \tag{15}$$

Besides answer prediction, UBER jointly optimizes evidence selection, bridge calibration, and cross-KG representation consistency. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{sel}}\mathcal{L}_{\mathrm{sel}} + \lambda_{\mathrm{br}}\mathcal{L}_{\mathrm{br}} + \lambda_e\mathcal{L}_e + \lambda_r\mathcal{L}_r, \tag{16}$$

where $\mathcal{L}_{\mathrm{br}}$ calibrates bridge confidence and uncertainty, while $\mathcal{L}_e$ and $\mathcal{L}_r$ encourage consistency only between reliable entity and relation correspondences. The discrete budget and evidence-selection operations are not treated as differentiable operators; their scoring functions are explicitly trained through $\mathcal{L}_{\mathrm{sel}}$. Complete target construction, calibration, and optimization details are given in Appendices A–B.

4 Experiments

We evaluate UBER on fragmented multi-KG completion, where evidence is distributed across incomplete source KGs connected by sparse and uncertain bridges. Accordingly, we compare UBER with non-federated multi-KG completion, cross-KG reasoning, and federated KGC methods under the same fragmented setting. Full-graph single-KG methods are excluded from the main comparison because they assume access to the complete original KG and therefore operate under a stronger information condition.

4.1 Datasets and Settings

We construct multi-KG benchmarks from FB15k-237 Toutanova et al. (2015), WN18RR Xiong et al. (2017), and YAGO3-10 Mahdisoltani et al. (2014). Each original KG is partitioned into $M \in \{2, 3, 4, 5\}$ source KGs by controlling entity overlap, relation overlap, and complementary evidence distribution. The entity-overlap ratio is defined as $\rho_e = |\mathcal{E}_{\mathrm{shared}}|/|\mathcal{E}|$, where $\mathcal{E}_{\mathrm{shared}}$ contains entities appearing in at least two source KGs. We evaluate $\rho_e \in \{10\%, 20\%, 30\%, 40\%\}$ while fixing the relation-overlap ratio at $\rho_r = 40\%$. Thus, "4KG–20%" denotes four source KGs with 20% entity overlap. The complete benchmark construction and leakage-control protocol is provided in the supplementary material.

4.2 Baselines and Evaluation Protocol

We compare UBER with two groups of methods. The first group contains non-federated multi-KG and cross-KG baselines, including KEnS Chen et al. (2020), AlignKGC Singh et al. (2021), SS-AGA Huang et al. (2022), JMAC Tong et al. (2022), LSMGA Tang et al. (2023), GC-PLM Song et al. (2023), IMKGC Sheng et al. (2026b), DMKGC Sheng et al. (2026a), and KL-GMoE Mao et al. (2025). The second group contains federated KGC baselines, including FedE Chen et al. (2021), FedR Zhang et al. (2022), FedEC Chen et al. (2022), FedLU Zhu et al. (2023), FKGE Peng et al. (2021), UniFE Chen et al. (2024), and PFLU Zhu et al. (2025). All metrics are computed under canonical filtered ranking. Source-local predictions are first aggregated by their canonical entities using the evaluation protocol described in Appendix C.2.

4.3 Main Results

Table 1 summarizes the main results under the 4KG–20% overlap setting. UBER achieves the best MRR and Hits@1 on all three benchmarks and obtains the strongest overall performance on most

metrics. These results demonstrate that query-aware source routing and reliability-controlled cross-KG reasoning are effective for identifying high-quality evidence and ranking the correct entity at top positions. Recent methods such as IMKGC and DMKGC remain competitive on several higher-order Hits metrics, indicating their strength in broad candidate coverage. In contrast, UBER shows more consistent advantages on the stricter ranking metrics, suggesting that uncertainty-aware bridge construction and reliability-gated fusion can better suppress irrelevant sources and noisy cross-KG propagation.

Method	FB15k-237-MKG				WN18RR-MKG				YAGO3-10-MKG			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
Non-federated multi-KG and cross-KG methods												
KEnS	0.321	0.236	0.354	0.506	0.452	0.411	0.476	0.541	0.501	0.421	0.548	0.671
AlignKGC	0.335	0.250	0.369	0.524	0.466	0.425	0.492	0.557	0.518	0.438	0.567	0.690
SS-AGA	0.339	0.252	0.373	0.529	0.468	0.428	0.494	0.560	0.521	0.441	0.570	0.692
JMAC	0.346	0.260	0.381	0.536	0.463	0.423	0.490	0.554	0.528	0.449	0.578	0.699
LSMGA	0.356	0.270	0.394	0.561	0.481	0.442	0.511	0.585	0.548	0.468	0.607	0.715
GC-PLM	0.353	0.281	0.389	0.547	0.488	0.452	0.517	0.579	0.554	0.472	0.604	0.714
KL-GMoE	0.359	0.278	0.402	0.558	0.492	0.450	0.520	0.582	0.552	0.476	0.605	0.718
IMKGC	0.369	0.277	0.396	0.567	0.504	0.449	**0.555**	0.583	0.559	0.470	0.604	0.719
DMKGC	0.376	0.279	0.399	0.565	0.510	0.451	0.549	0.584	0.565	0.474	0.606	**0.724**
Federated KGC methods												
FedE	0.307	0.221	0.338	0.491	0.438	0.398	0.464	0.528	0.489	0.408	0.535	0.656
FedR	0.324	0.238	0.356	0.510	0.461	0.421	0.487	0.551	0.516	0.435	0.564	0.682
FedEC	0.317	0.231	0.348	0.501	0.454	0.414	0.480	0.544	0.509	0.428	0.556	0.676
FKGE	0.331	0.245	0.363	0.519	0.446	0.406	0.472	0.536	0.501	0.421	0.548	0.669
FedLU	0.343	0.257	0.376	0.531	0.479	0.439	0.507	0.570	0.532	0.452	0.583	0.697
UniFE	0.336	0.250	0.369	0.526	0.472	0.432	0.499	0.563	0.541	0.461	0.592	0.704
PFLU	0.361	0.274	0.397	0.552	0.489	0.446	0.523	0.579	0.547	0.466	0.601	0.717
UBER	**0.386**	**0.318**	**0.450**	**0.579**	**0.517**	**0.475**	0.552	**0.607**	**0.571**	**0.508**	**0.613**	0.721

Table 1: Main results on the constructed multi-KG completion benchmarks under the 4KG–20% overlap setting. All methods are evaluated under the same fragmented multi-KG protocol. Best results are shown in bold and second-best results are underlined.

4.4 Ablation Study

Table 2 shows consistent trends across all three benchmarks. Removing RFR causes the largest MRR degradation (0.038–0.046), confirming that retrieved cross-KG evidence must be selectively propagated rather than simply fused. Removing UBC or the reliability gate also substantially weakens performance, highlighting the importance of controlling uncertain cross-KG connections.

Importantly, replacing uncertainty-aware reliability $\rho_{uv} = \mu_{uv} \exp(-\lambda_u \nu_{uv})$ with confidence alone $\rho_{uv} = \mu_{uv}$ consistently reduces MRR by 0.008–0.009 across the three datasets. This shows that the uncertainty term provides complementary information beyond matching confidence rather than merely rescaling bridge scores. The further gap between *Entity-Bridge Only* and the full model demonstrates the additional value of relation-level compatibility. Finally, removing QSR or DBA yields smaller but consistent degradation, supporting query-specific source selection and adaptive evidence allocation.

4.5 Hyperparameter Sensitivity

We investigate the sensitivity of UBER to four key hyperparameters: the reasoning depth L, total evidence budget B_{tot}, bridge reliability weight λ_b, and relation compatibility weight λ_c. Figure 3 reports results on FB15k-237-MKG under the 4KG–20% setting. Shaded regions indicate the best overall configurations, while stars mark metric-specific optima.

UBER performs best overall with $L = 3$, suggesting that moderate multi-hop propagation is sufficient to integrate complementary evidence; shallower reasoning limits the receptive field, whereas excessive depth introduces redundant or noisy messages. Similarly, performance improves with B_{tot} up to a moderate budget and then declines, supporting compact, query-relevant evidence retrieval over indiscriminate expansion.

Variant	FB15k-237-MKG				WN18RR-MKG				YAGO3-10-MKG			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
UBER	**0.386**	**0.318**	**0.450**	**0.579**	**0.517**	**0.475**	**0.552**	**0.607**	**0.571**	**0.508**	**0.613**	**0.721**
w/o QSR	0.360	0.284	0.418	0.548	0.493	0.451	0.528	0.584	0.542	0.481	0.584	0.692
w/o DBA	0.372	0.301	0.432	0.562	0.505	0.464	0.540	0.596	0.557	0.495	0.599	0.708
w/o UBC	0.354	0.279	0.412	0.540	0.486	0.443	0.521	0.576	0.531	0.470	0.572	0.681
w/o Uncertainty ($\rho = \mu$)	0.378	0.307	0.441	0.570	0.509	0.466	0.544	0.599	0.562	0.497	0.605	0.713
w/o RBM	0.373	0.302	0.434	0.564	0.506	0.465	0.541	0.597	0.554	0.492	0.596	0.705
Entity-Bridge Only	0.366	0.294	0.425	0.555	0.498	0.456	0.533	0.589	0.546	0.485	0.588	0.698
w/o RFR	0.347	0.272	0.405	0.531	0.479	0.436	0.514	0.568	0.525	0.464	0.565	0.674
w/o Gate	0.359	0.283	0.417	0.546	0.489	0.446	0.524	0.580	0.536	0.475	0.577	0.686

Table 2: Ablation study of UBER on FB15k-237-MKG, WN18RR-MKG, and YAGO3-10-MKG under the 4KG–20% setting. *w/o Uncertainty* removes the uncertainty head and uses $\rho_{uv} = \mu_{uv}$ for entity-bridge reliability while keeping the remaining bridge construction and reasoning components unchanged.

For the reliability weights, the best overall performance is obtained around $\lambda_b = 0.5$ and $\lambda_c = 0.5$. Small values provide insufficient control over unreliable bridges or incompatible relations, whereas overly large values may suppress useful cross-KG evidence. Performance remains relatively stable around these optima, indicating that UBER is not sensitive to narrowly tuned hyperparameters.

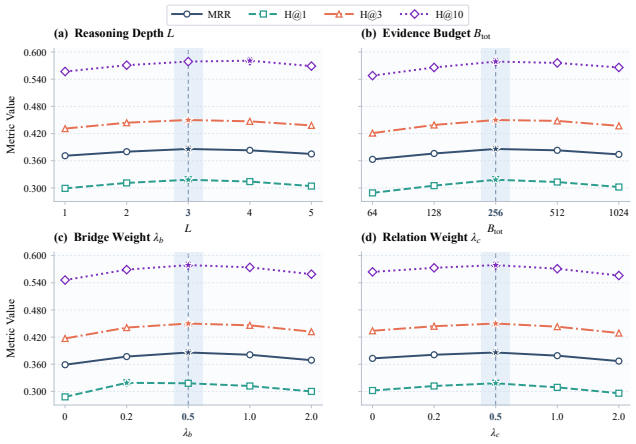

Figure 3: Hyperparameter sensitivity of UBER on FB15k-237-MKG under the 4KG–20% setting.

4.6 Further Experiments

Due to space limitations, the main paper focuses on the main comparison, ablation studies, and hyperparameter sensitivity analysis. The Appendix provides additional evaluations on benchmark validation, cross-KG scalability, bridge reliability and uncertainty analysis, robustness, efficiency, counterfactual faithfulness, interpretability, and implementation details.

5 Conclusion

We presented UBER, an uncertainty-aware bridge-structured evidence routing framework for multi-KG joint reasoning. Unlike alignment-first or global integration approaches, UBER constructs a compact query-specific evidence graph by dynamically selecting source KGs, allocating retrieval budgets, estimating entity- and relation-level bridge reliability, and controlling cross-KG propagation. Its three components: QSR, UBC, and RFR jointly enable query-aware, robust, and interpretable evidence composition without indiscriminately merging all source graphs.

Experiments on three constructed multi-KG completion benchmarks show that UBER consistently outperforms representative non-federated multi-KG, cross-KG, and federated baselines. The improvements remain stable with more source KGs, limited graph overlap, and incomplete or noisy bridges. Ablation, robustness, visualization, and counterfactual analyses further verify the effectiveness and faithfulness of its routing and bridge-aware reasoning mechanisms. Future work will extend UBER to open-world and multimodal multi-KG settings with more heterogeneous evidence sources.

AI USE STATEMENT

In this work, generative AI tools were used only for language polishing, including grammar correction, sentence refinement, and improving the clarity and readability of the manuscript. We did not use generative AI tools for research idea generation, methodology design, experimental implementation, data analysis, result interpretation, or scientific conclusion formulation. All AI-assisted revisions were manually reviewed and verified by the authors. We take full responsibility for the final content of the paper, including all scientific claims, results, and conclusions.

ETHICS STATEMENT

This work does not involve human-subject experiments, user studies, or the collection of new personal or sensitive data. Experiments are conducted on established knowledge graph benchmarks using constructed multi-KG partitions that preserve the original data splits and follow strict leakage-control protocols. As with other KG reasoning methods, UBER may inherit factual errors or biases from the underlying source graphs; therefore, real-world or high-stakes deployment should involve appropriate data validation and human oversight. We are not aware of additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce both UBER and the fragmented multi-KG benchmarks. The main paper specifies the model architecture, objectives, datasets, baselines, and evaluation protocol, while Appendices A–C provide the complete training procedure, observable feature construction, bridge calibration, benchmark-generation protocol, leakage controls, and dataset-specific hyperparameters.

REFERENCES

Muhammad Usman Akhtar, Jin Liu, Zhiwen Xie, Xiaohui Cui, Xiao Liu, and Bo Huang. Multilingual entity alignment by abductive knowledge reasoning on multiple knowledge graphs. *Engineering Applications of Artificial Intelligence*, 139:109660, 2025.

Deyu Chen, Hong Zhu, Jinguang Gu, Rusi Chen, and Meiyi Xie. Unaligned federated knowledge graph embedding. In *International Semantic Web Conference*, pp. 180–198. Springer, 2024.

Mingyang Chen, Wen Zhang, Zonggang Yuan, Yantao Jia, and Huajun Chen. Fede: Embedding knowledge graphs in federated setting. In *Proceedings of the 10th international joint conference on knowledge graphs*, pp. 80–88, 2021.

Mingyang Chen, Wen Zhang, Zonggang Yuan, Yantao Jia, and Huajun Chen. Federated knowledge graph completion via embedding-contrastive learning. *Knowledge-Based Systems*, 252:109459, 2022.

Xuelu Chen, Muhao Chen, Changjun Fan, Ankith Uppunda, Yizhou Sun, and Carlo Zaniolo. Multilingual knowledge graph completion via ensemble knowledge transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3227–3238, 2020.

Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, and Wei Wang. Multilingual knowledge graph completion with self-supervised adaptive graph alignment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 474–485, 2022.

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

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

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

Yifan Lu, Yigeng Zhou, Jing Li, Yequan Wang, Xuebo Liu, Daojing He, Fangming Liu, and Min Zhang. Knowledge editing with dynamic knowledge graphs for multi-hop question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 24741–24749, 2025.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *7th biennial conference on innovative data systems research*. CIDR Conference, 2014.

Cunli Mao, Xiaofei Gao, Ran Song, Shizhu He, Shengxiang Gao, Kang Liu, and Zhengtao Yu. Multilingual knowledge graph completion via efficient multilingual knowledge sharing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10882–10896, 2025.

ChangHao Men, Yu Han, Ping Wang, JianYuan Tao, and Cheng-Geng Huang. The interpretable reasoning and intelligent decision-making based on event knowledge graph with llms in fault diagnosis scenarios. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Hao Peng, Haoran Li, Yangqiu Song, Vincent W. Zheng, and Jianxin Li. Differentially private federated knowledge graphs embedding. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 1416–1425. ACM, 2021. doi: 10.1145/3459637.3482252.

Xiangqing Shen, Fanfan Wang, Zinong Yang, Bing Wang, Wenli Du, Chengqing Zong, and Rui Xia. Reason-align-respond: aligning llm reasoning with knowledge graphs for kgqa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jiawei Sheng, Taoyu Su, Xixun Lin, Xiaodong Li, and Tingwen Liu. Conditional diffusion guided knowledge transfer for multi-domain knowledge graph completion. In *Proceedings of the ACM Web Conference 2026*, pp. 3744–3754, 2026a.

Jiawei Sheng, Taoyu Su, Weiyi Yang, Linghui Wang, Yongxiu Xu, and Tingwen Liu. Information-theoretic minimal sufficient representation for multi-domain knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15707–15715, 2026b.

Harkanwar Singh, Soumen Chakrabarti, PRACHI JAIN, Sharod Roy Choudhury, et al. Multilingual knowledge graph completion with joint relation and entity alignment. In *3rd Conference on Automated Knowledge Base Construction*, 2021.

Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao Yu, and Jun Zhao. Multilingual knowledge graph completion from pretrained language models with knowledge constraints. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7709–7721, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.488. URL https://aclanthology.org/2023.findings-acl.488/.

Rongchuan Tang, Yang Zhao, Chengqing Zong, and Yu Zhou. Multilingual knowledge graph completion with language-sensitive multi-graph attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10508–10519, 2023.

Vinh Tong, Dat Quoc Nguyen, Trung Thanh Huynh, Tam Thanh Nguyen, Quoc Viet Hung Nguyen, and Mathias Niepert. Joint multilingual knowledge graph completion and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4646–4658, 2022.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1499–1509, 2015.

Shijie Wang, Wenqi Fan, Yue Feng, Lin Shanru, Xinyu Ma, Shuaiqiang Wang, and Dawei Yin. Knowledge graph retrieval-augmented generation for llm-based recommendation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27152–27168, 2025.

Bingbing Xie, Xiaoxiao Ma, Shan Xue, Jian Yang, Jia Wu, and Hao Fan. Contrastive multi-knowledge graph learning for fake news detection. *IEEE Transactions on Network Science and Engineering*, 2025.

Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 564–573, 2017.

Kai Zhang, Yu Wang, Hongyi Wang, Lifu Huang, Carl Yang, Xun Chen, and Lichao Sun. Efficient federated learning on knowledge graphs via privacy-preserving relation embedding aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 613–621, 2022.

Xiangrong Zhu, Guangyao Li, and Wei Hu. Heterogeneous federated knowledge graph embedding learning and unlearning. In *Proceedings of the ACM web conference 2023*, pp. 2444–2454, 2023.

Xiangrong Zhu, Yuexiang Xie, Yang Liu, Yaliang Li, and Wei Hu. Parameter-efficient federated knowledge graph embedding learning and unlearning. In *International Semantic Web Conference*, pp. 462–480. Springer, 2025.

A Complete Training Procedure

Algorithm 1 summarizes the training and inference procedure of UBER. During training, the positive target edge is masked before any query-specific graph reasoning operation. Hence, retrieval and message passing cannot directly access the fact being predicted.

B Additional Methodological Details

B.1 Observable Feature Construction

All graph-dependent features are constructed independently within each source KG. At inference time, they are computed from the available source training graphs. For a training query $q = (h, r_q, t)$, however, all query-dependent graph-derived quantities are computed from the masked support graph $\mathcal{G}_{-q}$ defined in Eq. (14).

In particular, relation-neighborhood profiles, local structural statistics, query-specific source summaries, and graph-dependent bridge features used by the current training query do not contain the masked positive edge. Purely textual features, such as entity names, descriptions, and relation names, are unaffected by this masking because they do not depend on graph edges.

Semantic representations. For entity v, we encode its observable surface name and description using a frozen text encoder f_{text}:

$$\mathbf{z}_v = \text{Norm}\left(\mathbf{W}_z f_{\text{text}}(\text{name}(v) \| \, \text{desc}(v)) + \mathbf{b}_z\right), \tag{17}$$

where only the surface name is used when no description is available.

A relation embedding combines a trainable representation with its textual representation:

$$\mathbf{e}_r = \text{LN}\left(\widetilde{\mathbf{e}}_r + \mathbf{W}_{rt} f_{\text{text}}(\text{name}(r))\right). \tag{18}$$

The text encoder remains frozen, whereas the projection parameters and $\widetilde{\mathbf{e}}_r$ are optimized jointly with UBER.

Algorithm 1 Training and Inference of UBER

Require: Source KGs $\{\mathcal{G}^{(m)}\}_{m=1}^{M}$, query $q = (h, r_q, ?)$, total budget B_{tot}, thresholds τ_e, τ_r, reasoning depth L
Ensure: Candidate ranking for query q
if training with target t **then**
  Identify the source s containing (h, r_q, t)
  Construct masked support graphs $\{\mathcal{G}_{-q}^{(m)}\}_{m=1}^{M}$ by removing (h, r_q, t) from source s
  if inverse edges are explicitly materialized **then**
    Also remove (t, r_q^{-1}, h)
  end if
else
  $\mathcal{G}_{-q}^{(m)} \leftarrow \mathcal{G}^{(m)}$ for all m
end if
Retrieve static semantic pre-candidates $\mathcal{C}_e^0$
Rerank $\mathcal{C}_e^0$ using graph-dependent features from the corresponding support graphs to obtain $\mathcal{C}_e(q)$
Compute source relevance $\{a_m\}_{m=1}^{M}$
Allocate integer source budgets $\{B_m\}_{m=1}^{M}$
for each source KG $\mathcal{G}_{-q}^{(m)}$ **do**
  Retrieve query-compatible evidence $\mathcal{G}_q^{(m)}$ under budget B_m
end for
Estimate entity-bridge confidence μ_{uv} and uncertainty ν_{uv}
Compute $\rho_{uv} = \mu_{uv} \exp(-\lambda_u \nu_{uv})$ and retain reliable entity bridges
Compute relation compatibility scores $\kappa_{rr'}$ and retain relation matches
Construct the query-specific fused graph $\mathcal{G}_q^{\text{fus}}$
for $\ell = 0, \ldots, L-1$ **do**
  Perform reliability-gated message passing
end for
Score candidate entities
if training **then**
  Construct observable candidate-based source and path targets from the masked support graphs
  Compute $\mathcal{L}_{\text{route}}$ and $\mathcal{L}_{\text{ret}}$
  Construct bridge pseudo-labels and compute $\mathcal{L}_{\text{br}}$
  Compute $\mathcal{L}_{\text{pred}}, \mathcal{L}_e, \mathcal{L}_r$
  Update parameters using the complete objective
end if
return Ranked candidate entities

Relation-neighborhood profile. Let $\mathcal{R}_{\text{out}}(v)$ and $\mathcal{R}_{\text{in}}(v)$ denote the multisets of outgoing and incoming relations of entity v. Its direction-aware relation profile is

$$\mathbf{p}_v = \text{Norm}\left(\mathbf{W}_p\Big[\text{Mean}_{r\in\mathcal{R}_{\text{out}}(v)}\,\mathbf{e}_r;\ \text{Mean}_{r\in\mathcal{R}_{\text{in}}(v)}\,\mathbf{e}_r\Big] + \mathbf{b}_p\right). \tag{19}$$

An empty incoming or outgoing neighborhood is represented by a zero vector.

Local structural information. For each entity, we construct

$$\mathbf{d}_v = \left[\log(1+\deg_{\text{in}}(v)); \log(1+\deg_{\text{out}}(v)); \log(1+\overline{\deg}_{\mathcal{N}(v)}); \frac{|\mathcal{R}(v)|}{|\mathcal{R}^{(m)}|}\right], \tag{20}$$

where $\overline{\deg}_{\mathcal{N}(v)}$ is the average degree of v's one-hop neighbors and $\mathcal{R}(v)$ is the set of distinct incident relations. Each dimension is standardized within its source KG.

The structural similarity between two cross-KG entities is

$$\text{sim}_{\text{str}}(u, v) = \exp\left(-\frac{\|\mathbf{d}_u - \mathbf{d}_v\|_1}{\tau_s}\right). \tag{21}$$

The final observable entity input is

$$\mathbf{x}_v = [\mathbf{z}_v; \mathbf{p}_v; \mathbf{d}_v]. \tag{22}$$

B.2 Candidate Generation and Discrete Router Supervision

Two-stage observable candidate generation. To avoid exhaustive cross-KG comparison while preventing the positive training edge from entering graph-dependent candidate features, we use a two-stage candidate-generation procedure.

First, we construct a static high-recall semantic pre-candidate pool offline using only edge-independent textual representations. For entity u in source KG i and entity v in source KG $j \neq i$, we define

$$\gamma_{uv}^{\mathrm{sem}} = \cos(\mathbf{z}_u, \mathbf{z}_v), \tag{23}$$

and retrieve the top-K_0 entities from every other source KG:

$$\mathcal{C}_e^0 = \left\{(u,v) : v \in \mathrm{TopK}_{K_0}^{\mathrm{sem}}(u) \ \vee \ u \in \mathrm{TopK}_{K_0}^{\mathrm{sem}}(v)\right\}. \tag{24}$$

Because $\mathbf{z}$ depends only on entity names and descriptions, $\mathcal{C}_e^0$ is independent of graph connectivity and can be precomputed once.

For each query, candidates in $\mathcal{C}_e^0$ are reranked using observable semantic and relation-profile information:

$$\gamma_{uv}(q) = \cos(\mathbf{z}_u, \mathbf{z}_v) + \beta_p \cos\left(\mathbf{p}_u^{-q}, \mathbf{p}_v^{-q}\right), \tag{25}$$

where $\mathbf{p}^{-q}$ is computed from the query-specific support graph. During training, this is the target-masked graph $\mathcal{G}_{-q}$; during inference, it is the available source graph.

The final candidate pool is obtained by retaining the top-K_c reranked candidates:

$$\mathcal{C}_e(q) = \left\{(u,v) \in \mathcal{C}_e^0 : v \in \mathrm{TopK}_{K_c}^{q}(u) \ \vee \ u \in \mathrm{TopK}_{K_c}^{q}(v)\right\}. \tag{26}$$

Exact surface-form matches are not deterministically inserted into either candidate pool, and latent canonical mappings, validation triples, and test triples are never used.

For routing supervision, we further define

$$\mathcal{C}_e^{\mathrm{rec}}(q) = \left\{(u,v) \in \mathcal{C}_e(q) : v \in \mathrm{TopK}_{K_c}^{q}(u) \ \wedge \ u \in \mathrm{TopK}_{K_c}^{q}(v)\right\}. \tag{27}$$

Observable training targets for source routing. Consider a training query $q = (h, r_q, t)$ whose observed positive triple belongs to source KG s. All routing supervision is constructed from the already masked support graphs $\{\mathcal{G}_{-q}^{(m)}\}_{m=1}^{M}$.

The source-local target $t^{(s)}$ is known only as the supervision label. For another source $m \neq s$, observable pseudo-targets are obtained from the coarse cross-KG candidate pool rather than from latent canonical mappings. We then search for supervision paths of at most H hops entirely inside the masked support graphs. For the source containing the observed training triple, we define

$$\mathcal{T}_s(q) = \{t^{(s)}\}. \tag{28}$$

For another source KG $m \neq s$, the model is not given the true source-local copy of $t^{(s)}$. Instead, we construct an observable pseudo-target set using only the reciprocal coarse candidates:

$$\mathcal{T}_m(q) = \left\{v \in \mathcal{E}^{(m)} : (t^{(s)}, v) \in \mathcal{C}_e^{\mathrm{rec}}(q)\right\}. \tag{29}$$

Hence, $\mathcal{T}_m(q)$ may contain zero, one, or multiple candidate counterparts. Importantly, it is determined entirely by the same observable semantic and relation-profile features available at inference time.

We then define the set of training-only supervision paths

$$\Pi_m^H(q) = \left\{\pi : \mathcal{A}^{(m)}(q) \rightsquigarrow v, \ v \in \mathcal{T}_m(q), \ |\pi| \leq H\right\}, \tag{30}$$

where all edges in π belong to the training graph after temporarily masking the target training triple.

No latent canonical entity mapping is accessed when constructing $\mathcal{T}_m(q)$ or $\Pi_m^H(q)$. Thus, cross-source routing supervision relies on observable candidate matching rather than ground-truth cross-KG identity correspondences.

Source-level routing supervision. For a path $\pi \in \Pi_m^H(q)$, we measure its compatibility with the query relation by

$$\psi(\pi, r_q) = \frac{1}{|\pi|} \sum_{r \in \pi} \cos\left(\bar{\mathbf{e}}_r, \bar{\mathbf{e}}_{r_q}\right), \tag{31}$$

where $\bar{\mathbf{e}}_r$ denotes a fixed relation representation used only for constructing training targets.

We define the set of sources with valid observable supervision paths as

$$\mathcal{S}_q = \left\{m : \Pi_m^H(q) \neq \emptyset\right\}. \tag{32}$$

For each feasible source $m \in \mathcal{S}_q$, its utility is

$$u_m(q) = \max_{\pi \in \Pi_m^H(q)} \left[\psi(\pi, r_q) - \lambda_h |\pi|\right]. \tag{33}$$

The target source distribution is normalized only over sources with valid supervision evidence:

$$\widehat{a}_m(q) = \begin{cases} \dfrac{\exp(u_m(q)/\tau_a)}{\sum_{j \in \mathcal{S}_q} \exp(u_j(q)/\tau_a)}, & m \in \mathcal{S}_q, \\ 0, & m \notin \mathcal{S}_q. \end{cases} \tag{34}$$

The source-routing loss is

$$\mathcal{L}_{\text{route}} = -\sum_{m=1}^{M} \widehat{a}_m(q) \log a_m(q). \tag{35}$$

If $\mathcal{S}_q = \emptyset$, we set $\mathcal{L}_{\text{route}} = \mathcal{L}_{\text{ret}} = 0$ for that query. The query remains supervised by $\mathcal{L}_{\text{pred}}$ and the other applicable objectives.

Path-level retrieval supervision. For each feasible source $m \in \mathcal{S}_q$, let

$$\pi_m^\star(q) = \arg \max_{\pi \in \Pi_m^H(q)} \left[\psi(\pi, r_q) - \lambda_h |\pi|\right] \tag{36}$$

denote its highest-utility supervision path.

The query-aware transition model is optimized with

$$\mathcal{L}_{\text{ret}} = -\sum_{m \in \mathcal{S}_q} \frac{1}{|\pi_m^\star(q)|} \sum_{(u,r,v) \in \pi_m^\star(q)} \log P_q^{(m)}(u, r, v). \tag{37}$$

The complete evidence-selection objective is

$$\mathcal{L}_{\text{sel}} = \mathcal{L}_{\text{route}} + \lambda_{\text{ret}} \mathcal{L}_{\text{ret}}. \tag{38}$$

This formulation directly supervises the source-relevance and relation-transition scoring functions while preserving hard integer budgets and top-B_m extraction at inference time. No policy-gradient estimator, straight-through estimator, or soft relaxation of the discrete selection operator is required.

B.3 Bridge Calibration and Relation Prototype Construction

Bridge confidence and uncertainty. For each $(u, v) \in \mathcal{C}_e(q)$, UBC uses Eq. 5 and predicts

$$\mu_{uv} = \sigma(\mathbf{w}_\mu^\top \boldsymbol{\phi}_{uv} + b_\mu), \qquad \nu_{uv} = \text{softplus}(\mathbf{w}_\nu^\top \boldsymbol{\phi}_{uv} + b_\nu). \tag{39}$$

To calibrate them without accessing latent canonical mappings, we construct observable pseudo-labels from the training data. A candidate pair is treated as a positive pseudo-match if it is a reciprocal top-1 semantic match and its similarity exceeds threshold τ_p. Other top-K_c candidates are treated as hard negatives and are supplemented with uniformly sampled cross-KG negatives.

Let $y_{uv} \in \{0, 1\}$ denote the pseudo-label. The uncertainty-aware bridge calibration objective is

$$\mathcal{L}_{\text{br}} = \frac{1}{|\mathcal{P}_{\text{br}}|} \sum_{(u,v) \in \mathcal{P}_{\text{br}}} \left[\exp(-\nu_{uv}) \, \text{BCE}(y_{uv}, \mu_{uv}) + \nu_{uv}\right]. \tag{40}$$

The first term trains matching confidence, while its uncertainty-dependent weight reduces the influence of ambiguous pseudo-labels. The additive term prevents uncertainty from increasing without penalty.

Relation structural summary. For relation r, let

$$\mathcal{H}_r = \{h : (h, r, t) \in \mathcal{T}_{\text{tr}}^{(m)}\}, \qquad \mathcal{T}_r = \{t : (h, r, t) \in \mathcal{T}_{\text{tr}}^{(m)}\}.$$

We define relation functionality and inverse functionality as

$$\text{fun}(r) = \frac{|\mathcal{H}_r|}{n_r}, \qquad \text{ifun}(r) = \frac{|\mathcal{T}_r|}{n_r}, \tag{41}$$

where $n_r = |\{(h, r, t) \in \mathcal{T}_{\text{tr}}^{(m)}\}|$.

Its source-local structural summary is

$$\mathbf{s}_r = \text{LN}\left(\mathbf{W}_{sr}\Big[\text{Mean}_{h\in\mathcal{H}_r}\,\mathbf{u}_h; \text{Mean}_{t\in\mathcal{T}_r}\,\mathbf{u}_t; \log(1+n_r); \text{fun}(r); \text{ifun}(r)\Big] + \mathbf{b}_{sr}\right). \tag{42}$$

The relation prototype $\mathbf{c}_r = \text{MLP}([\mathbf{e}_r; \mathbf{s}_r])$ therefore captures both relation semantics and source-local usage patterns.

B.4 Complete Reasoning and Optimization Details

Attention normalization. Given the reliability-aware logit $e_{uv}^{(\ell)}$ in Eq. 11, the normalized attention coefficient is

$$\alpha_{uv}^{(\ell)} = \frac{\exp(e_{uv}^{(\ell)})}{\sum_{u'\in\mathcal{N}(v)} \exp(e_{u'v}^{(\ell)})}. \tag{43}$$

The message aggregated at entity v is

$$\mathbf{m}_v^{(\ell)} = \sum_{u\in\mathcal{N}(v)} \alpha_{uv}^{(\ell)} \mathbf{W}_V [\mathbf{h}_u^{(\ell)}; \mathbf{e}_r], \tag{44}$$

which leads to the residual update in Eq. 12.

Entity consistency. Reliable cross-KG entity correspondences are regularized by

$$\mathcal{L}_e = \sum_{(u,v,\rho_{uv})\in\mathcal{B}_e^q} \rho_{uv} \left\| \mathbf{h}_u^{(L)} - \mathbf{h}_v^{(L)} \right\|_2^2. \tag{45}$$

The reliability weight prevents uncertain correspondences from being strongly forced into the same representation.

Relation consistency. For retained relation matches,

$$\mathcal{L}_r = \sum_{(r,r',\kappa_{rr'})\in\mathcal{B}_r} \kappa_{rr'} \left\| \mathbf{c}_r - \mathbf{c}_{r'} \right\|_2^2. \tag{46}$$

The compatibility weight similarly encourages consistent relations to remain close without imposing hard alignment on uncertain relation pairs.

Complete objective. Substituting the auxiliary objectives into Eq. 16 gives

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{sel}} (\mathcal{L}_{\text{route}} + \lambda_{\text{ret}} \mathcal{L}_{\text{ret}}) + \lambda_{\text{br}} \mathcal{L}_{\text{br}} \\ + \lambda_e \mathcal{L}_e + \lambda_r \mathcal{L}_r.\end{aligned} \tag{47}$$

All differentiable parameters are optimized jointly. The hard budget allocation and retrieved node sets are treated as discrete decisions, while their underlying scoring functions are trained through the explicitly supervised selection objective.

Dataset	#Ent.	#Rel.	Train	Valid	Test
FB15k-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134
YAGO3-10	123,182	37	1,079,040	5,000	5,000

Table 3: Statistics of the original benchmarks used to construct the fragmented multi-KG settings.

C Experiments

C.1 Dataset Statistics

Table 3 summarizes the statistics of the constructed multi-KG benchmarks. The numbers of entities, relations, and triples follow the corresponding original datasets, while the number of source KGs and overlap ratios vary across experimental settings.

C.2 Canonical Evaluation and Ranking Protocol

Although each source KG maintains independent local entity identifiers, entities originating from the same canonical entity are connected through the benchmark construction process. For evaluation, we report link prediction performance over canonical entities rather than individual source-specific copies.

For a test query $(h, r, ?)$, the query entity is first mapped to its corresponding source-local representation. After obtaining prediction scores over all source-local candidate entities, we aggregate scores belonging to the same canonical entity.

Specifically, for a canonical entity e, let $\mathcal{S}_e$ denote the set of its source-specific copies. The final score of e is computed as

$$s(q, e) = \max_{v \in \mathcal{S}_e} s(q, v), \tag{48}$$

where $s(q, v)$ denotes the prediction score assigned to a source-local entity copy v.

This max aggregation follows the assumption that multiple source copies provide alternative evidence views of the same underlying entity. It prevents duplicated source representations from artificially lowering the rank of the correct canonical answer.

We then perform filtered ranking over canonical entities. For each test query, all other known valid answers sharing the same query relation are removed before computing the rank of the target entity. Formally, the filtered candidate set is

$$\mathcal{E}_{\text{filter}} = \mathcal{E} \setminus \{e' : (h, r, e') \in \mathcal{G}_{train} \cup \mathcal{G}_{valid} \cup \mathcal{G}_{test}\}. \tag{49}$$

The resulting canonical ranking is used to compute MRR and Hits@K.

During testing, the canonical entity mappings are only used by the evaluation script and are never provided to UBER or any compared method during training or inference. Therefore, all models operate solely on source-local entity identifiers and observable source KG information.

C.3 Construction of Fragmented Multi-KG Benchmarks

We construct fragmented multi-KG benchmarks from FB15k-237, WN18RR, and YAGO3-10 while preserving their official training, validation, and test splits. Given $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}_{\text{tr}}, \mathcal{T}_{\text{va}}, \mathcal{T}_{\text{te}})$, we partition only $\mathcal{T}_{\text{tr}}$ into $M \in \{2, 3, 4, 5\}$ source KGs.

Overlap and initial partitioning. Let $S_e(v)$ and $S_r(r)$ denote the source memberships of entity v and relation r. The realized entity and relation overlap ratios are

$$\rho_e^{\text{act}} = \frac{|\{v : |S_e(v)| \geq 2\}|}{|\mathcal{E}|}, \qquad \rho_r^{\text{act}} = \frac{|\{r : |S_r(r)| \geq 2\}|}{|\mathcal{R}|}. \tag{50}$$

Dataset	Seed	$\lvert\mathcal{E}_{\mathrm{sh}}\rvert$	ρ_e^{act}	$\lvert\mathcal{R}_{\mathrm{sh}}\rvert$	ρ_r^{act}	Miss.	Dup.	Inf.	E-Mem.	R-Mem.
FB15k-237-MKG	202701	2,908	19.999%	95	40.084%	0	0	0	0	0
WN18RR-MKG	202702	8,189	20.001%	4	36.364%	0	0	0	0	0
YAGO3-10-MKG	202703	24,636	20.000%	15	40.541%	0	0	0	0	0

Table 4: Constraint audit for the main 4KG–20% setting. E-Mem. and R-Mem. denote entity- and relation-membership violations. All accepted partitions contain no missing, duplicated, infeasible, or membership-violating assignments. The deviation of ρ_r^{act} from 40% is caused by integer granularity of the relation vocabulary.

We vary the target entity overlap as $\rho_e \in \{10\%, 20\%, 30\%, 40\%\}$ and use a nominal relation overlap of $\rho_r = 40\%$. The corresponding numbers of shared entities and relations are obtained by nearest-integer rounding. Hence, the realized ratio can differ slightly from the nominal target for datasets with few relations; e.g., WN18RR contains 11 relations, for which the closest realization to 40% is $4/11 = 36.36\%$.

Each shared entity belongs to exactly two source KGs, while a non-shared entity belongs to one. Non-shared relations remain source-specific, whereas shared relations may occur in multiple sources. We first apply METIS 5.1.0 (`METIS_PartGraphKway`) to the undirected training graph and then select shared entities and relations by degree- and frequency-stratified sampling, respectively. The main 4KG–20% settings use construction seeds 202701, 202702, and 202703 for FB15k-237-MKG, WN18RR-MKG, and YAGO3-10-MKG.

Constraint-aware triple allocation. A training triple $\tau = (h, r, t)$ can be assigned to source m only if all three components occur in that source. Its feasible-source set is therefore

$$\mathcal{F}(\tau) = S_e(h) \cap S_r(r) \cap S_e(t), \tag{51}$$

and a valid partition requires $\mathcal{F}(\tau) \neq \emptyset$ for every $\tau \in \mathcal{T}_{\mathrm{tr}}$.

If the initial memberships violate this condition, we perform quota-preserving membership exchanges over the head entity, relation, and tail entity. A repair is accepted only if it preserves the target numbers of shared entities and relations, keeps every entity in at most two sources, does not invalidate previously feasible incident triples, and reduces the number of infeasible triples. When no admissible repair exists, the initialization is discarded and construction is restarted with the next deterministic seed.

After all triples become feasible, each triple is assigned to the least-loaded source in $\mathcal{F}(\tau)$. Consequently,

$$\mathcal{T}_{\mathrm{tr}}^{(i)} \cap \mathcal{T}_{\mathrm{tr}}^{(j)} = \emptyset \ (i \neq j), \qquad \bigcup_{m=1}^{M} \mathcal{T}_{\mathrm{tr}}^{(m)} = \mathcal{T}_{\mathrm{tr}}. \tag{52}$$

Thus, source KGs contain complementary training evidence while their union exactly preserves the original training graph.

Construction audit. Every generated partition is accepted only after an explicit invariant check verifying that (i) every training triple occurs in exactly one source, (ii) every assigned triple satisfies its entity/relation memberships, (iii) the target sharing counts are met, (iv) entity and relation membership constraints are satisfied, and (v) no validation or test triple is included in a source training graph. Let N_{miss}, N_{dup}, and N_{inf} denote the numbers of missing, duplicated, and infeasible training triples, respectively. All partitions used in our experiments satisfy

$$N_{\mathrm{miss}} = N_{\mathrm{dup}} = N_{\mathrm{inf}} = 0. \tag{53}$$

Per-source statistics. Table 5 reports the resulting source sizes under the main 4KG–20% setting. The sources remain balanced while preserving complementary evidence.

Validation, testing, and training-time leakage control. Only training triples are inserted into the source KGs. Validation and test triples are never used for source partitioning, graph-dependent feature construction, bridge candidate generation, or message passing and are used only as evaluation queries.

Dataset	Source	#Entities	#Relations	#Train Triples
FB15k-237-MKG	KG1	4,365	131	68,070
	KG2	4,360	131	67,996
	KG3	4,361	130	68,034
	KG4	4,363	130	68,015
WN18RR-MKG	KG1	12,290	6	21,730
	KG2	12,274	6	21,689
	KG3	12,283	6	21,712
	KG4	12,285	5	21,704
YAGO3-10-MKG	KG1	36,970	21	269,850
	KG2	36,942	21	269,694
	KG3	36,956	20	269,781
	KG4	36,950	20	269,715

Table 5: Per-source statistics after constraint-aware construction under the main 4KG–20% setting. Shared entities and relations are counted independently in each source, whereas training triples are disjoint across sources.

During optimization, each positive training triple is treated as the label of one training query. Before its forward pass, that triple is temporarily removed from its source support graph, together with its explicitly materialized inverse edge when applicable. All query-specific graph operations—including graph-derived feature construction, source routing, evidence retrieval, supervision-path search, bridge construction, fused-graph reasoning, and message passing—therefore operate without access to the target edge.

The removed triple is used only for computing the prediction objective after candidate scores have been produced. Cross-source routing targets are derived from observable candidates and never from latent canonical entity mappings.

C.4 Implementation Details and Hyperparameter Settings

Hardware and software environment. All experiments are conducted on a Linux server equipped with four NVIDIA RTX A6000 GPUs, each with 48 GB of memory (192 GB aggregate GPU memory). UBER is implemented in PyTorch and trained with CUDA acceleration. Our reference environment uses Python 3.10, PyTorch 2.4.1, CUDA 12.1, and cuDNN 9.x. All package versions are pinned in the released environment file.

Unless otherwise stated, a single experimental run uses one GPU, while independent random-seed runs and hyperparameter configurations are distributed across the four A6000 GPUs in parallel. No model or experiment requires model-parallel execution across GPUs.

Textual representations. Entity names, entity descriptions, and relation names are encoded with a frozen pretrained text encoder BERT. The encoder parameters remain fixed throughout training, and only the task-specific projection layers are optimized. All experiments use the same encoder checkpoint and preprocessing pipeline across datasets and baselines whenever textual features are required. Descriptions are obtained from the corresponding benchmark metadata; when no description is available, only the surface name is used.

Optimization. All learnable parameters are initialized with Xavier initialization and optimized using Adam. We train for at most 500 epochs and apply early stopping according to validation MRR with a patience of 30 epochs. The gradient norm is clipped to 1.0, and $\epsilon = 10^{-8}$ is used in logarithmic and normalization operations for numerical stability. Negative entities are sampled by randomly corrupting the head or tail entity while filtering known positive triples.

For each dataset, all model-selection decisions are made using the validation set only. Dataset-specific configurations are selected under the main 4KG–20% setting and then reused for experiments with different numbers of source KGs, overlap ratios, bridge perturbations, and scalability settings.

Randomness and repeated runs. Each reported model configuration is trained with five independent random seeds. The seeds control parameter initialization, mini-batch ordering, negative

Hyperparameter	FB15k-237-MKG	WN18RR-MKG	YAGO3-10-MKG
Training and representation			
Hidden dimension d	256	256	512
Learning rate	1×10^{-3}	5×10^{-4}	5×10^{-4}
Batch size	1024	1024	2048
Negative samples per query	256	256	512
Dropout rate	0.20	0.20	0.30
Query-aware routing and retrieval			
Reasoning depth L	3	2	4
Maximum supervision-path length H	3	2	4
Total evidence budget B_{tot}	256	512	1024
Minimum per-KG budget B_{min}	8	8	16
Retrieval temperature τ_q	0.50	0.50	0.50
Routing-target temperature τ_a	0.50	0.50	0.50
Path-length penalty λ_h	0.10	0.10	0.10
Retrieval supervision weight λ_{ret}	0.50	0.50	0.50
Selection loss weight λ_{sel}	0.10	0.10	0.05
Bridge construction and calibration			
Entity-bridge threshold τ_e	0.60	0.70	0.60
Relation-bridge threshold τ_r	0.50	0.40	0.60
Uncertainty coefficient λ_u	0.50	0.50	0.50
Coarse candidates per source K_c	20	20	30
Structural temperature τ_s	1.0	1.0	1.0
Pseudo-positive threshold τ_p	0.85	0.80	0.75
Relation-profile weight β_p	0.50	0.60	0.50
Bridge calibration weight λ_{br}	0.10	0.15	0.20
Reliability-gated reasoning			
Bridge reliability weight λ_b	0.50	1.00	0.20
Relation compatibility weight λ_c	0.50	0.20	1.00
Entity consistency weight λ_e	0.10	0.10	0.05
Relation consistency weight λ_r	0.10	0.05	0.10

Table 6: Dataset-specific hyperparameter settings of UBER. The routing-related parameters include the maximum training-only supervision-path length H, routing-target temperature τ_a, path-length penalty λ_h, retrieval supervision weight λ_{ret}, and overall selection-loss weight λ_{sel}.

sampling, and other stochastic training operations. Unless otherwise specified, tables report the arithmetic mean over the five runs. Benchmark construction uses separate fixed seeds from model-training seeds, so model stochasticity and benchmark-partition stochasticity are controlled independently.

Benchmark generation. The fragmented multi-KG benchmarks are generated by a deterministic pipeline given the construction seed. For every setting, we store the source-membership assignments, final entity and relation mappings, source-local training triples, overlap statistics, and constraint-audit outputs. The benchmark generator verifies that every training triple is assigned exactly once, no infeasible triple remains, and the requested entity and relation overlap constraints are satisfied before a partition is accepted.

Canonical entity and relation mappings are retained only by the benchmark generator and evaluator. They are never exposed to UBER or the compared models during training or inference.

Baseline reproduction. For each baseline, we use the authors' released implementation whenever available and follow its recommended optimization settings. Dataset-dependent hyperparameters are selected using validation MRR under the same fragmented multi-KG protocol as UBER. All methods receive the same source partitions, training/validation/test splits, and observable input information. No baseline is given access to latent canonical entity or relation mappings unless such information is also available to UBER under the corresponding experimental setting.

The final UBER hyperparameters are summarized in Table 6.

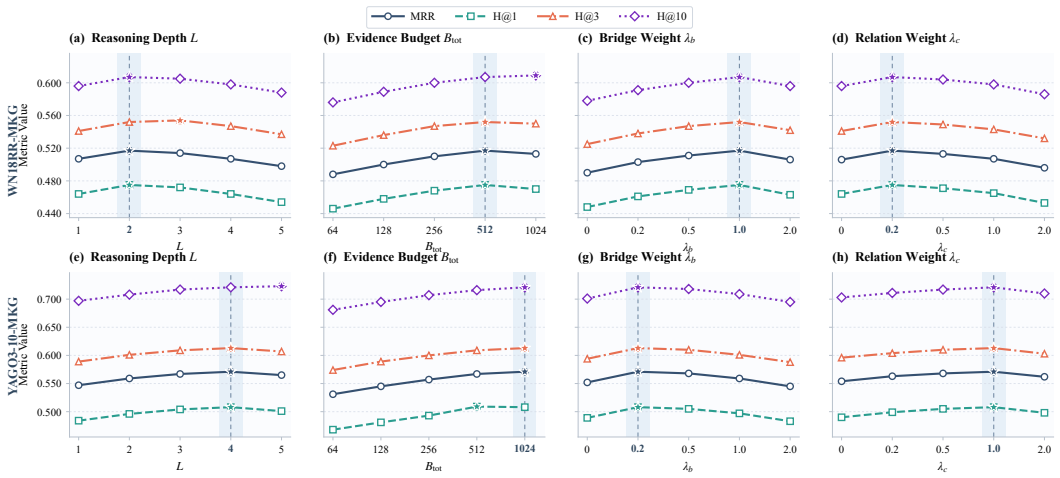

Figure 4: Additional hyperparameter sensitivity results on WN18RR-MKG (top row) and YAGO3-10-MKG (bottom row) under the 4KG–20% setting. The four columns correspond to the reasoning depth L, total evidence budget B_{tot}, bridge reliability weight λ_b, and relation compatibility weight λ_c. Shaded regions indicate configurations with the best overall performance, while stars denote metric-specific optima.

D Additional Hyperparameter Sensitivity Results

To complement the sensitivity analysis on FB15k-237-MKG in the main paper, we report additional results on WN18RR-MKG and YAGO3-10-MKG. All experiments follow the same 4KG–20% setting and evaluation protocol. Figure 4 presents the results, with WN18RR-MKG shown in the top row and YAGO3-10-MKG in the bottom row.

On WN18RR-MKG, the best overall configuration uses a relatively shallow reasoning depth of $L = 2$. Increasing the depth beyond this point provides little benefit and eventually degrades performance, suggesting that useful evidence can be captured through short reasoning paths on this benchmark. The model performs best overall with $B_{\text{tot}} = 512$, while Hits@10 slightly benefits from a larger budget. This indicates that additional evidence may improve broad candidate coverage, although the moderate budget yields better overall ranking quality. WN18RR-MKG also favors stronger bridge reliability control ($\lambda_b = 1.0$) and a relatively small relation compatibility weight ($\lambda_c = 0.2$), showing that reliable entity connections play a particularly important role in this setting.

YAGO3-10-MKG exhibits a different sensitivity pattern. Its best overall performance is obtained with a deeper reasoning network ($L = 4$) and the largest tested evidence budget ($B_{\text{tot}} = 1024$), reflecting the need to aggregate evidence from a larger and more structurally diverse graph. Hits@10 slightly favors an additional reasoning layer, while Hits@1 reaches its maximum at a moderately smaller evidence budget. Nevertheless, the selected settings provide the strongest joint performance across all four metrics. For cross-KG propagation, YAGO3-10-MKG favors a relatively small bridge weight ($\lambda_b = 0.2$) but a stronger relation compatibility weight ($\lambda_c = 1.0$), suggesting that relation-level semantic consistency is particularly useful for controlling evidence transfer in this larger benchmark.

Overall, the three datasets exhibit different optimal configurations, but share a consistent pattern: insufficient routing or reasoning capacity limits evidence utilization, whereas excessive propagation, retrieval, or reliability constraints introduce noise or suppress useful information. The smooth trends around the selected configurations further confirm that UBER remains stable across a broad range of practical hyperparameter choices.

Dataset	Reliability	Prec.	Rec.	F1	ECE↓	Brier↓	MRR	H@1
FB15k-237-MKG	$\rho = \mu$	.824	.700	.757	.074	.132	.378	.307
	$\rho = \mu e^{-\lambda_u \nu}$	**.857**	**.728**	**.788**	**.041**	**.105**	**.386**	**.318**
WN18RR-MKG	$\rho = \mu$	.851	.749	.797	.061	.118	.509	.466
	$\rho = \mu e^{-\lambda_u \nu}$	**.879**	**.774**	**.823**	**.035**	**.094**	**.517**	**.475**
YAGO3-10-MKG	$\rho = \mu$	.816	.669	.735	.083	.147	.562	.497
	$\rho = \mu e^{-\lambda_u \nu}$	**.849**	**.696**	**.765**	**.047**	**.116**	**.571**	**.508**

Table 7: Controlled comparison of confidence-only and uncertainty-aware bridge reliability under the 4KG–20% setting. Both variants retain identical bridge budgets, and latent mappings are used only for post-hoc bridge evaluation.

Reliability	0%	10%	20%	30%	50%
$\rho = \mu$	.378	.372	.365	.350	.305
$\rho = \mu e^{-\lambda_u \nu}$	**.386**	**.384**	**.381**	**.371**	**.340**

Table 8: MRR under injected bridge noise on FB15k-237-MKG. Both variants use matched retained-bridge counts.

E Effect of Explicit Bridge Uncertainty

UBER defines bridge reliability as

$$\rho_{uv} = \mu_{uv} \exp(-\lambda_u \nu_{uv}), \tag{54}$$

where μ_{uv} denotes matching confidence and ν_{uv} captures bridge uncertainty. To evaluate whether ν_{uv} provides information beyond confidence alone, we compare UBER with a controlled variant using $\rho_{uv} = \mu_{uv}$. The uncertainty head is removed in this variant, while all other components and training settings remain unchanged.

To eliminate the effect of different bridge quantities, we adopt a matched-retention protocol. For each query, both variants retain the same number of entity bridges; the confidence-only variant ranks bridges by μ_{uv}, whereas UBER uses $\mu_{uv} \exp(-\lambda_u \nu_{uv})$. Latent entity correspondences from benchmark construction are used only for post-hoc evaluation. We report bridge retrieval quality (Precision, Recall, and F1), calibration metrics (ECE and Brier), and downstream completion results. All calibration metrics are computed over the same candidate bridge pool.

Table 7 demonstrates that explicit uncertainty modeling improves both bridge selection and calibration. Under identical bridge budgets, UBER improves bridge F1 by 2.6–3.1 points and consistently reduces ECE and Brier scores. The corresponding gains in MRR and Hits@1 indicate that ν_{uv} provides complementary information beyond matching confidence.

We further evaluate robustness under injected bridge noise on FB15k-237-MKG. Noise is introduced only during inference following the same protocol as the bridge-robustness experiment.

As shown in Table 8, uncertainty-aware reliability becomes increasingly beneficial as bridge noise increases. The MRR improvement grows from 0.008 under clean bridges to 0.035 under 50% corruption, showing that ν_{uv} helps suppress plausible but unreliable correspondences.

Together, the matched-retention, calibration, and noise robustness results verify that the uncertainty head provides additional reliability signals beyond confidence estimation alone.

F Additional Experimental Results

F.1 Effect of Cross-KG Overlap

We further investigate the effect of cross-KG entity overlap by varying the overlap ratio from 10% to 40% while fixing the number of source KGs to four. Table 9 and Table 10 reports the results on all three benchmarks.

Method	10%		20%		30%		40%	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Non-federated multi-KG and cross-KG methods								
KEnS	0.303	0.482	0.321	0.506	0.333	0.521	0.338	0.528
AlignKGC	0.316	0.496	0.335	0.524	0.347	0.538	0.353	0.546
SS-AGA	0.320	0.501	0.339	0.529	0.349	0.542	0.356	0.550
JMAC	0.327	0.507	0.346	0.536	0.357	0.550	0.362	0.557
LSMGA	0.338	0.523	0.356	0.561	0.368	0.570	0.375	0.580
GC-PLM	0.335	0.514	0.353	0.547	0.370	0.565	0.378	0.573
KL-GMoE	0.346	0.532	0.359	0.558	0.372	0.574	0.381	0.583
IMKGC	0.354	0.548	0.369	0.567	0.379	0.581	0.392	**0.591**
DMKGC	0.362	0.550	0.376	0.565	0.384	0.580	0.389	0.588
Federated KGC methods								
FedE	0.288	0.468	0.307	0.491	0.319	0.504	0.326	0.512
FedR	0.305	0.486	0.324	0.510	0.337	0.525	0.343	0.532
FedEC	0.298	0.478	0.317	0.501	0.330	0.515	0.337	0.523
FKGE	0.310	0.493	0.331	0.519	0.341	0.531	0.349	0.540
FedLU	0.323	0.504	0.343	0.531	0.355	0.546	0.363	0.554
UniFE	0.317	0.499	0.336	0.526	0.351	0.541	0.367	0.560
PFLU	0.348	0.539	0.361	0.552	0.374	0.576	0.379	0.578
UBER	**0.374**	**0.560**	**0.386**	**0.579**	**0.391**	**0.586**	**0.394**	0.589

Table 9: Effect of the entity-overlap ratio on FB15k-237-MKG under the 4KG setting. The relation-overlap ratio is fixed at 40%. Best results are shown in bold and second-best results are underlined.

Method	WN18RR-MKG								YAGO3-10-MKG							
	10%		20%		30%		40%		10%		20%		30%		40%	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Non-federated multi-KG and cross-KG methods																
KEnS	0.431	0.520	0.452	0.541	0.463	0.554	0.469	0.561	0.478	0.646	0.501	0.671	0.516	0.688	0.525	0.699
AlignKGC	0.445	0.535	0.466	0.557	0.478	0.570	0.485	0.578	0.493	0.663	0.518	0.690	0.532	0.706	0.542	0.718
SS-AGA	0.448	0.538	0.468	0.560	0.480	0.573	0.487	0.581	0.497	0.667	0.521	0.692	0.535	0.709	0.545	0.721
JMAC	0.443	0.533	0.463	0.554	0.475	0.567	0.482	0.575	0.504	0.674	0.528	0.699	0.542	0.716	0.552	0.728
LSMGA	0.459	0.562	0.481	0.585	0.493	0.608	0.501	0.619	0.522	0.691	0.548	0.715	0.563	0.731	0.574	0.742
GC-PLM	0.467	0.558	0.488	0.579	0.500	0.593	0.507	0.602	0.529	0.690	0.554	0.714	0.569	0.730	0.580	0.741
KL-GMoE	0.472	0.561	0.492	0.582	0.504	0.596	0.511	0.605	0.531	0.696	0.552	0.718	0.567	0.735	0.578	0.748
IMKGC	0.485	0.565	0.504	0.583	0.514	0.601	0.521	0.611	0.539	0.699	0.559	0.719	0.573	0.738	0.584	0.752
DMKGC	0.492	0.568	0.510	0.584	0.519	0.603	0.525	0.613	0.547	0.707	0.565	**0.724**	0.579	0.741	0.590	0.756
Federated KGC methods																
FedE	0.416	0.507	0.438	0.528	0.449	0.541	0.456	0.549	0.466	0.631	0.489	0.656	0.504	0.673	0.513	0.684
FedR	0.440	0.529	0.461	0.551	0.473	0.564	0.480	0.572	0.492	0.657	0.516	0.682	0.530	0.699	0.540	0.711
FedEC	0.433	0.522	0.454	0.544	0.466	0.557	0.473	0.565	0.485	0.651	0.509	0.676	0.523	0.693	0.533	0.705
FKGE	0.425	0.514	0.446	0.536	0.458	0.549	0.465	0.557	0.478	0.644	0.501	0.669	0.516	0.686	0.525	0.697
FedLU	0.457	0.548	0.479	0.570	0.491	0.583	0.498	0.591	0.508	0.672	0.532	0.697	0.546	0.714	0.556	0.726
UniFE	0.450	0.541	0.472	0.563	0.484	0.576	0.491	0.584	0.517	0.679	0.541	0.704	0.555	0.721	0.565	0.733
PFLU	0.468	0.557	0.489	0.579	0.501	0.592	0.508	0.600	0.523	0.692	0.547	0.717	0.561	0.734	0.571	0.746
UBER	**0.501**	**0.584**	**0.517**	**0.607**	**0.525**	**0.616**	**0.530**	**0.620**	**0.556**	**0.711**	**0.571**	0.721	**0.584**	**0.744**	**0.594**	**0.760**

Table 10: Effect of the entity-overlap ratio on WN18RR-MKG and YAGO3-10-MKG under the 4KG setting. The relation-overlap ratio is fixed at 40%. Best results are shown in bold and second-best results are underlined.

As the overlap ratio increases, all methods generally benefit from stronger cross-KG connectivity and more accessible complementary evidence. However, the improvements gradually saturate, indicating that simply introducing more shared entities cannot fully resolve irrelevant or unreliable cross-KG propagation. UBER achieves the best MRR under every overlap ratio on all three benchmarks and obtains the strongest Hits@10 in most configurations. Its advantage is particularly evident in low-overlap settings, where reliable cross-KG links are scarce and conventional alignment or fusion methods struggle to identify useful evidence. This confirms that uncertainty-aware bridge construction enables UBER to exploit sparse correspondences while suppressing unreliable links. As the overlap ratio increases, the performance gap becomes smaller, and competing methods occasionally obtain slightly higher Hits@10. These isolated results suggest that dense entity overlap can improve broad candidate coverage. Nevertheless, UBER maintains more consistent overall ranking quality, demonstrating that query-aware source routing and reliability-gated reasoning remain beneficial even when abundant cross-KG connections are available.

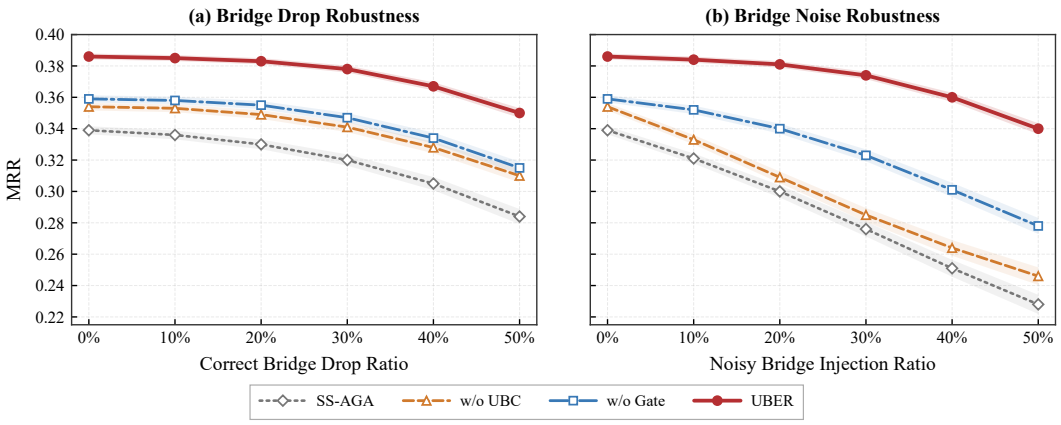

Figure 5: Robustness to incomplete and noisy cross-KG bridges on FB15k-237-MKG under the 4KG–20% setting. (a) MRR under correct-bridge dropping. (b) MRR under hard noisy-bridge injection. Shaded regions denote standard deviations over five random seeds.

F.2 Robustness to Incomplete and Noisy Bridges

We evaluate the robustness of UBER to imperfect cross-KG connectivity on FB15k-237-MKG under the 4KG–20% setting. For *bridge drop*, we randomly remove different proportions of correct entity bridges to simulate incomplete correspondences. For *bridge noise*, we inject hard negative bridges whose endpoints have similar surface forms or semantics but refer to different entities. The perturbation ratio ranges from 0% to 50%, and all perturbations are applied at inference time without retraining. We compare UBER with SS-AGA, w/o UBC, and w/o Gate, and report the average results over five random seeds.

As shown in Figure 5(a), UBER degrades more gracefully as correct bridges are removed. Its MRR decreases only from 0.386 to 0.383 at a 20% drop ratio and remains 0.350 when half of the bridges are unavailable, corresponding to a relative decrease of 9.3%. In comparison, SS-AGA drops by 16.2%. This result indicates that query-aware routing and relation-level compatibility enable UBER to better exploit the remaining cross-KG evidence under incomplete connectivity.

The advantage is more evident under bridge noise. As shown in Figure 5(b), UBER maintains an MRR of 0.381 at a 20% noise ratio and 0.340 at a 50% noise ratio, yielding a relative decline of 11.9%. By contrast, w/o Gate, w/o UBC, and SS-AGA decrease by 22.6%, 30.5%, and 32.7%, respectively. The sharper degradation of w/o UBC shows that uncertainty-aware bridge construction is essential for filtering unreliable correspondences, while the gap between UBER and w/o Gate confirms that propagation-stage reliability control remains necessary for suppressing residual noisy messages. Overall, UBC and RFR provide complementary robustness against corrupted cross-KG evidence.

F.3 Counterfactual Faithfulness Analysis

The visualized routing and bridge scores are meaningful only if they faithfully reflect the evidence used for prediction. We therefore conduct counterfactual interventions by selectively removing source KGs and cross-KG bridges according to their estimated importance. For source-level intervention, source KGs are ranked by the query-specific relevance scores produced by QSR. For bridge-level intervention, we rank each bridge (u, v) by its reliability-weighted reasoning contribution,

$$C^q_{uv} = \rho_{uv} \cdot \frac{1}{L} \sum_{\ell=1}^{L} \alpha^{(\ell)}_{uv}, \tag{55}$$

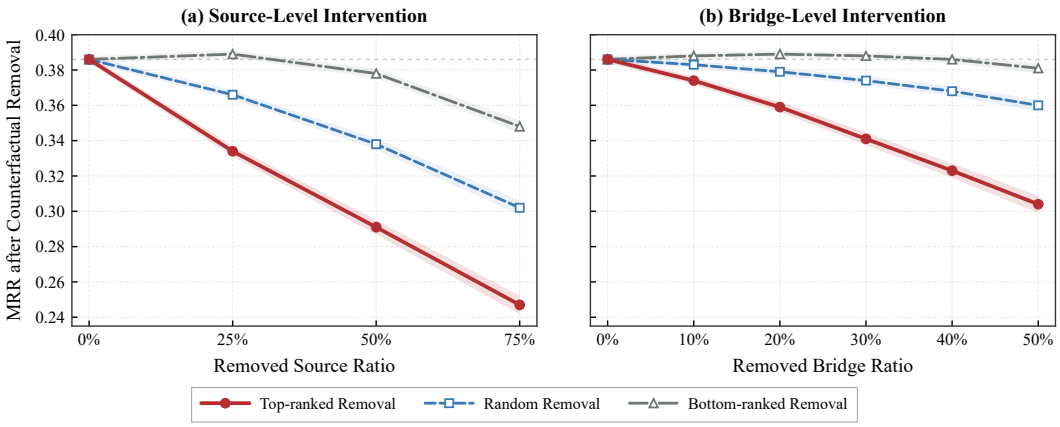

Figure 6: Counterfactual faithfulness analysis on FB15k-237-MKG under the 4KG–20% setting. (a) Source KGs are removed according to their QSR routing scores. (b) Cross-KG bridges are removed according to their reliability-weighted reasoning contributions. Shaded regions denote standard deviations over five random seeds.

where ρ_{uv} is the bridge reliability and $\alpha_{uv}^{(\ell)}$ is its attention weight at layer ℓ. We compare removing top-ranked, randomly selected, and bottom-ranked evidence on FB15k-237-MKG under the 4KG–20% setting.

As shown in Figure 6(a), removing highly ranked source KGs causes a substantially larger performance decrease than the other strategies. Removing the top-ranked source reduces MRR from 0.386 to 0.334, whereas random and bottom-ranked removal result in MRR values of 0.366 and 0.389, respectively. When half of the sources are removed, top-ranked removal further decreases MRR to 0.291, compared with 0.338 and 0.378 for random and bottom-ranked removal. The slight improvement after removing the lowest-ranked source suggests that QSR assigns low scores to sources containing mostly redundant or distracting evidence.

Figure 6(b) shows a similar pattern at the bridge level. Removing the top 20% of bridges decreases MRR to 0.359, while random removal only reduces it to 0.379. In contrast, removing the bottom 20% slightly improves MRR to 0.389, indicating that low-contribution bridges may contain residual noise. At a 50% removal ratio, the MRR values under top-ranked, random, and bottom-ranked removal are 0.304, 0.360, and 0.381, respectively. These consistent counterfactual gaps demonstrate that the routing scores and reliability-weighted bridge contributions are closely associated with the model's actual predictions, supporting the faithfulness of UBER's evidence-selection and reasoning process.

F.4 Interpretability Analysis

To better understand how UBER performs multi-KG joint reasoning, we provide a qualitative analysis of its intermediate behaviors in Figure 7. Different from conventional multi-KG approaches that mainly focus on learning unified representations after graph integration, UBER explicitly models query-specific source selection, uncertainty-aware bridge estimation, and relation-level compatibility. Therefore, we visualize the corresponding routing scores, bridge reliability distributions, and relation prototype space to examine whether these components provide meaningful reasoning signals.

Query-aware source routing. Figure 7(a) presents the source routing distribution generated by QSR for different queries. We observe that the routing scores vary across queries rather than remaining uniform over all source KGs. For some queries, UBER assigns higher weights to a small subset of KGs that contain more relevant evidence, while other queries activate multiple source KGs to aggregate complementary information. This adaptive behavior demonstrates that QSR avoids unnecessary retrieval from irrelevant graphs and supports query-dependent evidence allocation instead of static multi-KG fusion.

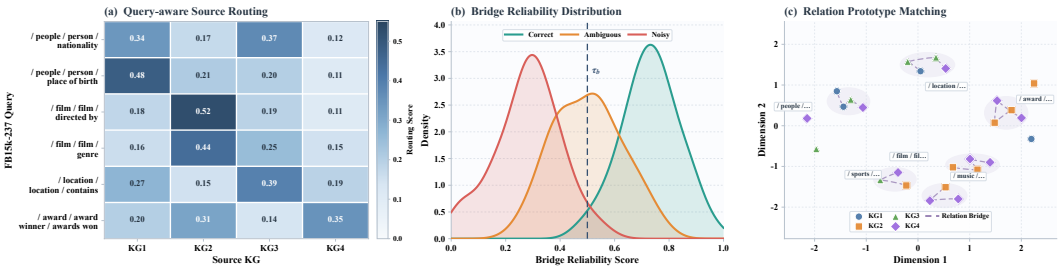

Figure 7: Interpretability analysis of UBER on FB15k-237-MKG. (a) Query-aware source-routing scores for representative queries, showing that different queries retrieve evidence from different combinations of source KGs. (b) Reliability-score distributions of correct, ambiguous, and noisy cross-KG bridges, where reliable correspondences receive consistently higher scores. (c) Learned relation prototypes and cross-KG relation bridges, illustrating that semantically compatible relations from different KGs are located in nearby regions.

Uncertainty-aware bridge estimation. Figure 7(b) illustrates the reliability score distributions of cross-KG bridges. Reliable bridges tend to concentrate in the high-score region, whereas noisy bridges receive lower confidence values. Ambiguous bridges exhibit intermediate distributions, reflecting the uncertainty of cross-KG correspondences. This separation indicates that UBC can provide discriminative reliability estimation for different bridge types, enabling UBER to selectively utilize trustworthy cross-KG connections while reducing the influence of unreliable links during evidence propagation.

Relation-level bridge matching. Figure 7(c) visualizes the learned relation prototype space across multiple KGs. Relations with similar semantic or structural roles are mapped into nearby regions, even when they originate from different source graphs. Meanwhile, the learned relation bridges connect compatible relation prototypes rather than relying solely on entity overlap. This observation verifies that RBM captures higher-level relation compatibility and provides complementary signals for cross-KG reasoning, especially in cases where heterogeneous KGs describe similar facts using different relation schemas.

Overall, these visualizations demonstrate that UBER achieves interpretability at multiple levels: it determines *where* to retrieve evidence through query-aware routing, estimates *which* cross-KG bridges are reliable through uncertainty modeling, and learns *how* heterogeneous relation semantics can be aligned through relation-level matching. These results further support that the performance improvements of UBER come from structured evidence composition rather than simple graph aggregation.

F.5 Scalability with More Source KGs

We evaluate the scalability of UBER as the number of source KGs increases from two to five. All experiments follow the same 20% entity-overlap setting and filtered ranking protocol. Tables 11 and 12 report the results on FB15k-237-MKG, WN18RR-MKG, and YAGO3-10-MKG, respectively.

As the number of source KGs increases, performance generally decreases because evidence becomes more fragmented and the number of potentially unreliable cross-KG connections grows. Nevertheless, UBER achieves the best MRR in all twelve dataset–KG configurations and the best Hits@10 in ten of them, demonstrating consistently strong top-ranking quality.

On FB15k-237-MKG, UBER maintains an MRR of 0.376 and a Hits@10 of 0.566 with five source KGs, outperforming the strongest competing results by 0.009 and 0.013, respectively. Moreover, its MRR decreases by only 0.016 from two to five KGs, compared with decreases of 0.018 for DMKGC and 0.021 for LSMGA. Similar trends are observed on WN18RR-MKG and YAGO3-10-MKG, where the corresponding MRR decreases of UBER are only 0.019 and 0.022.

UBER also achieves the best Hits@10 in all FB15k-237-MKG and WN18RR-MKG settings. On YAGO3-10-MKG, DMKGC obtains slightly higher Hits@10 under the 3KG and 4KG settings, sug-

Method	2KG		3KG		4KG		5KG	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Non-federated multi-KG and cross-KG methods								
KEnS	0.329	0.516	0.326	0.512	0.321	0.506	0.312	0.494
AlignKGC	0.345	0.535	0.340	0.530	0.335	0.524	0.326	0.512
SS-AGA	0.348	0.541	0.344	0.535	0.339	0.529	0.329	0.516
JMAC	0.354	0.545	0.351	0.541	0.346	0.536	0.337	0.524
LSMGA	0.367	0.573	0.362	0.567	0.356	0.561	0.346	0.548
GC-PLM	0.362	0.557	0.358	0.552	0.353	0.547	0.344	0.536
KL-GMoE	0.371	0.570	0.366	0.565	0.359	0.558	0.350	0.546
IMKGC	0.378	0.571	0.374	0.569	0.369	0.567	0.359	0.553
DMKGC	0.385	0.572	0.381	0.568	0.376	0.565	0.367	0.548
Federated KGC methods								
FedE	0.315	0.501	0.312	0.497	0.307	0.491	0.298	0.479
FedR	0.333	0.522	0.329	0.516	0.324	0.510	0.314	0.498
FedEC	0.326	0.512	0.322	0.507	0.317	0.501	0.308	0.489
FKGE	0.339	0.531	0.335	0.525	0.331	0.519	0.320	0.506
FedLU	0.352	0.543	0.348	0.538	0.343	0.531	0.334	0.520
UniFE	0.345	0.538	0.341	0.532	0.336	0.526	0.327	0.514
PFLU	0.374	0.565	0.368	0.559	0.361	0.552	0.352	0.541
UBER	**0.392**	**0.588**	**0.389**	**0.584**	**0.386**	**0.579**	**0.376**	**0.566**

Table 11: Effect of the number of source KGs on FB15k-237-MKG under the 20% overlap setting. Best results are shown in bold and second-best results are underlined.

Method	WN18RR-MKG								YAGO3-10-MKG							
	2KG		3KG		4KG		5KG		2KG		3KG		4KG		5KG	
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10
Non-federated multi-KG and cross-KG methods																
KEnS	0.464	0.554	0.458	0.548	0.452	0.541	0.442	0.529	0.514	0.684	0.508	0.678	0.501	0.671	0.491	0.659
AlignKGC	0.479	0.571	0.473	0.564	0.466	0.557	0.456	0.545	0.531	0.703	0.525	0.697	0.518	0.690	0.508	0.678
SS-AGA	0.481	0.574	0.475	0.567	0.468	0.560	0.458	0.548	0.534	0.705	0.528	0.699	0.521	0.692	0.511	0.680
JMAC	0.476	0.568	0.470	0.561	0.463	0.554	0.453	0.542	0.541	0.712	0.535	0.706	0.528	0.699	0.518	0.687
LSMGA	0.495	0.600	0.488	0.592	0.481	0.585	0.471	0.573	0.561	0.728	0.555	0.722	0.548	0.715	0.538	0.703
GC-PLM	0.502	0.593	0.495	0.586	0.488	0.579	0.478	0.567	0.567	0.727	0.561	0.721	0.554	0.714	0.544	0.702
KL-GMoE	0.506	0.599	0.499	0.594	0.492	0.582	0.482	0.570	0.565	0.733	0.558	0.726	0.552	0.718	0.542	0.710
IMKGC	0.517	0.598	0.511	0.591	0.504	0.583	0.494	0.571	0.572	0.732	0.566	0.725	0.559	0.719	0.549	0.707
DMKGC	0.523	0.597	0.517	0.590	0.510	0.584	0.500	0.575	0.578	0.735	0.572	**0.731**	0.565	**0.724**	0.555	0.709
Federated KGC methods																
FedE	0.450	0.540	0.444	0.534	0.438	0.528	0.428	0.516	0.501	0.669	0.495	0.663	0.489	0.656	0.479	0.644
FedR	0.473	0.563	0.467	0.557	0.461	0.551	0.451	0.539	0.528	0.695	0.522	0.689	0.516	0.682	0.506	0.670
FedEC	0.466	0.556	0.460	0.550	0.454	0.544	0.444	0.532	0.521	0.689	0.515	0.683	0.509	0.676	0.499	0.664
FKGE	0.458	0.548	0.452	0.542	0.446	0.536	0.436	0.524	0.513	0.682	0.507	0.676	0.501	0.669	0.491	0.657
FedLU	0.491	0.582	0.485	0.576	0.479	0.570	0.469	0.558	0.544	0.710	0.538	0.704	0.532	0.697	0.522	0.685
UniFE	0.484	0.575	0.478	0.569	0.472	0.563	0.462	0.551	0.553	0.717	0.547	0.711	0.541	0.704	0.531	0.692
PFLU	0.502	0.592	0.496	0.585	0.489	0.579	0.479	0.567	0.559	0.730	0.553	0.724	0.547	0.717	0.537	0.705
UBER	**0.526**	**0.618**	**0.522**	**0.613**	**0.517**	**0.607**	**0.507**	**0.596**	**0.584**	**0.737**	**0.578**	0.729	**0.571**	0.721	**0.562**	**0.712**

Table 12: Effect of the number of source KGs on WN18RR-MKG and YAGO3-10-MKG under the 20% overlap setting. Best results are shown in bold and second-best results are underlined.

gesting that diffusion-based transfer can occasionally improve broad candidate coverage. However, UBER consistently yields the best MRR and recovers the best Hits@10 under the more challenging 5KG setting. Overall, these results show that query-aware source routing limits retrieval from weakly relevant graphs, while uncertainty-aware bridge construction and reliability-gated fusion reduce noisy propagation as the number of sources increases.

F.6 Efficiency and Computational Complexity

We further evaluate the computational efficiency of UBER from both theoretical and empirical perspectives. The key advantage of UBER is that its online reasoning cost depends primarily on the query-specific evidence budget rather than the size of the complete multi-KG union.

B_{tot}	MRR	Latency	Queries/s	Memory
64	0.360	4.8	208	3.1
128	0.374	6.5	154	3.5
256	**0.386**	**9.6**	**104**	**4.4**
512	0.379	15.8	63	5.9
1024	0.367	28.7	35	8.7

Table 13: Efficiency–performance trade-off of UBER with different total evidence budgets on FB15k-237-MKG under the 4KG–20% setting.

Method	FB15k-237-MKG		WN18RR-MKG		YAGO3-10-MKG	
	Latency	Memory	Latency	Memory	Latency	Memory
SS-AGA	8.7	3.9	12.6	5.3	29.8	10.7
LSMGA	10.4	4.6	14.8	6.0	34.6	11.9
KL-GMoE	12.8	5.8	17.2	7.1	41.2	14.1
IMKGC	11.5	5.2	16.0	6.5	38.5	13.0
DMKGC	16.9	6.1	22.4	7.6	51.7	15.2
UBER	**9.6**	**4.4**	**13.7**	**5.7**	**31.9**	**11.8**

Table 14: Inference efficiency under the 4KG–20% entity-overlap setting. Latency is measured in milliseconds per query, and memory denotes peak GPU memory in GB. Lower values are better.

Computational complexity. Let

$$N_q = \sum_{m=1}^{M} |V_q^{(m)}| \le B_{\text{tot}} \tag{56}$$

be the number of retrieved entities, $F_q = |\mathcal{F}_q^{\text{fus}}|$ be the number of edges in the fused evidence graph, C_q be the number of candidate entity bridges, P_q be the number of candidate relation pairs, and $\bar{D}_q$ be the average degree examined during evidence retrieval.

With query projections cached, QSR and DBA require

$$O(Md^2 + B_{\text{tot}}\bar{D}_q d) \tag{57}$$

time for source scoring and query-aware evidence retrieval. UBC requires

$$O(C_q d + P_q d^2) \tag{58}$$

time for entity-reliability estimation and relation-bridge matching. The L-layer RFR has a complexity of

$$O(LF_q d^2). \tag{59}$$

The final all-entity scoring operation costs $O(|E^{\text{all}}|d)$, which is shared by the compared filtered ranking methods. The additional online memory cost of UBER is $O((N_q + F_q)d)$. Since N_q is bounded by B_{tot}, UBER avoids performing message passing over the complete union of all source KGs.

Evaluation protocol. All methods are evaluated on the 4KG–20% entity-overlap setting using the same hardware and software environment. Static entity features, relation features, and coarse bridge candidates are precomputed and cached for all methods when applicable. The reported online latency includes query routing, evidence retrieval, bridge construction, fused-graph reasoning, and candidate scoring, but excludes one-time data loading and index construction. We first execute 100 warm-up queries and then measure 1,000 randomly sampled test queries with a query batch size of one. CUDA synchronization is performed before and after each measurement. The reported latency and peak GPU memory are averaged over five runs.

Table 14 shows that UBER achieves a favorable efficiency–accuracy trade-off. Although its routing and reliability estimation introduce a modest overhead over the relatively simple SS-AGA, the additional latency is only 7.0–10.3% across the three datasets. Meanwhile, UBER improves MRR

by 0.047, 0.049, and 0.050 on FB15k-237-MKG, WN18RR-MKG, and YAGO3-10-MKG, respectively.

Compared with the stronger diffusion-based DMKGC, UBER reduces inference latency by 38.3–43.2% and peak GPU memory by 22.4–27.9%. This advantage becomes more evident on YAGO3-10-MKG, where globally transferring or propagating cross-source information is comparatively expensive. These results indicate that query-specific evidence composition provides stronger ranking performance without requiring indiscriminate reasoning over all available cross-KG evidence.

Efficiency under different evidence budgets. We next analyze how the total evidence budget affects online efficiency on FB15k-237-MKG. All settings use the same model parameters except for B_{tot}.

As shown in Table 13, inference latency and memory increase smoothly with the evidence budget, confirming that UBER provides explicit and predictable control over its online computational cost. A very small budget reduces latency but may omit useful complementary evidence. Increasing the budget from 64 to 256 improves MRR from 0.360 to 0.386, while maintaining a throughput of more than 100 queries per second. However, further increasing the budget introduces redundant or noisy evidence, resulting in both higher computational cost and lower ranking performance. Therefore, $B_{\mathrm{tot}} = 256$ provides the best efficiency–effectiveness balance on FB15k-237-MKG.

Overall, the empirical results are consistent with the complexity analysis: the cost of UBER grows with the compact query-specific evidence graph rather than the complete multi-KG union. The dynamic budget mechanism therefore serves not only as an evidence selection strategy but also as an explicit control over inference latency and memory consumption.