NativE: Multi-modal Knowledge Graph Completion in the Wild

Yichi Zhang
Zhejiang University
HangZhou, Zhejiang, China
zhangyichi2022@zju.edu.cn

Zhuo Chen
Zhejiang University
HangZhou, Zhejiang, China
zhuo.chen@zju.edu.cn

Lingbing Guo
Zhejiang University
HangZhou, Zhejiang, China
lbguo@zju.edu.cn

Yajing Xu
Zhejiang University
HangZhou, Zhejiang, China
yajingxu@zju.edu.cn

Binbin Hu
Ant Group
HangZhou, Zhejiang, China
bin.hbb@antfin.com

Ziqi Liu
Ant Group
HangZhou, Zhejiang, China
ziqiliu@antfin.com

Wen Zhang*
[1]Zhejiang University
[2]Zhejiang University-Ant Group Joint Laboratory of Knowledge Graph
[3]Alibaba-Zhejiang University Joint Institute of Frontier Technology
HangZhou, Zhejiang, China
zhang.wen@zju.edu.cn

Huajun Chen*
[1]Zhejiang University
[2]Zhejiang University-Ant Group Joint Laboratory of Knowledge Graph
[3]Alibaba-Zhejiang University Joint Institute of Frontier Technology
HangZhou, Zhejiang, China
huajunsir@zju.edu.cn

ABSTRACT

Multi-modal knowledge graph completion (MMKGC) aims to automatically discover the unobserved factual knowledge from a given multi-modal knowledge graph by collaboratively modeling the triple structure and multi-modal information from entities. However, real-world MMKGs present challenges due to their diverse and imbalanced nature, which means that the modality information can span various types (e.g., image, text, numeric, audio, video) but its distribution among entities is uneven, leading to missing modalities for certain entities. Existing works usually focus on common modalities like image and text while neglecting the imbalanced distribution phenomenon of modal information. To address these issues, we propose a comprehensive framework NativE to achieve MMKGC in the wild. NativE proposes a relation-guided dual adaptive fusion module that enables adaptive fusion for any modalities and employs a collaborative modality adversarial training framework to augment the imbalanced modality information. We construct a new benchmark called WildKGC with five datasets to evaluate our method. The empirical results compared with 21 recent baselines confirm the superiority of our method, consistently achieving state-of-the-art performance across different datasets and various scenarios while keeping efficient and generalizable. Our code and data are released at https://github.com/zjukg/NATIVE.

∗ Corresponding authors.

SIGIR 2024, July 14-18, 2024, Washington D.C., USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; *Semantic networks.*

KEYWORDS

Multi-modal Knowledge Graphs, Knowledge Graph Completion, Multi-modal Fusion, Adversarial Learning

ACM Reference Format:
Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang*, and Huajun Chen*. 2024. NativE: Multi-modal Knowledge Graph Completion in the Wild. In *Proceedings of The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

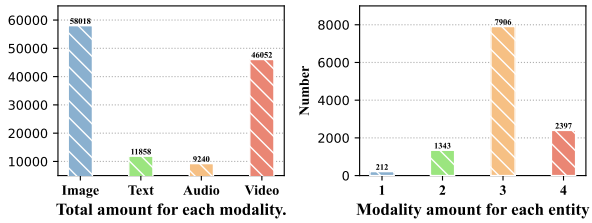
1 INTRODUCTION

Multi-modal knowledge graphs (MMKGs) [7, 8, 29] represent a rich extension of traditional knowledge graphs (KGs) [27, 47], enriching the structure triples in the form of *(head entity, relation, tail entity)* with comprehensive multi-modal entity attributes. MMKGs are representative and have become the significant infrastructure in many AI fields such as recommender systems [48], computer vision [15], and nature language processing [57, 61].

Investigating how to leverage structural knowledge in KGs and MMKGs is beneficial for many areas of AI research. However, KGs and MMKGs usually face critical **incompleteness issues** as they entail many unobserved factual knowledge. This phenomenon makes knowledge graph completion (KGC) [27] a significant task to automatically discover new knowledge in the given KGs. Conventional KGC methods [3, 37, 40, 55] generally emphasize learning structural embeddings to model the triple structure and measure the triple plausibility. Additionally, MMKGs accomplish the task more holistically by incorporating information from different modalities such as images and text into the KGC model, extending the task to multi-modal KGC (MMKGC) [5, 24, 54].

Dataset	Image	Text	Numeric	Audio	Video
WN9-IMG [51]	✓	✓	✗	✗	✗
MMKG [25]	✓	✓	✓	✗	✗
Richpedia [41]	✓	✓	✗	✗	✗
MKG-W/Y [52]	✓	✓	✗	✗	✗
Kuaipedia [27]	✓	✓	✗	✓	✓
TIVA [46]	✓	✓	✗	✓	✓
MMpedia [50]	✓	✓	✗	✗	✗

(a) Statistics of modalities in different MMKGs.



(b) Modality information distribution overall TIVA.

Figure 1: The diversity and imbalance nature in MMKGs. We report the modalities included in each MMKG in (a) and the statistical information about the modality information distribution across dataset/entity in TIVA in (b).

Existing MMKGC methods [24, 34, 45, 53] typically treat the multi-modal information of entities as auxiliary multi-modal embeddings and incorporate them to enhance the entity representations. However, these methods neglect two vital problems for MMKGC in real-world scenarios: the **diversity problem** and the **imbalance problem**. The diversity problem arises from the diverse modalities in contemporary information media. As the use of varied modalities such as numeric [43], audio [49], and video [31] data increases in MMKG construction, a summary in Figure 1(a) highlights the range of modalities included in the different MMKGs. However, the existing MMKGC methods are primarily designed for the prevalent image and text modalities, limiting their generalization to diverse modalities. The imbalance problem arises from the imbalanced distribution of modality information, implying that some important modality information would be missing in the real-world KGs. Figure 1(b) shows the total amount and entity-wise distribution of different modalities in the TIVA [49], confirming the uneven distribution of modalities across the datasets. Existing works do not focus on this problem or solve it by naive initialization [53], leading to inadequate utilization of modality information.

Aiming to solve these key issues of MMKGC in the wild, we propose a novel framework NativE, which can process and fuse **N**umeric, **A**udio, **T**ext, **I**mage, **V**ideo, and any other modalities into the **E**mbedding space with adaptive fusion and adversarial augmentation. NativE comprises two key modules called **Relation-guided Dual Adaptive Fusion** (ReDAF) module and **Collaborative Modality Adversarial Training** (CoMAT) module respectively. The ReDAF facilitates diverse multi-modal fusion with any input modalities with relation guidance to moderate the weight of each modality. The CoMAT designs an adversarial training strategy to augment the imbalanced modality information with Wasserstein distance [1] based objective. In addition, we undertake a theoretical analysis to prove the rationale of our designs. We construct a new benchmark called **WildKGC** with five MMKG datasets to evaluate our method against 21 recent baselines with further exploration. Our contributions can be summarized as three-fold:

- **Innovative framework**. We propose a new framework called NativE to address the diversity and imbalance problems of MMKG in the wild. NativE can achieve adaptive fusion with any modality with relation guidance to address the diversity problem of MMKG while augmenting the imbalanced modality information by collaborative modality adversarial training.
- **Theoretical analysis**. We performed a theoretical analysis to prove the legitimacy and soundness of our design.
- **Comprihensive experiments**. We construct a new benchmark to evaluate the MMKGC task in the wild and conduct comprehensive experiments to demonstrate the effectiveness, efficiency, and generalization of our NativE framework.

2 RELATED WORKS

2.1 Knowledge Graph Completion

Knowledge graph completion (KGC) [27] is an essential task in the community, aiming to discover the unobserved triples in the given KG. Conventional KGC methods are usually embedding-based, which embed the entities and relations of KGs into the continuous vector space and learn the embeddings based on the existing triple structure. This technique is also called **knowledge graph embedding (KGE)** [47]. Typically, conventional KGE models are designed with different score functions to measure the plausibility of triples with a general target to assign higher scores for positive triples and lower scores for negative triples.

The existing KGE models can be divided into two main categories: translation-based methods and semantic matching methods. Translation-based methods such as TransE [3], TransD [18], RotatE [37], OTE [38], and PairRE [6] modeling the triple structure as relational translation from the head entity to the tail entity, which design distance-based score functions as the plausibility measurement. Semantic matching methods such as DistMult [55], ComplEx [40], TuckER [2] exploit similarity-based scoring functions based on tensor decomposition. Some methods [10, 25, 26, 56, 63, 66] also attempt to extract structural semantics with deep neural networks.

2.2 Multi-modal Knowledge Graph Completion

Multi-modal knowledge graph completion (MMKGC) further considers utilizing the complex multi-modal information in the MMKGs to benefit the KGC. Current mainstream methods usually extend the conventional KGE models with more flexible multi-modal embeddings of entities such as visual embeddings and textual embeddings. These embeddings are extracted with pre-trained models and represent the entity feature from multi-views. In our taxonomy, the MMKGC methods can be further divided into three categories. The first category is the modal fusion methods. These methods [5, 22, 32, 34, 45, 51, 53, 62] design elegant approaches to achieve multi-modal fusion in the same representation space. The second category is modal ensemble methods. These methods [24, 65] learn the respective models for each modality and make joint predictions with ensemble learning. The third category is the negative sampling (NS) enhanced methods. As mentioned before,

NS is an important technology for KGE model training. Therefore, some methods [54, 60, 64] attempt to enhance the NS process and generate high-quality negative samples by utilizing the multi-modal information. The problem scenarios of existing MMKGC methods are relatively simple and usually consider the text and image modalities, while some of the work just considers the numerical modality. Also, they do not consider the data imbalance problem in MMKGs. In our work, we plan to accomplish the MMKGC task with a more unified perspective for more modalities and more complex multi-modal data distribution in the wild.

2.3 Generative Adversarial Networks

Generative adversarial networks (GAN) [14] is a milestone progress in the field of deep learning. GAN proposes to train a pair of discriminator and generator by playing a min-max game between them and achieve better performance, which has been widely used in various fields such as computer vision [1, 19], natural language processing [9, 58], information retrieval [44, 59], and recommender systems [50, 52, 58]. In the KGC field, there are some methods [4, 30, 39, 46] employs a GAN-based framework to enhance the negative sampling [3] process. For example, KBGAN [4] utilizes reinforcement learning (RL) with GAN to learn a better sampling policy. MMKRL [30] designs an adversarial training strategy for MMKGC. However, these methods tend to be oriented towards conventional KGC and often require RL, which leads to a more limited effect. Our work is the first to involve multi-modal entity information in MMKGCs and propose a unified framework to enhance MMKGC models.

3 TASK DEFINITION

A KG can be typically represented as $\mathcal{K} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ where $\mathcal{E}, \mathcal{R}$ are the entity set, the relation set respectively. $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the triple set. Furthermore, MMKGs have a modality set denoted as $\mathcal{M}$, encapsulating different modalities (image I, text T, numeric N, audio A, video V) in the MMKGs. For a given modality $m \in \mathcal{M}$, the set of modal information is denoted as $\mathcal{X}_m$. For an entity $e \in \mathcal{E}$, its modality information m is denoted as $\mathcal{X}_m(e)$, which is an empty set if the corresponding modal information is missing. For different modalities, the elements in it have different forms. For instance, $\mathcal{X}_m(e)$ can be a set of images when $m = I$ and some video clips when $m = V$. Note that the graph structure (S) is also an intrinsic modality for each entity and the structural information is already embodied in the triple set $\mathcal{T}$.

The general purpose of the MMKGC task is to learn a score function $\mathcal{F} : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$ to discriminate the plausibility of a given triple (h, r, t). In this context, a higher score implies a more plausible triple. To enable differentiable computations with gradient-based optimization, entities and relations are embedded into continuous vector spaces, which is called knowledge graph embedding (KGE). The embeddings of entity e and relation r are denoted as $\boldsymbol{e} \in \mathbb{R}^{d_e}$ and $\boldsymbol{r} \in \mathbb{R}^{d_r}$, where d_e, d_r are the embedding dimensions. Convectional KGE models focus solely on the structural information of triple, rendering these embeddings as **structural embeddings**. The score function $\mathcal{F}(h, r, t)$ leverages the structural embeddings to calculate the triple scores. For MMKGs, the multi-modal entity information should also be considered in the score function, implying additional embeddings $\boldsymbol{e}_m$ for varying modalities $m \in \mathcal{M}$ to represent the multi-modal feature of an entity e. This expansion consequently complicates the score function by considering multi-modal embedding integration. During the training stage, negative sampling (NS) [3] is widely used to construct manual negative triples for contrastive learning as KGs usually consist of only observed positive triples. Given a positive triple (h, r, t), the head h or tail t is randomly replaced by another entity $e \in \mathcal{E}$ in the NS process. The negative triple set can be denoted as $\mathcal{T}' = \{(h', r, t) \mid (h, r, t) \in \mathcal{T} \cap h' \in \mathcal{E} \setminus \{h\}\} \cup \{(h, r, t') \mid (h, r, t) \in \mathcal{T} \cap t' \in \mathcal{E} \setminus \{t\}\}$. During the inference stage, the KGC model is usually evaluated with the link prediction task [3]. The target of link prediction is to predict the missing head or tail entity in the given query $(?, r, t)$ or $(h, r, ?)$. For instance, in tail prediction, the entire set of entities $\mathcal{E}$ will be the candidate set during evaluation. For each $e \in \mathcal{E}$, the plausibility score of the triple (h, r, e) is calculated and then ranked across the entire candidate set. A higher rank of the ground truth (h, r, t) represents better model performance.

4 METHODOLOGY

In this section, we will introduce the proposed MMKGC framework in detail. We refer to our model as NativE, designed to represent and combine multiple data modalities (**N**umeric, **A**udio, **T**ext, **I**mage, **V**ideo, and more) from MMKGs into multi-modal **E**mbeddings with adversarial augmentation. This ability readies NativE to deliver robust prediction capabilities in the wild while facing diversity and imbalance problems. In NativE, we design two new modules called relation-guided dual adaptive fusion and collaborative modality adversarial training to address the problems mentioned before.

4.1 Modality Encoding

To leverage the modality information, we first perform modality encoding to capture modality feature, which is a very common step in different MMKGC methods. The raw feature f_m of the entity e in the modality m is extracted as:

$$f_m = \frac{1}{|\mathcal{X}_m(e)|} \sum\nolimits_{x_{e,m} \in \mathcal{X}_m(e)} \mathbf{PE}_m(x_{e,m}) \tag{1}$$

where $x_{e,m}$ is one of the modality m elements for entity e and $\mathbf{PE}_m$ is the corresponding pre-trained encoder. The encoders, typically pre-trained on extensive datasets, extract deep semantic characteristics distinctive to each modality. For example, we can employ BERT [11] as the textual encoder and VGG16 [35] as the image encoder. The numeric information can be also extended as a sequence and employ BERT to capture its modality information. The pre-trained encoders are frozen to capture modality features and will not be fine-tuned during the training process. Moreover, the modal embedding for entity $e \in \mathcal{E}$ of modality $m \in \mathcal{M}$ is indicated as $\boldsymbol{e}_m = \mathcal{P}_m(f_m) \in \mathbb{R}^{d_e}$, where d_e is the embedding dimension and $\mathcal{P}_m$ is a projection layer for the modality m, aiming to project the different modal embeddings into the same vector space. Each $\mathcal{P}_m$ comprises a two-layer MLP with ReLU [13] as the activation function.

The modality encoding process yields distinct modal embeddings $\boldsymbol{e}_m$ for each entity e. In the MMKG context, embeddings from other modalities serve as auxiliary information to enhance the structure of the entity. For uniform representation of multi-modal information, we assign the structural modality as S and the structural embeddings $\boldsymbol{e}$ as $\boldsymbol{e}_S$. However, structural embeddings for entities deviate

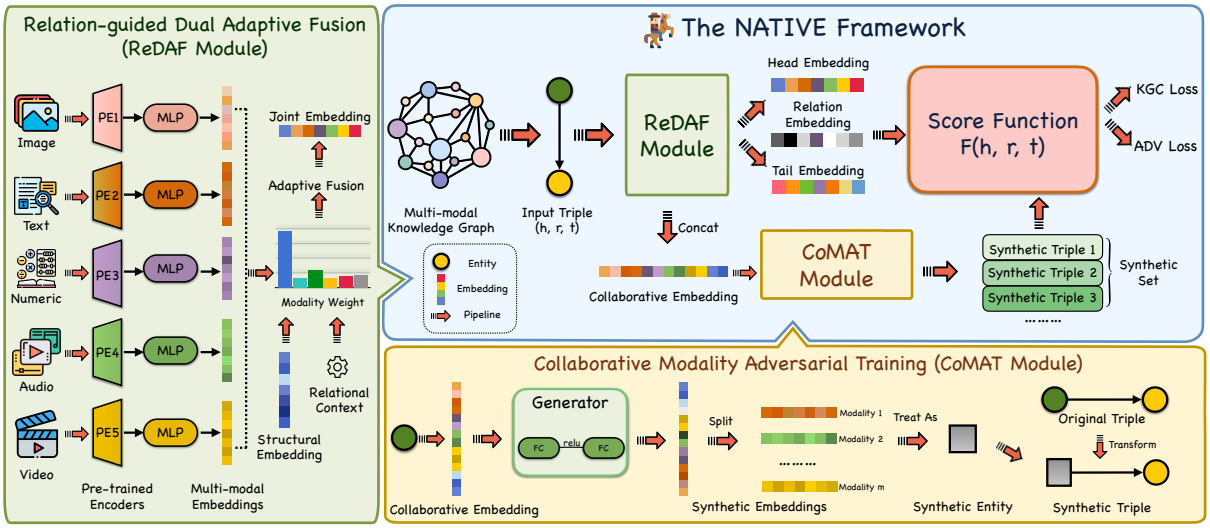


Figure 2: The overview of our NativE framework. NativE consists of two main modules called relation-guided dual adaptive fusion (ReDAF) module and collaborative modality adversarial training (CoMAT) module respectively. ReDAF is designed to fuse any input modality with modality adaptive weights and relational guidance. CoMAT aims to augment the imbalanced modality information in an adversarial manner by constructing synthetic triples to play a min-max game.

from other modalities. They are set to be **learnable parameters** to fit the triple structural information rather than relying on the modality encoding offered by pre-trained encoders.

4.2 Relation-guided Dual Adaptive Fusion

To extract the rich semantic information contained within entities, it is customary to fuse these multi-modal embeddings. Existing methods often employ fusion mechanisms such as concatenation [34], dot product [24], or gating [45] to accomplish modality fusion. Nevertheless, these methods are primarily designed for specific modalities such as text and images, which may not adequately serve scenarios intertwined with a richer mix of modalities. Furthermore, given the inherent imbalance in KGs, different modalities need to play different roles and offer diverse evidence for robust prediction within varying contexts.

To address the mentioned problems, we propose a **Relation-guided Dual Adaptive Fusion** (abbreviated as **ReDAF**) module within our framework. ReDAF includes an adaptive fusion mechanism using a set of adaptive weights ω_m for different modalities, which can be dynamically adjusted when the modal information is imbalanced. For example, when a certain modal is missing and the corresponding modal embedding is randomly initialized, the adaptive weight will be decreased for this modality. Meanwhile, we design a **relational-wise temperature** ζ_r to further modulate the weight distribution, thereby providing relational context for entity representation. This leads to dynamic weight adjustments populated both over different entities and under different relational contexts, which is the reason we name this module dual adaptive fusion. In more specific terms, the head entity h of the triple (h, r, t) has a relational context r and the adaptive weight for each modal m of h is denoted as:

$$\omega_m(h,r) = \frac{\exp(\mathcal{V} \odot \text{Tanh}(\boldsymbol{h}_m)/\sigma(\zeta_r))}{\sum_{n \in \mathcal{M} \cup \{S\}} \exp(\mathcal{V} \odot \text{Tanh}(\boldsymbol{h}_n)/\sigma(\zeta_r))} \tag{2}$$

where $\mathcal{V}$ is a learnable vector and $\odot$ is the point-wise operator. Tanh() is the tanh function. σ represents the sigmoid function to limit the relational-wise temperature in $(0, 1)$, aiming to amplify the differences between different modal weights. With the adaptive weights, the joint embedding of the head entity h is aggregated as:

$$\boldsymbol{h}_{joint} = \sum\nolimits_{m \in M \cup \{S\}} \omega_m(h,r)\boldsymbol{h}_m \tag{3}$$

The joint embedding $\boldsymbol{t}_{joint}$ of tail t can be also obtained similarly.

Classical MMKGC methods typically employ entity-centric modality fusion methods, regardless of their triple context. The strength of our design lies in its ability to dynamically modulate the modality weights for different entities across different triples, permitting these weights to be dynamically adjusted by their relational context. Therefore, ReDAF facilitates dual-adaptive multi-modal fusion, considering both the modal information of the entity and the relational context. Furthermore, ReDAF is a versatile modality fusion module, capable of processing **an unlimited number of input modalities** rather than considering specified modalities.

Once obtaining the joint embedding of entities, we employ a RotatE [37] score function to discriminate the triple plausibility, the score function is denoted as:

$$\mathcal{F}(h,r,t) = -||\boldsymbol{h}_{joint} \circ \boldsymbol{r} - \boldsymbol{t}_{joint}|| \tag{4}$$

where $\circ$ is the rotation operator in complex space. A higher score represents higher triple plausibility. We chose RotatE as the score function as RotatE can model the most common relational patterns. During training, we employ a negative sampling-based loss function to optimize the parameters, which can be represented as:

$$\begin{aligned}\mathcal{L}_{kgc} = &\sum\nolimits_{(h,r,t) \in \mathcal{T}} -\log \sigma(\gamma + \mathcal{F}(h,r,t)) \\ &- \sum_{i=1}^{K} p(h'_i, r'_i, t'_i) \log \sigma(-\mathcal{F}(h'_i, r'_i, t'_i) - \gamma)\end{aligned} \tag{5}$$

where σ is the sigmoid function; γ is a fixed margin; $(h'_i, r'_i, t'_i) \in \mathcal{T}', (i = 1, 2, \ldots, K)$ are K negative samples for triple (h, r, t). Besides, $p(h'_i, r'_i, t'_i)$ is the self-adversarial weight [37] proposed in RotatE, it can be denoted as:

$$p(h'_i, r'_i, t'_i) = \frac{\exp\left(\beta \mathcal{F}(h'_i, r'_i, t'_i)\right)}{\sum_{j=1}^{k} \exp\left(\beta \mathcal{F}(h'_j, r'_j, t'_j)\right)} \tag{6}$$

where β is a temperature to control the negative triple weight. This is a setting commonly used in KGC models.

4.3 Collaborative Modality Adversarial Training

The key function of the proposed ReDAF framework, despite achieving dual adaptive multi-modal fusion amongst imbalanced and diverse multi-modal data, is essentially feature selection for prediction, leaving the original imbalanced modality information intact. Inspired by the idea of generative adversarial networks [14, 41], we propose a **Collaborative Modality Adversarial Training** (abbreviated as **CoMAT**) module to augment the modality embeddings. The CoMAT module enhances the multi-modal embeddings through adversarial training, using entity-specific collaborative modality data to balance the multi-modal information distribution.

Drawing from the design principles of classical Wasserstein GAN (WGAN) [1], our goal is to establish a min-max game between the discriminator $\mathcal{D}$ and the generator $\mathcal{G}$ as:

$$\max_{\mathcal{D}} \min_{\mathcal{G}} \mathbb{E}_{x \sim p_{real}}[\mathcal{D}(x)] - \mathbb{E}_{x' \sim \mathcal{G}}[\mathcal{D}(x')] \tag{7}$$

where x is the data sample. In this min-max game, the discriminator $\mathcal{D}$ is assigned to discriminate the input data sample x with a score while the generator $\mathcal{G}$ aims to generate a synthetic data sample x'. With the adversarial training, the generator $\mathcal{G}$ adapts to the actual distribution of the data x and the discriminator $\mathcal{D}$ can learn to judge the plausibility of the input data.

In our scenario, the data x corresponds to the entity embeddings in the KG. As mentioned previously, each entity is represented by several modal embeddings, which can denoted as $\boldsymbol{e}_{real} = \{\boldsymbol{e}_{m_1}, \boldsymbol{e}_{m_2} \ldots; \boldsymbol{e}_{m_N}\}$ where $m_i \in \mathcal{M} \cup \{S\}$ is each modality. We expect to design a pair of $\mathcal{G}$ and $\mathcal{D}$ to learn the joint distribution of the multi-modal embeddings and augment the imbalanced multi-modal information. Therefore, $\mathcal{G}$ seeks to produce a collection of **synthetic embeddings**, defined as: $\boldsymbol{e}_{syn} = \{\boldsymbol{e}'_{m_1}, \boldsymbol{e}'_{m_2} \ldots; \boldsymbol{e}'_{m_N}\}$, and these embeddings can form a new **synthetic entity** denoted as e^*. By facilitating adversarial training between real and synthetic entities, we can enrich the multi-modal embeddings of the entity.

4.3.1 **The Design of Generator**. In the design of CoMAT, the generator $\mathcal{G}$ is implemented by a two-layer MLP. $\mathcal{G}$ processes the input random noisy z and the **concated** real multi-modal embedding $\boldsymbol{e}_{real}$ to generate augmented synthetic embeddings:

$$\boldsymbol{e}_{syn} = \mathcal{G}(\boldsymbol{e}_{real}, z) = \mathbf{MLP}_2\left(\delta\left(\mathbf{MLP}_1[\boldsymbol{e}_{real}, z]\right)\right) \tag{8}$$

where $\boldsymbol{e}_{real}$ is obtained by concatenating the different multi-modal embeddings collaboratively (as illustrated in Figure 2) and the generated synthetic embedding $\boldsymbol{e}_{syn}$ has the same shape of $\boldsymbol{e}_{real}$. We can split $\boldsymbol{e}_{syn}$ by dimension to yield the generated embedding $\boldsymbol{e}'_m$ of each modality generated by $\mathcal{G}$. The ultimate goal of this design is to emulate the probability distribution of the multimodal embedding for each entity. Given a triple (h, r, t), synthetic embeddings for both head and tail can be generated for both head and tail, denoted as $\boldsymbol{h}_{gen}$ and $\boldsymbol{t}_{gen}$ respectively. Besides, the corresponding synthetic entities are denoted as h^*, t^*.

4.3.2 **The Design of Discriminator**. The next step is to design the discriminator $\mathcal{D}$ to scoring the synthetic embeddings. In existing works in other fields [1, 28], $\mathcal{D}$ is usually implemented with another two-layer MLP. However, in the KGC task, the score function $\mathcal{F}$ mentioned in Equation 4 serves as a natural alternative for the discriminator. For a given triple (h, r, t) and the synthetic entities h^*, t^* generated by $\mathcal{G}$, we can construct a series of synthetic triples denoted as $\{(h^*, r, t), (h, r, t^*), (h^*, r, t^*)\}$. These synthetic triples' scores can be computed with the ReDAF module and the score function $\mathcal{F}$, indicating the plausibility of the synthetic embeddings in the triple context. We can denote these synthetic triples as a **synthetic set** $\mathcal{S}(h, r, t)$. Therefore, in the design of CoMAT, the adversarial training loss can be denoted as:

$$\mathcal{L}_{adv} = \sum_{(h,r,t) \in \mathcal{T}} \left(-\mathcal{F}(h, r, t) + \frac{1}{|\mathcal{S}|} \sum_{\substack{(h^*,r,t^*) \\ \in \mathcal{S}(h,r,t)}} \mathcal{F}(h^*, r, t^*) \right) \tag{9}$$

Note that the synthetic entities, analogous to real entities, also have several multi-modal embeddings like the real entities. They can obtain their joint embeddings with the ReDAF and calculate the final score subsequently. Therefore, the final min-max game between $\mathcal{D}$ and $\mathcal{G}$ is represented as:

$$\min_{\mathcal{D}} \max_{\mathcal{G}} \mathcal{L}_{adv} \tag{10}$$

According to the previous setting, the parameters of $\mathcal{G}$ encompass all the parameters in the two-layer MLP, while the parameters of $\mathcal{D}$ include the existing real embeddings, the extra parameters mentioned in ReDAF. The two parts are **iteratively** optimized with the adversarial loss to achieve convergence.

4.4 Overall Training Objective

The final training objective of NativE combines the above-mentioned KGC training loss $\mathcal{L}_{kgc}$ and the adversarial loss $\mathcal{L}_{adv}$ together. The discriminator $\mathcal{D}$ would minimize $\mathcal{L}_{kgc}$ and discriminate the synthetic entities generated by $\mathcal{G}$. Conversely, the generator $\mathcal{G}$ aims to generate high-score synthetic entities. Therefore, the training objective of $\mathcal{D}$ and $\mathcal{G}$ can be expressed separately as:

$$\begin{cases} \mathcal{L}_{\mathcal{D}} = \mathcal{L}_{kgc} + \lambda_1 \mathcal{L}_{adv} \\ \mathcal{L}_{\mathcal{G}} = -\mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{gp} \end{cases} \tag{11}$$

where λ_1, λ_2 are the loss weights, $\mathcal{L}_{gp}$ is the gradient penalty [16], commonly used for stable WGAN training, denoted as:

$$\mathcal{L}_{gp} = -\sum_{(h^*,r,t^*) \in \mathcal{S}(h^*,r,t^*)} \left(||\nabla \mathcal{F}(h^*, r, t^*)||_2 - 1 \right)^2 \tag{12}$$

These two losses will be iteratively minimized during training.

4.5 Theoretical Analysis

In the previous sections, we give an intuitive motivation and design for the CoMAT module. We further provide theoretical analysis

Table 1: Statistical information of the five MMKGs in our WildKGC benchmark. We report the statistical information in the KGs. For each modality, we list its feature dimension and the number of entities with the corresponding modal information.

Dataset	#Entity	#Relation	#Train	#Valid	#Test	Image		Text		Numeric		Audio		Video	
						Num	Dim	Num	Dim	Num	Dim	Num	Dim	Num	Dim
MKG-W [54]	15000	169	34196	4276	4274	14463	383	14123	384	-	-	-	-	-	-
MKG-Y [54]	15000	28	21310	2665	2663	14244	383	12305	384	-	-	-	-	-	-
DB15K [29]	12842	279	79222	9902	9904	12818	4096	9078	768	11022	768	-	-	-	-
TIVA [49]	11858	16	20071	2000	2000	11636	2048	11858	300	-	-	2441	128	10269	2048
KVC16K [31]	16015	4	180190	22523	22525	14822	768	14822	768	-	-	14822	768	14822	768

to discuss the rationale for our design. The success of the WGAN restriction discriminator's scoring function relies on its adherence to the K-Lipschitz condition, typically accomplished through a two-layer MLP with gradient penalty [1, 28]. However, our design employs the RotatE score function $\mathcal{F}$ as a discriminator. We can prove that function $\mathcal{F}$ is also a K-Lipschitz function for h and t.

Proposition. Let function $\mathcal{F} = -||\boldsymbol{h} \circ \boldsymbol{r} - \boldsymbol{t}||$. Then for any h_1, h_2 in the function domain, there exists a scalar K which satisfies:

$$||\mathcal{F}(h_1, r, t) - \mathcal{F}(h_2, r, t)|| \leq K||\boldsymbol{h}_1 - \boldsymbol{h}_2|| \tag{13}$$

Proof. According to the properties of trigonometric inequalities, we have the following derivations:

$$\begin{aligned} LHS &= \Big|\big||\boldsymbol{h}_1 \circ \boldsymbol{r} - \boldsymbol{t}|| - ||\boldsymbol{h}_2 \circ \boldsymbol{r} - \boldsymbol{t}||\big|\Big| \\ &\leq ||(\boldsymbol{h}_1 \circ \boldsymbol{r} - \boldsymbol{t}) - (\boldsymbol{h}_2 \circ \boldsymbol{r} - \boldsymbol{t})|| \\ &= ||(\boldsymbol{h}_1 - \boldsymbol{h}_2) \circ r|| = ||\boldsymbol{r}|| \times ||\boldsymbol{h}_1 - \boldsymbol{h}_2|| \end{aligned} \tag{14}$$

Besides, in the RotatE score function, the relation embeddings $\boldsymbol{r}$ are limited with unit modulus [37] which means $||\boldsymbol{r}|| \leq 1$ for any $r \in \mathcal{R}$. Let $K = 1$, the proof is complete. This establishes that $\mathcal{F}$ is K-Lipschitz for head h. A similar argument confirms the same for the tail t. From the preceding analyses, it emerges that our WGAN-based design, harnessing the score function from the MMKGC model, is theoretically sound and rational for scoring for the synthetic entities. It makes the design of the whole framework more collaborative and natural. Indeed, CoMAT serves as a versatile framework that can be extendable to other embedding-based MMKGC models as well. We will go through the empirical evaluation to further demonstrate its effectiveness and generalization.

5 EXPERIMENTS AND EVALUATION

In this section, we first introduce our experimental procedure and settings, followed by a comprehensive discussion of extensive experiments to highlight the strengths of our method across a variety of scenarios. The following six research questions (RQ) are the key questions that we explore in the experiments.

- **RQ1.** Can our model NativE outperform the existing baseline and make substantial progress in the MMKGC task?
- **RQ2.** Can NativE maintain robust performance in the MMKGC task when the modality information is imbalanced?
- **RQ3.** How much do each module in the NativE contribute to the final results? Are these modules reasonably designed?
- **RQ4.** Is the CoMAT strategy we designed universal and general enough to be applied in other MMKGC models?
- **RQ5.** How does the training efficiency of our model compare to existing methods?
- **RQ6.** Are there intuitive cases to straightly demonstrate the effectiveness of NativE?

5.1 Experiment Settings

5.1.1 **Datasets**. To better explore the MMKGC tasks in a more complex and diverse environment, we construct a new MMKGC benchmark including five different datasets. While four of the datasets are from prior studies, one is completely new. Our benchmark includes the following datasets:

- **MKG-Y** and **MKG-W** [54] are two MMKGs derived from YAGO [36] and Wikidata [42] with images and texts proposed by [54].
- **DB15K** [29] is an MMKG with image, text, and numerical information proposed by [29], which is a subset of DBpedia [23].
- **TIVA** [49] is an MMKG with image, text, audio, and video information modality with 12K entities.
- **KVC16K** [31] is modified from KuaiPedia [31], a video concept encyclopedia. We reorganize it into an MMKG and leverage the image/text/audio/video features provided by the original authors.

We refer to our benchmark as WildKGC with the 5 MMKGs. Detailed information about the datasets is presented in Table 1. The raw modality features are inherited from the original datasets.

5.1.2 **Task and Evaluation Protocol**. We conduct link prediction [3] task on the five datasets, which is a significant task of KGC. We have introduced the setting of link prediction in Section 3. Following existing works, we use rank-based metrics [37] like mean reciprocal rank (MRR) and Hit@K (K=1, 3, 10) to evaluate the results.

Besides, the filter setting [3] is applied to remove the candidate triples already existing in the training set for fair comparisons.

5.1.3 **Baseline Methods for Comparisons**. In our experiments, we employ 21 different state-of-the-art KGC and MMKGC models as our baselines for a comprehensive comparison and analysis. The baselines can be divided into four categories: uni-modal (conventional) KGC methods, multi-modal KGC methods, negative sampling methods, and numeric-aware KGC methods.

i) Uni-modal KGC methods. We select 5 state-of-the-art uni-modal KGC methods including **TransE** [3], **DistMult** [55], **ComplEx** [40], **RotatE** [37], **PairRE** [6], which design elegant score functions and learn the structural embeddings of the given KG without any multi-modal information.

ii) Multi-modal KGC methods. We employ 10 different MMKGC methods that consider both multi-modal information and the triple structural information including **IKRL** [53], **TBKGC** [34], **TransAE** [51], **RSME** [45], **MMKRL** [30], **VBKGC** [64], **OTKGE** [5], **IMF** [24], **QEB** [49] and **VISTA** [22]. Comparison with these methods can verify the effectiveness of our model NativE. Meanwhile, among these methods, **MMKRL** design an adversarial training framework. The design concepts of both methods bear some resemblance to our design of CoMAT.

Table 2: The main MMKGC results on WildKGC benchmark. We list the modalities considered by each method where S/I/T denotes Structure/Image/Text. "All" represents that the method can process any number of input modalities. The best results in baselines are underlined and we highlight the SOTA results in bold. The flag † denotes the adversarial training baselines.

Method	Modality	MKG-W		MKG-Y		DB15K			KVC16K			TIVA		
		MRR	Hit@1	MRR	Hit@1	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
Uni-modal KGC Methods														
TransE	S	29.19	21.06	30.73	23.45	24.86	12.78	47.07	8.54	0.64	23.42	83.85	83.20	84.15
DistMult	S	20.99	15.93	25.04	19.33	23.03	14.78	39.59	6.37	3.03	12.61	82.27	81.15	84.22
ComplEx	S	24.93	19.09	28.71	22.26	27.48	18.37	45.37	12.85	7.48	23.18	80.67	77.67	86.10
RotatE	S	33.67	26.80	34.95	29.10	29.28	17.87	49.66	14.33	8.25	26.17	84.59	83.47	86.95
PairRE	S	34.40	28.24	32.01	25.53	31.13	21.62	49.30	-	-	-	-	-	-
Multi-modal KGC Methods														
IKRL	S+I	32.36	26.11	33.22	30.37	26.82	14.09	49.09	11.11	5.42	22.39	67.71	63.72	75.67
TBKGC	S+I+T	31.48	25.31	33.99	30.47	28.40	15.61	49.86	5.39	0.35	15.52	81.57	78.75	86.05
TransAE	S+I	30.00	21.23	28.10	25.31	28.09	21.25	41.17	10.81	5.31	21.89	79.57	74.95	88.07
MMKRL†	S+I+T	30.10	22.16	36.81	31.66	26.81	13.85	49.39	8.78	3.89	18.34	85.03	81.92	90.10
RSME	S+I	29.23	23.36	34.44	31.78	29.76	24.15	40.29	12.31	7.14	22.05	40.01	30.55	51.35
VBKGC	S+I+T	30.61	24.91	37.04	33.76	30.61	19.75	49.44	14.66	8.28	27.04	74.07	66.87	85.85
OTKGE	S+I+T	34.36	28.85	35.51	31.97	23.86	18.45	34.23	8.77	5.01	15.55	35.28	30.45	41.98
IMF	S+I+T	34.50	28.77	35.79	32.95	32.25	24.20	48.19	12.01	7.42	21.01	55.46	41.87	77.57
QEB	All	32.38	25.47	34.37	29.49	28.18	14.82	51.55	12.06	5.57	25.01	74.25	66.10	88.35
VISTA	S+I+T	32.91	26.12	30.45	24.87	30.42	22.49	45.94	11.89	6.97	21.27	76.07	70.67	86.60
Negative Sampling Methods														
KBGAN †	S	29.47	22.21	29.71	22.81	25.73	9.91	51.93	13.72	7.54	25.88	85.44	82.45	90.10
MANS	S+I	30.88	24.89	29.03	25.25	28.82	16.87	49.26	10.42	5.21	20.45	85.70	82.70	90.62
MMRNS	S+I+T	35.03	28.59	35.93	30.53	32.68	23.01	51.01	13.31	7.51	24.68	83.12	83.05	83.25
NativE	S+I+T	36.58	29.56	39.04	34.79	36.74	26.87	**54.65**	15.26	8.56	28.29	91.27	90.57	91.27
NativE	All	**36.58**	**29.56**	**39.04**	**34.79**	**37.16**	**28.01**	54.13	**15.76**	**9.23**	**28.55**	**92.10**	**91.40**	**92.85**
Improvement		+4.42%	+3.39%	+5.40%	+3.05%	+13.71%	+15.75%	+5.24%	+7.50%	+11.47%	+5.58%	+7.46%	+10.05%	+2.46%

iii) Negative sampling methods. We employ 3 different negative sampling methods to compare with our method, including **KBGAN** [4], **MANS** [60], **MMRNS** [54]. Among these methods, **KBGAN** [4] is an adversarial negative sampling method designed for conventional KGC, which applies reinforcement learning to optimize the models. **MANS** [60] and **MMRNS** [54] are two negative sampling strategies designed for MMKGC, which utilize the multi-modal information to enhance the negative sampling process.

iv) **Numeric-aware KGC methods.** We employ 3 popular numeric-aware KGC methods including **KBLRN** [12], **LiteralE** [21], and **KGA** [43]. These methods consider the numerical information to enhance the KGC models. However, their design can not be generalized to other modalities such as image and text as they are designed only for numerical modality augmentation.

All of the selected baselines are embedding-based KGC and MMKGC methods. Other methods such as text-based methods [33, 56] or GNN-based methods [66] are not considered as they are orthogonal to our design.

5.1.4 **Implemention Details**. We implement our NativE framework based on OpenKE [17], a famous open-source KGC library. We conduct each experiment on a Linux server with Ubuntu 20.04.1 operating system and a single NVIDIA A800 GPU.

In the NativE, we fix the batch size to 1024 and set the training epoch to 1000. The embedding dimensions d_e, d_r are tuned from $\{150, 200, 250\}$ and the negative sampling number K is tuned from $\{32, 64, 128\}$. The margin γ is tuned from $\{3, 6, 9, 12\}$ and the tempreature β is set to 2. In the CoMAT, the random noise dimension is set to 64, and the coefficient λ_1, λ_2 are tuned from $\{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$. We optimize the model with Adam [20] and the learning rate is tuned from $\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ for both $\mathcal{G}$ and $\mathcal{D}$ in the adversarial training setting. For baselines, we reproduce the results on WildKGC following the methodology and parameter setting described in the original papers and their open-source official code. Some of the baseline results refer to MMRNS [54].

5.2 Main Results (RQ1)

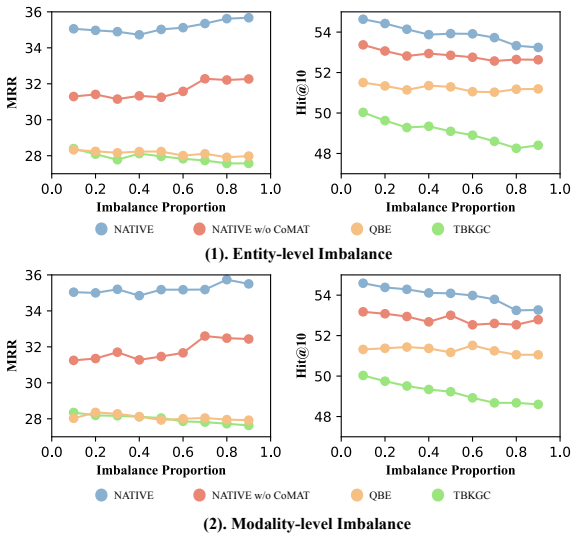
The main results of MMKGC experiments are presented in Table 2. We first compare our method NativE with 18 different KGC baselines. For equitable comparisons, we account for all modalities leveraged by each method and conduct two different sets of experiments on our model. The first set considers only modalities common to all datasets (structure/image/text) and the second set considers all the modalities inherent to each specific dataset.

Table 2 shows that NativE outperforms all the existing baselines and achieves new SOTA results. Specifically, NativE significantly surpasses the baselines by a large margin on Hit@1 on DB15K (15.75%), KVC16K (11.47%), and TIVA (10.05%), indicating a marked improvement in its accurate reasoning ability. When employing only the common modalities (structure/image/text) considered by most mainstream methods, NativE can still perform better than the baseline methods, signifying its ability to efficiently utilize information from different modalities. The performance comparison with two adversarial-based methods MMKRL [30] and KBGAN [4] demonstrates the superior effectiveness of our adversarial module design which can fully unleash the power of multi-modal information of the entities.

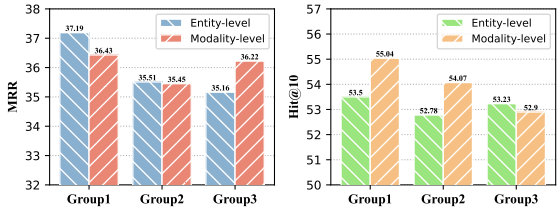
Besides, none of the MMKGC baselines in Table 2 has considered the numerical information presented in the DB15K. Accordingly, we conduct another experiment to compare our model NativE with other numerical KGC baselines. These methods' design principles

Table 3: Comparasions among numeric-aware methods on DB15K. S/N denote the structure/numeric modalitites.

Model	Modality	MRR	Hit@1	Hit@3	Hit@10
KBLN	S+N	24.68	19.26	27.27	34.89
LiteralE	S+N	31.29	24.24	34.59	44.98
KGA	S+N	33.62	25.82	37.73	47.85
NativE	S+N	36.18	27.27	41.25	52.51
NativE	All	**37.16**	**28.01**	**42.25**	**54.13**



(a) Imbalance MMKGC Experiments



(b) Group-wise Analysis for Modality-missing Triples

Figure 3: The imbalance MMKGC results. We report the MRR and Hit@10 results on the DB15K datasets. Further, we divide the test triples into three groups according to whether there was complete modal information and tally their experimental results separately, where: Group1 (both h and t are modality-complete); Group2 (one of h, r is modality-missing); Group3 (both h and t are modality-missing).

significantly differ from typical MMKGCs and lack generalizability to other modalities, so we compare them with numerical methods only on the DB15K dataset. From Table 3 we can observe that NativE still outperforms all baselines when considering only the structure and numeric modalities. As the diversity of modal information increases, the effectiveness of the model is further enhanced.

5.3 Imbalanced MMKGC Experiments (RQ2)

To better illustrate the performance of NativE in complex modality imbalance scenarios, we perform a series of MMKGC experiments

Table 4: The ablation study results on DB15K. We conduct three groups (G1/G2/G3) of experiments to validate the effectiveness of different modalities, the ReDAF module, and the CoMAT module respectively.

	Setting	MRR	Hit@1	Hit@10
G1	**w/o image**	36.88	27.08	54.63
	w/o text	36.09	26.86	52.92
	w/o numeric	36.74	26.87	54.65
G2	**w/o $\mathcal{P}_m$**	35.58	26.12	52.69
	w/o ζ_r	35.84	25.79	53.96
	w/o ω_m	35.65	25.23	54.10
	w/ concat	26.01	11.16	50.66
	w/ product	33.98	24.97	50.47
G3	**w/o CoMAT**	31.17	18.25	53.52
	w/o $\mathcal{L}_{gp}$	37.14	27.93	54.21
	w/ vanilla GAN	36.91	27.02	54.74
	w/ MLP as $\mathcal{D}$	36.34	27.12	53.31
Full NativE Model		37.16	28.01	54.13

in imbalance scenarios. We introduce a new parameter **imbalanced proportion** denoted as η to quantify the imbalance in datasets, representing the percentage of entities lacking modalities information. For instance, $\eta = 0.3$ means that there is 30% missing modality information in the given MMKG.

Specifically, we perform a random division of the original dataset randomly selecting the corresponding proportion η of modal information. Besides, we distinguish between two imbalance stratifications: entity-level imbalance and modality-level imbalance. For the entity-level imbalance division, we randomly drop all the modality information of a selected entity. For the modality-level imbalance, we randomly drop the modality information for a proportion η across all entities, implying an entity might exhibit varying modal information quantities. The missing modality information dropped in these settings will be initiated randomly and the MMKGC experiments on the imbalanced datasets are shown in Figure 3.

The results from the imbalanced experiments present that NativE maintains its performance when the modality information is imbalanced. As imbalance intensifies, the performance of the baseline models TBKGC [34] and QBE [49] experiences a significant downturn, while our methods show relative stability. Further, the impact of imbalanced modality is more noticeable on coarse-grained metrics such as Hit@10, suggesting that the completeness of modality information has a greater influence on coarse-grained entity ranking. Additionally, a surplus of irrelevant noise within the modality information hampers the accurate inference of the model, leading to a marginal improvement in the performance of NativE when the imbalance rate is increased. It is also noticeable that the CoMAT design has a significant effect on model performance, retaining its usefulness even in unbalanced situations.

We evaluate the imbalanced modality impact on different triples by counting separately the results of triple at different levels of missing entity modal information, which is shown in Figure **??**(b).

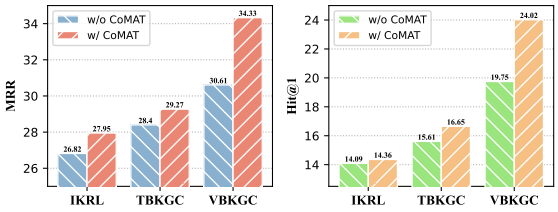


Figure 4: The generalization experiments of the CoMAT module on three different MMKGC models. We report the MRR and Hit@1 results on the DB15K dataset.

In this figure, the proportion $\eta = 0.5$. Interestingly, the model performance is not necessarily worse when modality information is fully missing (Group 3) in the triples compared to a partial missing (Group 2). This outcome arises because modality-level missing is much less likely to result in a complete lack of entity modal information, preserving some modality information instrumental for the final prediction, thereby leading to a relatively better performance of the model in this case (MRR of Group 3). Conversely, entity-level missing scenario deprives some entities of all modality information, causing a complete loss of valid information in the modality, leading to poor performance on coarse-grained metrics like Hit@10.

5.4 Ablation Study (RQ3)

We conduct extensive ablation studies to validate the effectiveness of each module in our design. In the ablation study, we conduct three groups of experiments to validate the contribution of each module in NativE by removing the corresponding key component. Each of the three groups of experiments targets the different modalities, the ReDAF module, and the CoMAT module. The ablation study results are summarized in Table 4. From the results in G1, we can conclude that each modality contributes to the overall performance, despite the image modal seemingly having a lesser role in comparison to the text. G2 shows the essential role of key components in the ReDAF module, as their removal leads to a marked performance decline. Besides, compared with other modality fusion strategies widely used by other MMKGC models like concat and dot-product, our NativE module still outperforms. From the results in G3, we can find that the CoMAT leads to a huge performance boost, especially on the precision metrics like Hit@1. Besides, when removing the gradient penalty loss or changing the adversarial loss to a vanilla GAN version with log-likelihoods [14] loss, the model performance still decreases. We also try to replace the discriminator $\mathcal{D}$ with a two-layer MLP, but the final result is worse than the original NativE. This suggests that the WGAN [1] framework employed by us works better for the MMKGC scenario, affirming the correctness of our prior analyses and proofs in the MMKGC setting.

5.5 Further Analysis (RQ4 & RQ5)

5.5.1 **Generalization Analysis (RQ4).** As we mentioned before, CoMAT is a general framework for enhancing MMKGC models based on embedding. To prove this point, we apply CoMAT on more MMKGC models (IKRL [53], TBKGC [34], and VBKGC [64]) and demonstrate the results in Figure 4.

We can make a conclusion that the MMKGC models trained w/ CoMAT can obtain significant performance gain. This suggests that CoMAT can be used as a general adversarial enhancement framework in different MMKGC models.

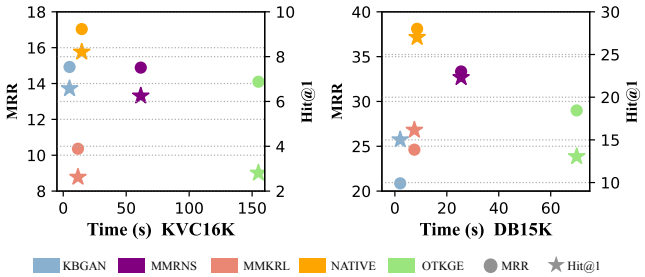
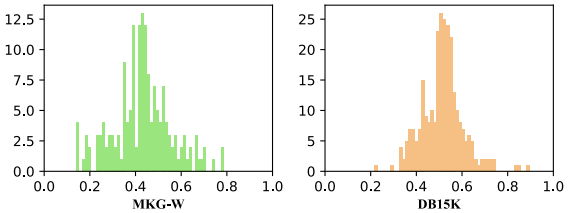


Figure 5: The results of the efficiency experiment. We report the MRR and Hit@1 results on the KVC16K/DB15K datasets.

5.5.2 **Efficency Analysis (RQ5).** Another important concern is the efficiency of the adversarial methods. This is because the iterative training strategies would lead to more computation. Therefore, we evaluate the efficiency of several MMKGC methods. The methods we have chosen cover both non-adversarial (OTKGE [5], MMRNS [54], IMF [24]) and adversarial approaches (KBGAN [4], MMKRL [30]). We evaluate the efficiency and performance of different models with the same batch size of 1024 and dimension of 250 on the same device. As shown in Figure 5, we can observe that NativE makes a good trade-off between efficiency and performance, achieving the best results while keeping relatively fast and efficient. Though the adversarial training module CoMAT slows down the training to some extent, this latency is within acceptable limits and there are significant gains in the model performance.

5.6 Case Study (RQ6)

This section presents clear examples to further elucidate our design. We previously emphasize that the relation guidance in ReDAF can zoom in on the differences between the different modalities and sift through them to find the useful parts, as the relation-wise temperatures are limited in $(0, 1)$ with a sigmoid function. We first demonstrate the distribution of the temperatures as shown in Figure 6. We can observe that the distribution is diversified



(a) The distrubution histgram of the $\sigma(\zeta_r)$.

Modality	Relations
Structure	GovType, Headquarter. SisterStation, SpokenIn
Image	(Film)Genre, Capital, Producer, Position
Text	Affiliation, Nationality, DeathPlace, IsPartOf
Numeric	Location, Leader, LeaderName, LeaderParty

(b) Relations raising up each modality weight.

Figure 6: The relation-wise temperature distribution and high-frequency relations that can raise each modality weight.

among relations. To better illustrate this, we filter out and list the most frequent relations that pull the weights of the corresponding modality higher for each modality, which can be found in Figure 6 (b). We can find that each different relationship will have a different dependence on the different modal information of the entity when making predictions. For example, image information can provide some intuitive visual information. When predicting movie *genre*, such image information would greatly benefit the result. Meanwhile, the *affiliation*, *nationality*, and *death place* of a person can be usually found in the text descriptions which makes text modality valuable when predicting the results of these relations.

6 CONCLUSION

In this paper, we propose a novel MMKGC framework NativE to rectify the diversity and imbalance issues of the MMKGs. NativE consists of two core designs called ReDAF and CoMAT. ReDAF proposes a relation-guided dual adaptive fusion method to incorporate adaptive features of any modalities to obtain joint representations while CoMAT aims to enhance the information about the imbalanced modes by a training strategy based on WGAN. We perform an in-depth theoretical analysis to justify our design's rationale. We construct a new MMKGC benchmark and conduct comprehensive experiments on it against 21 baselines to show the effectiveness, generalization, and efficiency of our framework. In the future, we think the MMKGC task and the downstream application of MMKG in more complex real scenarios remain great challenges to be solved.

ACKNOWLEDGMENTS

This work is founded by National Natural Science Foundation of China (NSFC62306276 / NSFCU23B2055 / NSFCU19B2027 / NSFC91846204), Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020017), Ningbo Natural Science Foundation (2023J291), Yongjiang Talent Introduction Programme (2022A-238-G), Fundamental Research Funds for the Central Universities (226-2023-00138).

REFERENCES

[1] Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR* abs/1701.07875 (2017).
[2] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 5184–5193.
[3] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*. 2787–2795.
[4] Liwei Cai and William Yang Wang. 2018. KBGAN: Adversarial Learning for Knowledge Graph Embeddings. In *Proc. of NAACL*.
[5] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport. In *NeurIPS*.
[6] Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. PairRE: Knowledge Graph Embeddings via Paired Relation Vectors. In *Proc. of ACL*.
[7] Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z. Pan, Wenting Song, and Huajun Chen. 2023. MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid. In *ACM Multimedia*. ACM, 3317–3327.
[8] Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. 2024. Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. *CoRR* abs/2402.05391 (2024).
[9] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples. In *ACL*. Association for Computational Linguistics, 2114–2119.
[10] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2D Knowledge Graph Embeddings. In *AAAI*. AAAI Press, 1811–1818.
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.
[12] Alberto García-Durán and Mathias Niepert. 2018. KBlrn: End-to-End Learning of Knowledge Base Representations with Latent, Relational, and Numerical Features. In *UAI*. AUAI Press, 372–381.
[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *AISTATS (JMLR Proceedings, Vol. 15)*. JMLR.org, 315–323.
[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proc. of NeurIPS*.
[15] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. 2023. Text with Knowledge Graph Augmented Transformer for Video Captioning. In *CVPR*. IEEE, 18941–18951.
[16] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. In *NIPS*. 5767–5777.
[17] Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. OpenKE: An Open Toolkit for Knowledge Embedding. In *Proc. of EMNLP*.
[18] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *ACL (1)*. The Association for Computer Linguistics, 687–696.
[19] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. Computer Vision Foundation / IEEE, 4401–4410.
[20] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
[21] Agustinus Kristiadi, Mohammad Asif Khan, Denis Lukovnikov, Jens Lehmann, and Asja Fischer. 2019. Incorporating Literals into Knowledge Graph Embeddings. In *ISWC (1) (Lecture Notes in Computer Science, Vol. 11778)*. Springer, 347–363.
[22] Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. 2023. VISTA: Visual-Textual Knowledge Graph Representation Learning. In *EMNLP (Findings)*. Association for Computational Linguistics, 7314–7328.
[23] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* (2015).
[24] Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. 2023. IMF: Interactive Multimodal Fusion Model for Link Prediction. In *WWW*. ACM, 2572–2580.
[25] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. 2024. Knowledge Graph Contrastive Learning Based on Relation-Symmetrical Structure. *IEEE Trans. Knowl. Data Eng.* 36, 1 (2024), 226–238.
[26] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. 2023. Learn from Relational Correlations and Periodic Events for Temporal Knowledge Graph Reasoning. In *SIGIR*. ACM, 1559–1568.
[27] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, X Liu, and F Sun. 2022. A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multimodal. (2022).
[28] Juncheng Liu, Zequn Sun, Bryan Hooi, Yiwei Wang, Dayiheng Liu, Baosong Yang, Xiaokui Xiao, and Muhao Chen. 2022. Dangling-Aware Entity Alignment with Mixed High-Order Proximities. In *NAACL-HLT (Findings)*. Association for Computational Linguistics, 1172–1184.
[29] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *ESWC (Lecture Notes in Computer Science, Vol. 11503)*. Springer, 459–474.
[30] Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. 2022. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.* 52, 7 (2022), 7480–7497.
[31] Haojie Pan, Yuzhou Zhang, Zepeng Zhai, Ruiji Fu, Ming Liu, Yangqiu Song, Zhongyuan Wang, and Bing Qin. 2022. Kuaipedia: a Large-scale Multi-modal Short-video Encyclopedia. *CoRR* abs/2211.00732 (2022).
[32] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *EMNLP*. Association for Computational Linguistics, 3208–3218.
[33] Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-Sequence Knowledge Graph Completion and Question Answering. In *ACL (1)*. Association for Computational Linguistics, 2814–2828.
[34] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In **SEM@NAACL-HLT*. Association for Computational Linguistics, 225–234.
[35] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[36] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*. ACM, 697–706.

[37] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR (Poster)*. OpenReview.net.

[38] Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding. In *Proc. of ACL*.

[39] Zhenwei Tang, Shichao Pei, Zhao Zhang, Yongchun Zhu, Fuzhen Zhuang, Robert Hoehndorf, and Xiangliang Zhang. 2022. Positive-Unlabeled Learning with Adversarial Data Augmentation for Knowledge Graph Completion. In *Proc. of IJCAI*.

[40] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML (JMLR Workshop and Conference Proceedings, Vol. 48)*. JMLR.org, 2071–2080.

[41] J. D. Tygar. 2011. Adversarial Machine Learning. *IEEE Internet Comput.* 15, 5 (2011), 4–6.

[42] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[43] Jiang Wang, Filip Ilievski, Pedro A. Szekely, and Ke-Thia Yao. 2022. Augmenting Knowledge Graphs for Better Link Prediction. In *IJCAI*. ijcai.org, 2277–2283.

[44] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *SIGIR*. ACM, 515–524.

[45] Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective. In *ACM Multimedia*. ACM, 2735–2743.

[46] Peifeng Wang, Shuangyin Li, and Rong Pan. 2018. Incorporating GAN for Negative Sampling in Knowledge Representation Learning. In *Proc. of AAAI*.

[47] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.* 29, 12 (2017), 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

[48] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *KDD*. ACM, 950–958.

[49] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and Audio. In *ACM Multimedia*. ACM, 2391–2399.

[50] Xiang Wang, Yaokun Xu, Xiangnan He, Yixin Cao, Meng Wang, and Tat-Seng Chua. 2020. Reinforced Negative Sampling over Knowledge Graph for Recommendation. In *WWW*. ACM / IW3C2, 99–109.

[51] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal Data Enhanced Representation Learning for Knowledge Graphs. In *IJCNN*. IEEE, 1–8.

[52] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *WWW*. ACM, 790–800.

[53] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *IJCAI*. ijcai.org, 3140–3146.

[54] Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced Negative Sampling for Multimodal Knowledge Graph Completion. In *ACM Multimedia*. ACM, 3857–3866.

[55] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR (Poster)*.

[56] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for Knowledge Graph Completion. *CoRR* abs/1909.03193 (2019).

[57] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *NAACL-HLT*. Association for Computational Linguistics, 535–546.

[58] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *AAAI*. AAAI Press, 2852–2858.

[59] Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022. Adversarial Retriever-Ranker for Dense Text Retrieval. In *ICLR*. OpenReview.net.

[60] Yichi Zhang, Mingyang Chen, and Wen Zhang. 2023. Modality-Aware Negative Sampling for Multi-modal Knowledge Graph Embedding. In *IJCNN*. IEEE, 1–8.

[61] Yichi Zhang, Zhuo Chen, Yin Fang, Lei Cheng, Yanxi Lu, Fangming Li, Wen Zhang, and Huajun Chen. 2023. Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering. *CoRR* abs/2311.06503 (2023).

[62] Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. 2024. Unleashing the Power of Imbalanced Modality Information for Multi-modal Knowledge Graph Completion. *CoRR* abs/2402.15444 (2024).

[63] Yichi Zhang, Zhuo Chen, Wen Zhang, and Huajun Chen. 2023. Making Large Language Models Perform Better in Knowledge Graph Completion. *CoRR* abs/2310.06671 (2023).

[64] Yichi Zhang and Wen Zhang. 2022. Knowledge Graph Completion with Pre-trained Multimodal Transformer and Twins Negative Sampling. *CoRR* abs/2209.07084 (2022).

[65] Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. 2022. MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion. In *EMNLP*. Association for Computational Linguistics, 10527–10536.

[66] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2021. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. In *NeurIPS*. 29476–29490.