Agree to Disagree: Multimodal Autonomous Negotiation and Calibration for Entity Representation Learning

Chenyi Xiong[1], **Yan Zhang**[1,2], **Yueluan Huang**[1], **Lu Wang**[1],
Kui Xiao[1,2], **Wenxin Huang**[1,2], **Zhifei Li**[1,2,*]
[1]School of Computer Science, Hubei University
[2]Hubei Key Laboratory of Big Data Intelligent Analysis and Application, Hubei University
{xiongchenyi, yueluan, luwang}@stu.hubu.edu.cn
wenxinhuang_wh@163.com,{zhangyan,xiaokui,zhifei1993}@hubu.edu.cn

Abstract

Learning high-quality multimodal entity representations is essential for advancing reasoning tasks such as multimodal knowledge graph completion (MMKGC). However, rigid alignment in existing fusion strategies can bias representations toward a dominant anchor modality, weakening fine-grained complementary cues and propagating noisy signals into relational reasoning. To address these limitations, we propose **MANoR**, a **M**ultimodal **A**utonomous **N**ego**o**tiation framework for calibrated **R**epresentation learning. MANoR treats modalities as autonomous semantic agents that negotiate before fusion. Specifically, MANoR preserves modality-specific semantics through centroid-guided subspace contraction, which reduces unstable centroid-relative dispersion while retaining entity-level residual semantics. On this stabilized basis, negotiated attention enables selective cross-modal interaction without collapsing modality boundaries. As modalities can still differ in reliability after interaction, uncertainty calibrated fusion estimates dimension-level aleatoric uncertainty and modality-level reliability to compute relation-aware fusion weights. In this way, MANoR couples modality autonomy, selective interaction, and calibrated integration within a unified representation learning framework. Experiments on MKG-W, MKG-Y, TIVA, and KVC16K show that MANoR consistently outperforms strong MMKGC baselines, achieving a relative Hits@1 gain of up to 17.72% on KVC16K. Our code can be found at `https://github.com/HubuKG/MANoR`.

1 Introduction

Multimodal Knowledge Graphs (MMKGs) [19] extend conventional knowledge graphs by associating entities with heterogeneous modalities, including images, audio, videos, and textual descriptions. Such information is increasingly important for building AI systems that reason over entities not only as symbolic identifiers, but also as objects embedded in real-world multimodal contexts. By connecting structured relational semantics with unstructured perceptual signals, MMKGs provide a representation substrate for multimodal reasoning, recommendation, and knowledge-enhanced foundation models [7, 36, 11, 5].

Learning effective entity representations over MMKGs is a central problem for multimodal knowledge graph completion (MMKGC) [38], where the model must infer missing triples from both relational structure and multimodal content. Unlike unimodal KGC [1, 34], MMKGC requires the model to handle heterogeneous modalities whose semantic granularity, noise level, and relation-specific

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

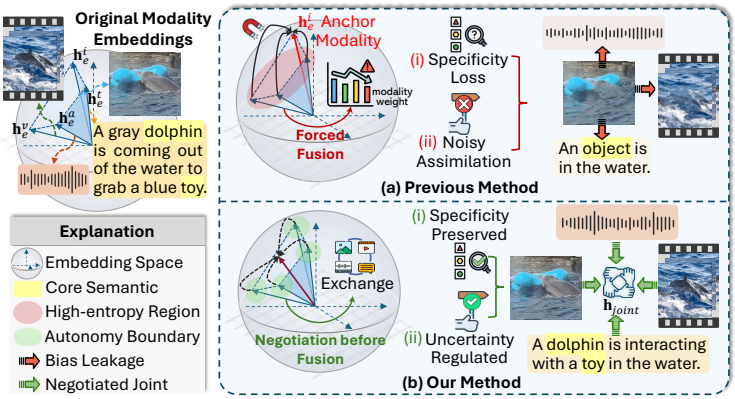

Figure 1: Comparison between (a) traditional fusion strategies may incur specificity loss and noisy assimilation from their rigid alignment, and (b) our proposed negotiation-based-fusion, which maintains a modality autonomy boundary to preserve inherent strengths and calibrate uncertainty.

usefulness can differ substantially. Therefore, an effective MMKGC model needs to preserve modality-specific semantics, exploit useful cross-modal complementarity, and prevent unreliable signals from dominating entity representations during relational reasoning.

Most existing MMKGC methods [16, 9] address these challenges through attention-based aggregation. While effective in many settings, these mechanisms often impose a rigid alignment among modalities. In practice, such alignment can bias the representation space toward a dominant anchor modality [12]. Once a modality becomes dominant, other modalities are encouraged to adapt to its geometry, even when they contain complementary or more fine-grained information. This tendency is related to the path-of-least-resistance behavior observed in multimodal learning [21, 18], where optimization favors signals that are easier to fit rather than those that are more semantically faithful.

This rigid alignment leads to two recurring failure modes, as illustrated in Figure 1(a). The first is **specificity loss**. When non-anchor modalities are forced to match the anchor geometry, their distinctive semantic cues may be diluted. For instance, a detailed textual description such as "*a gray dolphin grabbing a blue toy*" may be reduced to a generic concept such as "*an object*". The second is **noisy assimilation**. If the dominant modality is ambiguous or noisy, other modalities may absorb this uncertainty during fusion, producing representations that appear confident but are semantically unreliable. These problems become more severe when modalities differ significantly in noise level or semantic resolution.

To address these challenges, we propose a novel framework called MANoR, which rethinks fusion as a negotiation process rather than rigid geometric alignment, as shown in Figure 1(b). Our key insight is that distinct modalities should be treated as independent agents capable of "agreeing to disagree." Specifically, we introduce the Modality Autonomous Negotiation (MoA) module, which leverages subspace centroid projection to distill the semantic essence from noisy raw features, and then regulates cross-modal information exchange through attention bootstrapping. This design prevents semantic drift while preserving fine-grained core information. To further enhance robustness against low-quality or noisy data, we devise the Uncertainty Calibrated Fusion (UnCF) module. By modeling aleatoric uncertainty, UnCF acts as a probabilistic filter that dynamically penalizes high-entropy features and calibrates global modality biases. Finally, we integrate these negotiated and calibrated representations into a generative framework to jointly optimize the MMKGC task. Extensive experiments on multiple benchmarks demonstrate the effectiveness of MANoR.

Our contributions are summarized as follows:

- We propose MANoR, a novel framework that redefines multimodal fusion as an uncertainty-calibrated autonomous negotiation process. This framework enables modalities to preserve their intrinsic semantics while interacting selectively, thereby ensuring effective fusion.
- We develop two synergistic modules: Modality Autonomous Negotiation and Uncertainty Calibrated Fusion. They jointly combine centroid-guided subspace contraction with dynamic uncertainty calibration.

- We conduct extensive experiments on standard benchmarks, demonstrating that MANoR substantially outperforms strong baselines. Specifically, on the KVC16K dataset, it achieves gains of 17.41% in MRR, 17.72% in Hits@1, and 10.47% in Hits@10.

2 Related Work

Multimodal Knowledge Graph Completion. MMKGC aims to infer missing links by jointly exploiting relational structure and heterogeneous modality information [35]. Early translation-based methods, such as IKRL [31] and TransAE [29], projected visual and textual features into a unified vector space to constrain structural embeddings. Subsequent works explore more expressive fusion mechanisms to handle complex multimodal interactions, such as IMF [17] and AdaMF [38]. To address the varying importance of modalities across relational contexts, MoMoK [39] and MoCME [16] introduced mixture-of-experts architectures, allowing different experts to specialize in different relational contexts. Other works focus on more specific challenges in multimodal representation learning. MyGO [40] addresses granularity mismatch by modeling modality information across semantic levels, while APKGC [14] introduces attention penalties to reduce over-confident reliance on noisy modality signals. Recent approaches further explore contrastive learning and unified representation spaces to improve cross-modal alignment [20, 41].

Uncertainty Quantification. Uncertainty quantification (UQ) has been widely studied for improving the robustness of multimodal learning [3, 13, 4]. Classical Bayesian methods [8] and deep ensembles [15] provide principled estimates of predictive uncertainty, but their computational overhead can be substantial for large-scale multimodal knowledge graphs. As a result, recent work has increasingly focused on lightweight uncertainty estimation mechanisms that can guide feature selection, modality weighting, or dynamic fusion. Variational methods have been used to estimate modality confidence and adjust fusion weights [30, 23], while Gaussian uncertainty regressors have been introduced to calibrate latent distributions in multimodal fusion [25]. Temporal-invariant learning has also been explored to reduce redundant or unstable multimodal features [33].

Unlike methods that use UQ mainly as a regularizer or post-hoc confidence score, MANoR uses uncertainty to calibrate the fusion process itself, allowing unreliable feature dimensions and modalities to contribute less during relation-aware fusion.

3 Methodology

In this section, we propose MANoR, a multimodal representation learning framework illustrated in Figure 2. MANoR combines Modality Autonomous Negotiation for preserving specificity and Uncertainty Calibrated Fusion for mitigating noise propagation. Additionally, a generative adversarial scheme is employed to further strengthen the robustness of the learned representations.

3.1 Preliminaries

Problem Formulation. A KG is defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{T}$ denote sets of entities, relations, and triples (h, r, t), respectively. Furthermore, MMKGs have a modality set denoted as $\mathcal{M}$, encapsulating different modalities (image I, text T, audio A, video V) in the MMKGs. For each entity $e \in \mathcal{E}$, its raw data in modality $m \in \mathcal{M}$ is denoted as $Z_m(e)$. The graph structure itself constitutes an intrinsic modality, whose information is embodied in the triple set $\mathcal{T}$.

MMKGC aims to learn a scoring function $S(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$ to measure triple plausibility. Training employs negative sampling by corrupting entities in valid triples to form a negative set $\mathcal{T}'$. During inference, the model is evaluated via link prediction [1], ranking candidates for queries $(?, r, t)$ or $(h, r, ?)$ [38].

Pre-trained Multimodal Representation. Following [37], we employ pre-trained modality encoders to extract initial features for each modality. For entity e and modality m, the initial feature is computed as:

$$\mathbf{e}_m = \frac{1}{|Z_m(e)|} \sum_{x \in Z_m(e)} \Phi_m(x), \tag{1}$$

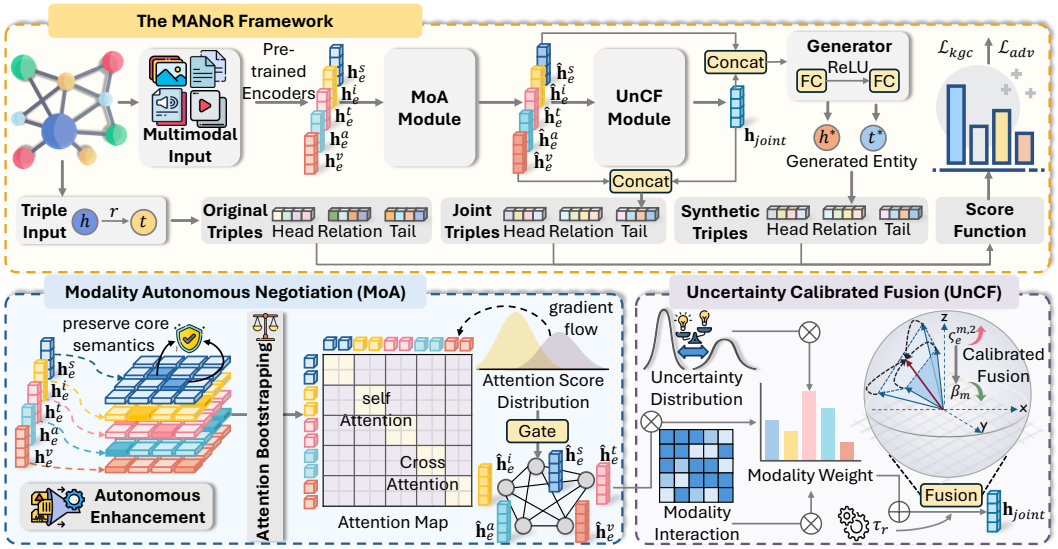

Figure 2: The overview of MANoR. The framework takes multimodal information and triples as inputs, integrates the Modality Autonomous Negotiation module to preserve modality-specific information and support cross-modal interaction, and the Uncertainty Calibrated Fusion module to down-weight noise via uncertainty estimation. Joint embeddings are finally optimized by a generative adversarial scheme for robust MMKGC.

where Φ_m is the corresponding pre-trained encoder, such as BERT [6] for text features. To unify representations across modalities, we project features into a shared semantic space via a lightweight adapter $\mathcal{P}_m$, implemented as a two-layer MLP with ReLU activation: $\mathbf{h}_e^m = \mathcal{P}_m(\mathbf{e}_m) \in \mathbb{R}^{d_e}$.

3.2 Modality Autonomous Negotiation

To handle modality heterogeneity without sacrificing semantic specificity, we propose Modality Autonomous Negotiation (MoA). MoA models modalities as independent agents that selectively leverage external context to enhance representations without distorting their internal feature manifolds.

Autonomous Enhancement Raw modality features contain both stable semantic factors and unstable variations. Directly mixing them across modalities may amplify semantic drift, especially when one modality is noisy or overly dominant [10]. Therefore, MoA stabilizes each modality within its own principal semantic subspace before cross-modal interaction. For modality m, we estimate a modality-specific basis U_m from the centered covariance of projected entity features. This basis retains the dominant covariance directions and defines the subspace where the main modality-specific variations are represented:

$$U_m = \arg\max_{U^\top U = I} \mathrm{Tr}(U^\top C_m U), \tag{2}$$

where C_m denotes the empirical covariance matrix. For simplicity, $\mathcal{E}$ denotes the training entities with available modality-m features in this subsection. The projected semantic factor and its modality centroid are:

$$\mathbf{s}_e^m = U_m^\top \mathbf{h}_e^m, \qquad \boldsymbol{\mu}^m = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbf{s}_e^m. \tag{3}$$

MoA then refines each entity by contracting its centroid-relative deviation in the principal subspace:

$$\tilde{\mathbf{h}}_e^m = \mathbf{h}_e^m - \lambda_m U_m (\mathbf{s}_e^m - \boldsymbol{\mu}^m), \tag{4}$$

where $\tilde{\mathbf{h}}_e^m$ is the enhanced representation, and $U_m(\mathbf{s}_e^m - \boldsymbol{\mu}^m)$ maps the centroid-relative deviation back to the original feature space. The coefficient $\lambda_m = \sigma(\theta_m) \in (0, 1)$ is a learnable modality-specific contraction coefficient, where $\theta_m \in \mathbb{R}$ is unconstrained and $\sigma(\cdot)$ is the sigmoid function. Larger λ_m yields stronger contraction toward the modality centroid.

Since $U_m^\top U_m = I$, the refined projected factor satisfies $\tilde{\mathbf{s}}_e^m - \boldsymbol{\mu}^m = (1-\lambda_m)(\mathbf{s}_e^m - \boldsymbol{\mu}^m)$, where $\tilde{\mathbf{s}}_e^m = U_m^\top \tilde{\mathbf{h}}_e^m$. Thus, the update reduces centroid-relative dispersion only within the selected subspace, while leaving the orthogonal residual $(I - U_m U_m^\top)\mathbf{h}_e^m$ unchanged. This controlled contraction stabilizes modality-specific representations without collapsing entity-level discriminative information.

Attention Bootstrapping After autonomous enhancement, each modality obtains a stabilized representation $\tilde{\mathbf{h}}_e^m$ but remains isolated from other modalities. The goal of attention bootstrapping is not immediate fusion, but to probe potential cross-modal relevance while maintaining modality autonomy. For entity e, we define modality-wise latent probes via linear transformations:

$$\mathbf{q}_e^m = \mathbf{W}_q^m \tilde{\mathbf{h}}_e^m, \quad \mathbf{k}_e^m = \mathbf{W}_k^m \tilde{\mathbf{h}}_e^m, \quad \mathbf{v}_e^m = \mathbf{W}_v^m \tilde{\mathbf{h}}_e^m, \tag{5}$$

where $\mathbf{W}_q^m, \mathbf{W}_k^m \in \mathbb{R}^{d_k \times d_e}$ are learnable projection matrices, while $\mathbf{W}_v^m \in \mathbb{R}^{d_e \times d_e}$ preserves the semantic dimension and d_k is the attention head dimension. We then define a soft relevance matrix capturing cross-modal affinity:

$$\alpha_e^{m \to n} = \frac{\exp\left(\langle \mathbf{q}_e^m, \mathbf{k}_e^n \rangle / \sqrt{d_k}\right)}{\sum_{n' \in \mathcal{M}} \exp\left(\langle \mathbf{q}_e^m, \mathbf{k}_e^{n'} \rangle / \sqrt{d_k}\right)}. \tag{6}$$

This mechanism allows each modality to identify complementary evidence from other modalities. Rather than directly overwriting modality features, we interpret this attention distribution as a bootstrap signal. The bootstrapped response is computed as:

$$\mathbf{b}_e^m = \sum_{n \in \mathcal{M}} \alpha_e^{m \to n} \mathbf{v}_e^n. \tag{7}$$

Negotiated Update Directly injecting the bootstrapped response may dilute modality-specific semantics. We therefore use a dimension-wise negotiation gate $\boldsymbol{\gamma}_e^m$ to control the incorporation of cross-modal information:

$$\boldsymbol{\gamma}_e^m = \sigma\Big(\mathbf{W}_g^m \left[\tilde{\mathbf{h}}_e^m \,\|\, \mathbf{b}_e^m\right] + \mathbf{b}_g^m\Big), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function, $[\cdot \,\|\, \cdot]$ denotes concatenation and $\mathbf{W}_g^m \in \mathbb{R}^{d_e \times 2d_e}$ is the gating projection. The updated representation is computed as:

$$\hat{\mathbf{h}}_e^m = \boldsymbol{\gamma}_e^m \odot \mathbf{b}_e^m + (1 - \boldsymbol{\gamma}_e^m) \odot \tilde{\mathbf{h}}_e^m, \tag{9}$$

where $\odot$ denotes the Hadamard product. Smaller values of $\boldsymbol{\gamma}_e^m$ preserve the autonomous representation, whereas larger values incorporate more cross-modal information.

3.3 Uncertainty Calibrated Fusion

Although negotiated representations incorporate complementary cross-modal evidence, their reliability is still uneven across entities, feature dimensions, and modalities. For example, visual signals can be informative for some relations but ambiguous for others. To reduce such noise propagation, we propose the Uncertainty Calibrated Fusion (UnCF) module, which estimates aleatoric uncertainty for each negotiated modality representation and uses it to calibrate relation-aware modality fusion.

Aleatoric Uncertainty Estimation After MoA, each modality obtains a negotiated representation $\hat{\mathbf{h}}_e^m$ that preserves modality-specific semantics while incorporating selective cross-modal evidence. We associate this representation with a diagonal Gaussian distribution:

$$\mathbf{z}_e^m \sim \mathcal{N}(\hat{\mathbf{h}}_e^m, \mathrm{diag}(\boldsymbol{\varsigma}_e^{m,2})), \tag{10}$$

where $\mathbf{z}_e^m$ denotes the latent probabilistic embedding. The diagonal variance is estimated as:

$$\boldsymbol{\varsigma}_e^{m,2} = \mathrm{Softplus}(\mathbf{W}_\varsigma^m \hat{\mathbf{h}}_e^m + \mathbf{b}_\varsigma^m) + \epsilon, \tag{11}$$

where $\boldsymbol{\varsigma}_e^{m,2}$ captures dimension-wise aleatoric uncertainty, $\mathbf{W}_\varsigma^m$ and $\mathbf{b}_\varsigma^m$ are learnable parameters, and ϵ is a small constant for numerical stability. A larger variance in a specific dimension corresponds to higher marginal uncertainty of that semantic dimension. Such dimensions are therefore assigned lower reliability in the subsequent fusion stage.

Calibrated Fusion To mitigate noise propagation, UnCF first constructs a reliability-aware scoring feature by attenuating uncertain dimensions:

$$\tilde{\mathbf{u}}_e^m = \hat{\mathbf{h}}_e^m \odot \frac{1}{1+\boldsymbol{\varsigma}_e^{m,2}}, \tag{12}$$

where $\tilde{\mathbf{u}}_e^m$ is the calibrated feature used for modality scoring. The attenuation is applied element-wise, so dimensions with larger predicted variance contribute less to the fusion weight. Besides dimension-level uncertainty, each modality may exhibit a persistent reliability bias during training. We summarize this bias by an exponential moving average:

$$y_e^m = \frac{1}{d_e}\sum_{j=1}^{d_e}\log\left(1+\boldsymbol{\varsigma}_{e,j}^{m,2}\right), \quad \bar{a}_m^{(t)} = \eta\bar{a}_m^{(t-1)} + (1-\eta)\frac{1}{|\mathcal{B}|}\sum_{e\in\mathcal{B}} y_e^m, \tag{13}$$

where y_e^m summarizes the average uncertainty, t is the training iteration, η is the momentum coefficient, and $\mathcal{B}$ is the current mini-batch. For relation-aware fusion, the calibrated modality weight is computed as:

$$\beta_m(e,r) = \frac{\exp\left(\frac{\alpha_r}{1+\bar{a}_m^{(t)}}\mathbf{V}_m^\top \tilde{\mathbf{u}}_e^m\right)}{\sum_{k\in\mathcal{M}}\exp\left(\frac{\alpha_r}{1+\bar{a}_k^{(t)}}\mathbf{V}_k^\top \tilde{\mathbf{u}}_e^k\right)}, \tag{14}$$

where $\mathbf{V}_m$ is a learnable modality-specific scoring vector, and $\alpha_r = \mathrm{Softplus}(\tau_r) + \epsilon$ is a relation-specific inverse temperature with learnable scalar $\tau_r \in \mathbb{R}$. The factor $1+\bar{a}_m^{(t)}$ down-weights globally unreliable modalities, while α_r controls the sharpness of modality selection for relation r. The final joint representation is obtained by aggregating the negotiated modality representations:

$$\mathbf{h}_{joint} = \sum_{m\in\mathcal{M}} \beta_m(e,r)\hat{\mathbf{h}}_e^m. \tag{15}$$

Thus, UnCF combines dimension-level uncertainty attenuation, modality-level reliability normalization, and relation-aware evidence selection in a unified fusion process.

3.4 Overall Training Objective

To jointly optimize structural reasoning and multimodal representation learning, we train MANoR with a link prediction objective and an adversarial generation objective. Given the joint entity embeddings $\mathbf{h}_{joint}$ and $\mathbf{t}_{joint}$, we use the RotatE scoring function [24] to measure the plausibility of a triple (h,r,t). The scoring function is defined as:

$$F(h,r,t) = -\|\mathbf{h}_{joint}\circ\mathbf{r} - \mathbf{t}_{joint}\|, \tag{16}$$

where $\circ$ is the rotation operator. The link prediction loss is optimized via self-adversarial negative sampling:

$$\mathcal{L}_{kgc} = -\log\sigma(\delta + F(h,r,t)) - \sum_{i=1}^{K} p(h_i',r,t_i')\log\sigma(-F(h_i',r,t_i')-\delta), \tag{17}$$

where δ is the margin and K is the number of negative samples. The negative set $\mathcal{T}' = \{(h_i',r,t_i')\}_{i=1}^K$ is generated by corrupting the head or tail entity of a positive triple (h,r,t), and the sampling weights $p(h_i',r,t_i')$ follow the self-adversarial strategy of RotatE [24].

Following NativE [37], we further introduce a two-layer MLP generator $\mathcal{G}$ to synthesize hard negative entities. The generator maps real entity embeddings together with random noise into the latent space. Given a triple (h,r,t), $\mathcal{G}$ generates synthetic entities h^* and t^* to construct the set of synthetic triples $\mathcal{S}(h,r,t) = \{(h^*,r,t),(h,r,t^*),(h^*,r,t^*)\}$. The adversarial loss encourages the discriminator to assign higher plausibility to real triples than to generated synthetic triples:

$$\mathcal{L}_{adv} = \sum_{(h,r,t)\in\mathcal{T}}\left(-F(h,r,t) + \frac{1}{|\mathcal{S}|}\sum_{(h^*,r,t^*)\in\mathcal{S}(h,r,t)} F(h^*,r,t^*)\right). \tag{18}$$

Method	MKG-W			MKG-Y			TIVA			KVC16K		
	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10
TransE [1]	29.19	21.06	44.23	30.73	23.45	43.37	83.85	83.20	84.15	8.54	0.64	23.42
DistMult [34]	20.99	15.93	30.86	25.04	19.33	35.95	82.27	81.15	84.22	6.37	3.03	12.61
ComplEx [26]	24.93	19.09	36.73	28.71	22.26	40.93	80.67	77.67	86.10	12.85	7.48	23.18
RotatE [24]	33.67	26.80	46.73	34.95	29.10	45.30	84.59	83.47	86.95	14.33	8.25	26.17
IKRL [31]	32.36	26.11	44.07	33.22	30.37	38.26	67.71	63.72	75.67	11.11	5.42	22.39
TransAE [29]	30.00	21.23	44.72	28.10	25.31	33.03	79.57	74.95	88.07	10.81	5.31	21.89
RSME [27]	29.23	23.36	40.43	34.44	31.48	39.09	40.01	30.55	51.35	12.31	7.14	22.05
OTKGE [2]	34.36	28.85	44.88	35.51	31.97	41.38	35.28	30.45	41.98	8.77	5.01	15.55
IMF [17]	34.50	28.77	45.44	35.79	32.95	40.63	55.46	41.87	77.57	12.01	7.42	21.01
QEB [28]	32.38	25.47	45.32	34.37	29.49	42.32	74.25	66.10	88.35	12.06	5.57	25.01
MMRNS [32]	35.03	28.59	47.47	35.93	30.53	45.47	83.12	83.05	83.25	13.31	7.51	24.68
AdaMF [38]	34.27	27.21	47.21	38.06	33.49	45.48	89.60	88.43	91.27	15.05	8.96	27.10
NativE [37]	36.58	29.56	49.39	39.04	34.79	46.25	92.27	91.40	92.85	15.76	9.23	28.55
MoMoK [39]	35.89	30.38	45.53	37.91	35.09	42.90	89.21	87.62	91.55	16.87	10.53	29.20
MYGO [40]	36.10	29.78	47.75	38.44	35.01	44.19	-	-	-	-	-	-
MCKGC [9]	36.88	31.32	47.43	38.92	35.49	45.21	-	-	-	-	-	-
DSoM [42]	40.90	33.56	55.28	40.02	35.67	47.71	-	-	-	-	-	-
MoCME [16]	37.79	30.81	-	40.37	36.21	-	94.58	93.32	95.61	18.32	11.23	30.76
MANoR (Ours)	**44.05**	**37.20**	**60.54**	**45.13**	**40.79**	**53.26**	**96.03**	**95.42**	**97.56**	**21.51**	**13.22**	**33.98**
Improvement ↑	**7.70%**	**10.85%**	**9.52%**	**11.79%**	**12.65%**	**11.63%**	**1.53%**	**2.25%**	**2.04%**	**17.41%**	**17.72%**	**10.47%**

Table 1: Link prediction results on four benchmark datasets. The best results are highlighted in **bold**, while the second-best results are underlined.

To stabilize adversarial training, we apply a gradient penalty on synthetic triples:

$$\mathcal{L}_{gp} = \sum_{(h^*,r,t^*)\in\mathcal{S}} \left(\|\nabla F(h^*, r, t^*)\|_2 - 1\right)^2. \tag{19}$$

The model is optimized via an alternating strategy, where the discriminator minimizes the KGC and adversarial losses, while the generator minimizes the adversarial loss regularized by the gradient penalty:

$$\begin{cases} \mathcal{L}_D = \mathcal{L}_{kgc} + \lambda_1 \mathcal{L}_{adv}, \\ \mathcal{L}_G = -\mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{gp}, \end{cases} \tag{20}$$

where λ_1 and λ_2 are hyperparameters balancing knowledge retention and adversarial generation.

4 Experiments

4.1 Experimental Setup

Datasets and Baselines. To evaluate the generalization capability and robustness of MANoR, we conduct experiments on four benchmark datasets: MKG-W [32], MKG-Y [32], TIVA [28], and KVC16K [22]. Detailed dataset statistics are provided in Appendix A.1, Table 3. We compare MANoR with strong baselines from two categories. (i) **Unimodal KGC Methods**: TransE [1], DistMult [34], ComplEx [26], and RotatE [24]. (ii) **Multimodal KGC Methods**: IKRL [31], TransAE [29], RSME [27], OTKGE [2], IMF [17], QEB [28], MMRNS [32], AdaMF [38], NativE [37], MoMoK [39], MYGO [40], MCKGC [9], DSoM [42] and MoCME [16].

Evaluation Metrics. Following standard protocols in knowledge graph completion [24], we evaluate the performance of our method using two widely adopted metrics: Mean Reciprocal Rank (MRR) and Hits@K (K=1, 10). To ensure fair evaluation, we adopt the filtered setting [1] by removing all valid triples from the candidate space before ranking.

4.2 Overall Performance

Table 1 reports the link prediction results on four benchmark datasets. MANoR achieves the best performance on all reported metrics. Compared with the strongest baseline on each dataset, it improves MRR by 7.70% on MKG-W, 11.79% on MKG-Y, 1.53% on TIVA, and 17.41% on KVC16K. The gains cover both two-modality and four-modality datasets, suggesting that MANoR generalizes across different modality configurations. On KVC16K, where image, text, audio, and

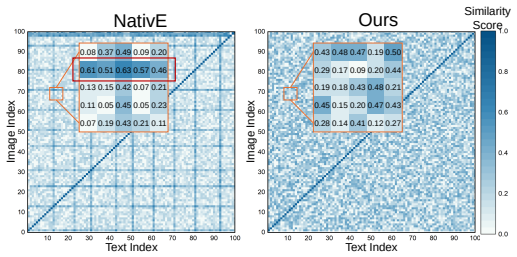

Figure 3: Cross-modal interaction matrices on MKG-Y under image-text matching.

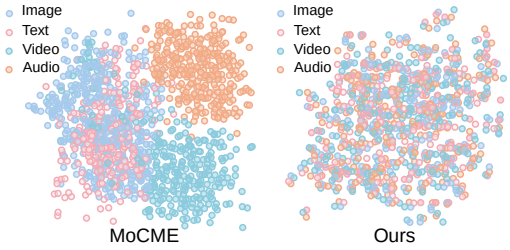

Figure 4: t-SNE visualization of learned modality distributions after fusion.

video features are all available, MANoR obtains the largest relative gain, improving MRR from 18.32 to 21.51 and Hits@1 from 11.23 to 13.22. This suggests that when multimodal information is richer but also more heterogeneous, directly aggregating modalities is less sufficient, and calibrated modality selection becomes more important. On TIVA, where the strongest baseline is already highly competitive, MANoR still improves all metrics, showing that the gain is not limited to low-performing settings. The consistent Hits@1 improvements further show that MANoR is effective in improving top-ranked predictions, which are especially important in practical prediction scenarios.

4.3 Modality Specificity Analysis

We visualize image-text interaction matrices on MKG-Y to examine whether fusion preserves modality-specific correspondence. This analysis directly targets the main failure mode discussed in Figure 1: if a fusion method collapses different modalities toward a generic anchor, then multiple non-matching samples may obtain high cross-modal similarity, producing horizontal or vertical striations in the interaction matrix. As shown in Figure 3, NativE exhibits such off-diagonal striations. This indicates that some embeddings are assigned high similarity to multiple non-matching samples, suggesting weaker pairwise discrimination between image and text features. In contrast, MANoR yields a more diagonally concentrated interaction matrix, where matched image-text pairs obtain higher similarity and non-matching pairs are less correlated. This pattern indicates that MANoR preserves more discriminative image-text correspondence, rather than simply increasing global cross-modal similarity. By mitigating early homogenization, MANoR maintains cross-modal discriminability and establishes a more reliable foundation for MMKGC.

4.4 Representation Visualization

To further examine how different methods organize multimodal representations, we visualize the learned feature distributions in Figure 4. The baseline MoCME exhibits a distinct modality gap. While visual and textual features partially align, audio and video features remain relatively isolated. This suggests that the fused space still retains strong modality-dependent separation, which may limit effective information transfer across modalities. Conversely, MANoR yields a more connected distribution where representations from all modalities are effectively interwoven. This demonstrates that our model successfully abstracts away from the heterogeneity of input signals. MANoR reduces unnecessary modality separation while ensuring that the learned representations are driven by intrinsic semantic content rather than modality-specific variations. Consequently, even highly disparate modalities are effectively integrated into the joint representation.

4.5 Ablation Study

Table 2 reports the ablation results on KVC16K. In G1, removing any single modality reduces performance, indicating that MANoR does not rely on a single dominant modality but benefits from complementary multimodal information. In G2, removing MoA causes the largest degradation, with MRR dropping from 21.51 to 18.97, showing that direct fusion without autonomous negotiation substantially weakens the learned representation. Removing $\boldsymbol{\mu}^m$ also decreases MRR to 19.78, indicating that the centroid-guided subspace operation is not redundant. Replacing the negotiation gate with a residual update further reduces performance. This comparison is important because both variants still allow cross-modal information flow, but the residual variant lacks selective control over

Table 2: The ablation study results on KVC16K. We divide experiments into four groups to isolate the impact of modalities (G1), MoA (G2), UnCF (G3), and the adversarial training strategy (G4). "Δ (%)" denotes the relative MRR drop compared with the full model.

Group	Setting	MRR	Δ (%)	Hits@1	Hits@10
G1	w/o image modality	19.56	↓ 9.07	10.51	30.34
	w/o text modality	19.12	↓ 11.11	11.37	31.96
	w/o audio modality	20.68	↓ 3.86	13.02	33.21
	w/o video modality	20.13	↓ 6.42	12.25	32.65
G2	w/o MoA	18.97	↓ 11.81	11.45	31.30
	w/o $\boldsymbol{\mu}^m$	19.78	↓ 8.04	11.98	33.25
	w/ Residual as $\boldsymbol{\gamma}_e^m$	20.35	↓ 5.39	12.57	32.91
G3	w/ Attn. as UnCF	19.92	↓ 7.39	11.56	33.83
	w/o $\bar{a}_m$	20.54	↓ 4.51	12.69	33.77
G4	w/o Adv. Training	21.17	↓ 1.58	13.10	33.90
Full MANoR Model		**21.51**		**13.22**	**33.98**

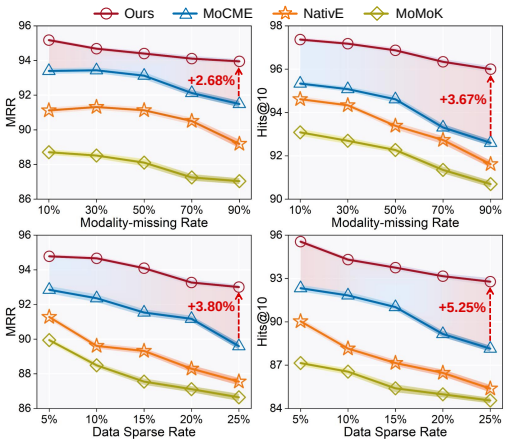

Figure 5: Performance comparison on TIVA under modality missing and data sparsity scenarios.

what is incorporated. The gap therefore points to the necessity of controlled, rather than unconditional, cross-modal updates. In G3, replacing UnCF with standard attention reduces MRR to 19.92, while removing $\bar{a}_m$ reduces MRR to 20.54. These results show that attention alone can model feature similarity, but it does not explicitly account for unreliable dimensions or persistently noisy modalities. G4 shows that removing adversarial training leads to only a moderate decline.

4.6 Complex Scenarios Analysis

To evaluate robustness under incomplete and limited multimodal information, we test MANoR under two challenging settings: modality missing and data sparsity. The results are summarized in Figure 5.

The top row of Figure 5 shows the results under different modality missing rates. As the missing rate escalates from 10% to 90%, baselines suffer catastrophic degradation, betraying their structural dependency on a dominant anchor. MANoR consistently degrades more slowly than the baselines and outperforms the strongest baseline by 2.68% in MRR at the extreme 90% setting. This suggests that MANoR reduces structural dependency on data completeness by fostering modality autonomy, enabling the reconstruction of robust semantics exclusively from the available evidence.

The bottom row of Figure 5 highlights the model's superiority in low-resource settings. As sparsity levels increase, the performance gap widens, with MANoR achieving a 5.25% gain in Hits@10 at 25% sparsity. Unlike baselines that overfit to sparse noise, MANoR dynamically down-weights uncertain signals. This prevents the model from memorizing spurious patterns, maintaining robust reasoning capabilities even in data-scarce environments.

4.7 Case Study

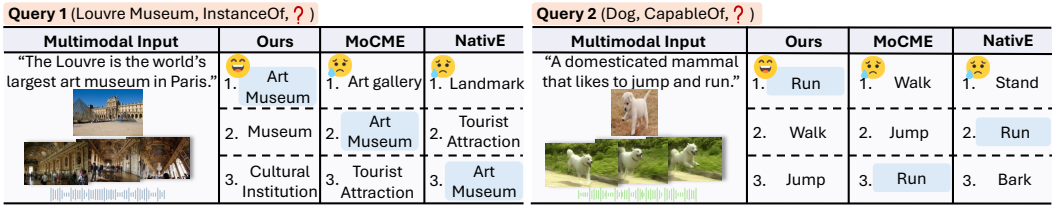

Figure 6: Qualitative comparison of Top-3 predictions under multimodal inputs.

Figure 6 presents a qualitative analysis of real cases from the TIVA dataset. In the first case, NativE biases towards visually salient but semantically generic concepts (e.g., *Landmark* or *Tourist Attraction*), while MoCME struggles to discriminate *Art Museum* from the closely related *Art Gallery*, reflecting semantic over-smoothing caused by forced fusion strategies. However, our method accurately identifies *Art Museum* as the top prediction, followed by semantically coherent candidates,

demonstrating its capacity to preserve fine-grained semantic distinctions. Similarly, for the query *(Dog, CapableOf, ?)*, our method prioritizes the specific action *Run*, whereas baselines output generic or static concepts like *Walk* or *Stand*. These cases suggest that explicitly modeling modality autonomy and uncertainty is essential for achieving precise and semantically faithful multimodal learning.

5 Conclusion

In this paper, we present MANoR, a multimodal representation learning framework that formulates MMKGC fusion as uncertainty-calibrated autonomous negotiation. MANoR stabilizes modality-specific representations before cross-modal interaction and uses estimated reliability to calibrate the subsequent integration process. Experiments on four MMKGC benchmarks show consistent improvements over strong baselines, including settings with missing modalities and sparse training data. A limitation is that MANoR is mainly evaluated on MMKGC benchmarks. Future work will investigate the extension of this approach to broader multimodal reasoning and foundation model settings, preserving intrinsic modal semantics against text-centric homogenization to ground linguistic reasoning.

Acknowledgments and Disclosure of Funding

Funding. This work was supported in part by the National Natural Science Foundation of China (No. 62677024, 62677025, 62377009), the Hubei Provincial Key Research and Development Program (No. 2026BCB079), and the Open Fund of Hubei Key Laboratory of Big Data Intelligent Analysis and Application, Hubei University (No. 2025BDIAA01).

Competing Interests. The authors declare no competing interests.

References

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2013.

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

[3] Lei Chen, Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Uncertainty-aware representation learning for action segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 820–826, 2022.

[4] Luke Chen, Junyao Wang, Trier Mortlock, Pramod Khargonekar, and Mohammad Abdullah Al Faruque. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22306–22316, 2025.

[5] Zhuo Chen, Wen Zhang, Yufeng Huang, Mingyang Chen, Yuxia Geng, Hongtao Yu, Zhen Bi, Yichi Zhang, Zhen Yao, Wenting Song, Xinliang Wu, Yi Yang, Mingyi Chen, Zhaoyang Lian, Yingying Li, Lei Cheng, and Huajun Chen. Tele-knowledge pre-training for fault analysis. In *39th IEEE International Conference on Data Engineering*, pages 3453–3466, 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[7] Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 2417–2429, 2024.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

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

[10] Eleonora Grassucci, Giordano Cicchetti, Aurelio Uncini, and Danilo Comminiello. Semantic compression via multimodal representation learning. *CoRR*, abs/2509.24431, 2025.

[11] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. Improving sequential recommendation with knowledge-enhanced memory networks. In *Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 505–514, 2018.

[12] Minoh Jeong, Min Namgung, Zae Myung Kim, Dongyeop Kang, Yao-Yi Chiang, and Alfred Hero. Anchors aweigh! sail for optimal unified multi-modal representations. *CoRR*, abs/2410.02086, 2024.

[13] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 23262–23271, 2023.

[14] Yue Jian, Xiangyu Luo, Zhifei Li, Miao Zhang, Yan Zhang, Kui Xiao, and Xiaoju Hou. APKGC: noise-enhanced multi-modal knowledge graph completion with attention penalty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15005–15013, 2025.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Lijian Li. Complementarity-driven representation learning for multi-modal knowledge graph completion. *CoRR*, abs/2507.20620, 2025.

[17] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference*, pages 2572–2580, 2023.

[18] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[19] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. MMKG: multi-modal knowledge graphs. In *Proceedings of the 16th International Conference on The Semantic Web*, pages 459–474, 2019.

[20] Yuxing Lu, Weichen Zhao, Nan Sun, and Jinzhuo Wang. Enhancing multimodal knowledge graph representation learning through triple contrastive learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5963–5971, 2024.

[21] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 26752–26762, 2024.

[22] Haojie Pan, Yuzhou Zhang, Zepeng Zhai, Ruiji Fu, Ming Liu, Yangqiu Song, Zhongyuan Wang, and Bing Qin. Kuaipedia: a large-scale multi-modal short-video encyclopedia. *CoRR*, abs/2211.00732, 2022.

[23] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the International Conference on Computer Vision*, pages 6301–6310, 2019.

[24] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations*, 2019.

[25] Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 805–822, 2023.

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

[27] Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. Is visual context really helpful for knowledge graph? A representation learning perspective. In *Proceedings of the ACM International Conference on Multimedia*, pages 2735–2743, 2021.

[28] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the ACM International Conference on Multimedia*, pages 2391–2399, 2023.

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

[30] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[32] Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the ACM International Conference on Multimedia*, pages 3857–3866, 2022.

[33] Guoyang Xu, Junqi Xue, Yuxin Liu, Zirui Wang, Min Zhang, Zhenxi Song, and Zhiguo Zhang. Semantic-guided multimodal sentiment decoding with adversarial temporal-invariant learning. *CoRR*, abs/2409.00143, 2024.

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

[35] Yichi Zhang and Wen Zhang. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. *CoRR*, abs/2209.07084, 2022.

[36] Yichi Zhang, Zhuo Chen, Yin Fang, Yanxi Lu, Li Fangming, Wen Zhang, and Huajun Chen. Knowledgeable preference alignment for llms in domain-specific question answering. In *Findings of the Association for Computational Linguistics*, pages 891–904, 2024.

[37] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 91–101, 2024.

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

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

[40] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 13322–13330, 2025.

[41] Lan Zhao, Boyue Wang, Junbin Gao, Xiaoyan Li, Yongli Hu, and Baocai Yin. Multi-modal entity in one word: Aligning multi-level semantics for multi-modal knowledge graph completion. *IEEE Transactions on Big Data*, 2025.

[42] Yu Zhao, Ying Zhang, Xuhui Sui, Baohang Zhou, Haoze Zhu, Jeff Z Pan, and Xiaojie Yuan. Dark side of modalities: Reinforced multimodal distillation for multimodal knowledge graph reasoning. In *Proceedings of the ACM International Conference on Multimedia*, pages 2506–2515, 2025.

A Additional Experimental Details

A.1 Dataset Statistics

We provide the detailed statistics of the datasets used in our experiments in Table 3. MKG-W and MKG-Y contain image and text features, while TIVA and KVC16K provide richer multimodal information including image, text, audio, and video features. These datasets cover different relation scales, modality combinations, and feature dimensions, allowing us to evaluate the generalization ability of MANoR under both two-modality and four-modality MMKG settings. In the table, Num. denotes the number of entities with available features for a given modality, and Dim. denotes the corresponding feature dimension. A dash indicates that the modality is not available in the dataset.

Table 3: Statistics of the experimental datasets. Num. and Dim. denote the number of entities with available modality features and the feature dimension, respectively.

Dataset	#Entity	#Relation	#Train	#Valid	#Test	Image		Text		Audio		Video	
						Num.	Dim.	Num.	Dim.	Num.	Dim.	Num.	Dim.
MKG-W [32]	15000	169	34196	4276	4274	14463	383	14123	384	-	-	-	-
MKG-Y [32]	15000	28	21310	2665	2663	14244	383	12305	384	-	-	-	-
TIVA [28]	11858	16	20071	2000	2000	11636	2048	11858	300	2441	128	10269	2048
KVC16K [22]	16015	4	180190	22523	22525	14822	768	14822	768	14822	768	14822	768

A.2 Implementation Details

We implement our method in PyTorch and operate it on an NVIDIA A100 GPU. For all datasets, we set the batch size to 1024, the number of attention heads to 4, and the momentum coefficient η to 0.9. We initialize each learnable relation-specific parameter τ_r to 0.01. We tune key hyperparameters over: learning rate $\{1e^{-4}, 1e^{-5}\}$, embedding dimension d_e {150, 250}, margin δ {3, 4, 9}, and λ_1, λ_2 $\{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$. We fix the optimizer as Adam.

A.3 Efficiency Analysis

Despite introducing the MoA and UnCF modules, MANoR maintains high efficiency with minimal computational overhead. Table 4 shows that incorporating these components leads to only a marginal increase in training time compared to the lightweight baseline NativE. This slight overhead is an expected consequence of the adversarial training mechanism employed to bolster robustness against modal incompleteness. Crucially, MANoR avoids the heavy computational burden observed in MoCME, training significantly faster while achieving the lowest memory consumption on large-scale datasets. Overall, our framework strikes an optimal balance, delivering state-of-the-art performance with competitive time and space efficiency.

Table 4: Efficiency comparison on different datasets. We report the training time per epoch (minutes) and GPU memory usage (MB).

Method	TIVA		KVC16K	
	Time	Memory	Time	Memory
MMRNS	8.51	24582	15.52	25600
NativE	1.98	9723	4.10	9983
MoMoK	3.50	11264	8.36	12288
MoCME	6.55	17408	12.59	21491
Ours	**2.17**	**8920**	**4.68**	**9831**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions of MANoR. These claims are supported by the methodology in Section 3 and the experimental results in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We acknowledge the current scope of MANoR in the conclusion. The experiments focus on MMKGC benchmarks, and future work will explore whether the proposed modality autonomy and uncertainty calibration principles generalize to broader multimodal reasoning and foundation model settings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present theoretical results, theorems, or formal proofs. The equations in Section 3 define the MANoR architecture, fusion process, and training objective.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model architecture, datasets, baselines, evaluation metrics, and main experimental protocol. Additional dataset statistics, implementation details, and efficiency results are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide publicly released code and running instructions in the supplementary material to reproduce the main experiments. The datasets are publicly available MMKGC benchmarks, and their original sources are cited in Section 4.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes the datasets, baselines, and evaluation metrics, including the filtered link prediction protocol. Appendix A provides dataset statistics, implementation details, optimizer choice, and hyperparameter search ranges.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

Justification: We do not report error bars or statistical significance tests due to the computational cost of running all MMKGC baselines multiple times. To improve reliability, we evaluate MANoR on four benchmark datasets, compare against a broad set of strong baselines, and include ablation, robustness, efficiency, and qualitative analyses.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the compute environment in the appendix A, including the NVIDIA A100 GPU used for experiments. We also provide an efficiency comparison with training time per epoch and GPU memory usage on TIVA and KVC16K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that this work conforms to it. The experiments use publicly available MMKGC benchmark datasets and do not involve collecting private data, human-subject experiments, or personally identifiable information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the scope and limitation of MANoR in Section 5. The method may benefit multimodal knowledge graph completion and downstream knowledge-enhanced systems, while its predictions can still be affected by noisy or incomplete modality features. Therefore, MMKGC outputs should be treated as decision-support signals rather than definitive judgments in sensitive applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release a high-risk generative model, scraped dataset, or dual-use asset. The released materials consist of publicly released code for reproducing MANoR on publicly available MMKGC benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources of the MMKGC benchmark datasets, pretrained encoders, and baseline methods used in this paper. The experiments rely on publicly released assets, and we use them for academic evaluation following their original release terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide publicly released code for MANoR with documentation, training scripts, and instructions for reproducing the main experimental results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: This paper does not involve crowdsourcing or newly conducted human-subject experiments. All experiments are conducted on existing publicly available MMKGC benchmark datasets.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: This work does not involve new data collection from human participants, crowdsourcing, user studies, or human-subject experiments. All experiments are conducted on existing publicly available MMKGC benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not used as an important, original, or non-standard component of MANoR. The proposed method does not rely on LLM-based prediction, calibration, or generation modules.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.