Towards Multimodal Inductive Learning: Adaptively Embedding MMKG via Prototypes

Anonymous Author(s)

Abstract

Multimodal Knowledge Graphs (MMKG) models integrate multimodal contexts to improve link prediction performance. All existing MMKG models follow the transductive setting with a fixed predefined set, meaning that all the entities, relations, and multimodal information in the test graph are observed during training. This hinders their generalization to real-world MMKG with unseen entities and relations. Intuitively, a MMKG model trained on DBpedia cannot infer on Freebase. To address above limitations, we make the first attempt towards inductive learning for MMKG and propose a multimodal **Ind**uctive **MMKG** model (**IndMKG**) that is ***universal*** and ***transferable*** to any MMKG. Distinct from existing transductive methods, our model does not rely on specific trained embeddings; instead, IndMKG generates adaptive embeddings conditioned on any new MMKG via multimodal prototypes. Specifically, we construct class-adaptive prototypes to appropriately characterize the multimodal feature distribution of the given graph and equip IndMKG with robust adaptability to multimodal information across MMKGs. In addition, IndMKG learns non-specific structural embeddings based on meta relations. Such strategies tackle the challenge of notable multimodal feature discrepancies in cross-graph induction and allow the pre-trained IndMKG model to effectively zero-shot generalize to any MMKG. The strong performance in both inductive and transductive settings, across more than 20+ different scenarios, confirms the effectiveness and robustness of IndMKG. Our code is released at https://anonymous.4open.science/r/IndMKG.

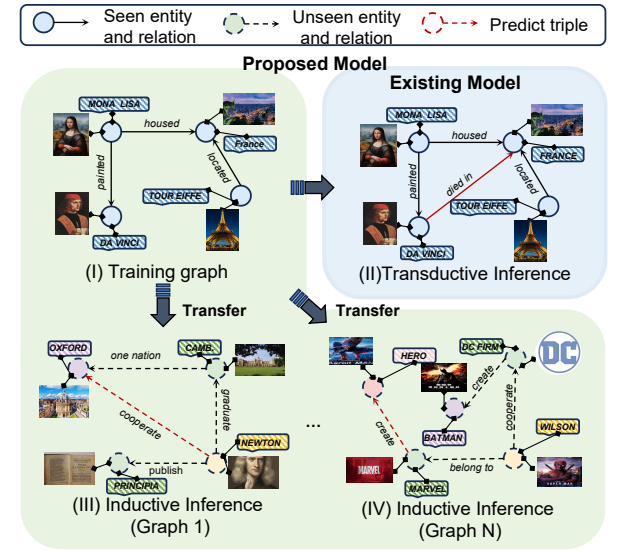
Keywords

Inductive Learning, Multimodal Knowledge Graph, Link Prediction, Knowledge Representation Learning and Embedding

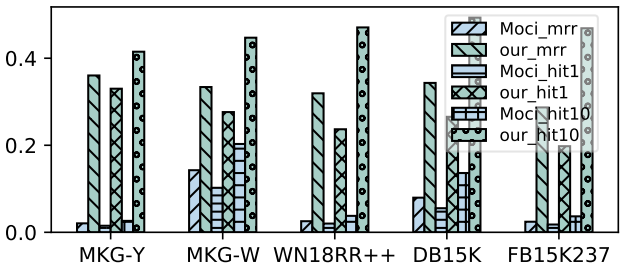
1 INTRODUCTION

Multimodal Knowledge graph (MMKG)[17] extends the representational richness of traditional knowledge graphs by integrating and utilizing the comprehensive multimodal attributes of given entities, typically including text, pictures, topologies, and is crucial for accurately revealing complex patterns of relations between potentially related entities. At present, MMKG has been extensively researched and applied in various fields, including intelligent search[37], personalized recommendation[29], bioinformatics[18], etc.

Link prediction[20, 23, 32] as a pivotal task in MMKG, aims to infer missing triples by integrating multimodal contexts[13, 22] to enhance the completeness of knowledge graphs, such as predicting the head or tail entity, namely ⟨?, r, t⟩ or ⟨h, r, ?⟩. However, existing MMKG models follow the transductive setting with a fixed predefined set, meaning that all the entities, relations, and multimodal information in the test graph are observed during training, as shown in Fig. 1a (I) and (II). In this setting, the entities, relations, and multimodal information in both the training and inference sets are the same, represented as $\mathcal{E}_{train} = \mathcal{E}_{inf}, \mathcal{R}_{train} = \mathcal{R}_{inf}$). Existing MMKG models make predictions rely on the specific trained embeddings, which means that a MMKG model trained on DBpedia[1] cannot infer on Freebase[2]. Consequently, they costly retrain whenever a new graph is introduced [8, 9, 27], since these models cannot handle new entities, relations, and multimodal information. This limitation hinders their generalization to real-world MMKG scenarios involving unseen entities and relations. Fig. 1b presents the performance of current MMKG models in inductive settings, revealing that even leading models struggle with effective cross-graph inductive inference.



(a) Examples of transductive and inductive inference



(b) Inductive Performance: Our Model vs. Existing SOTA MoCi[35]

Figure 1: (a) Task Comparison (b) Performance Comparison: Existing SOTA MMKG models achieve up to 46% of our model's induction performance (trained on YAGO 15K).

To overcome the limitations of existing transductive MMKG models, this paper presents the first attempt towards inductive

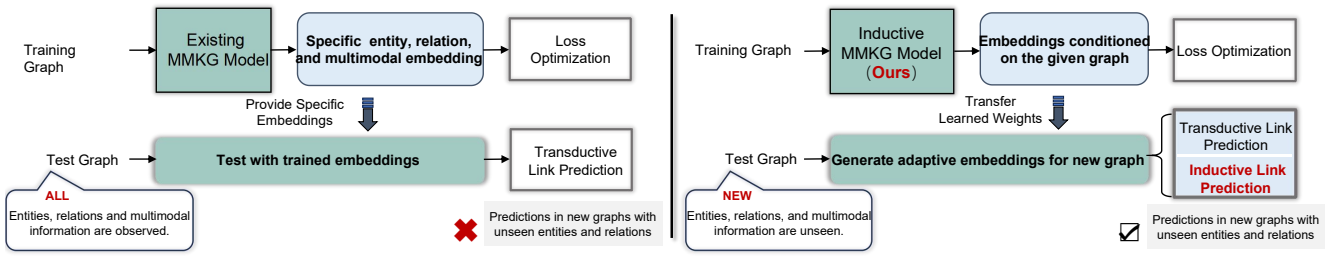


Figure 2: Framework comparison between existing transductive models and our inductive model.

learning for MMKG. We aim to propose a multimodal inductive MMKG framework (IndMKG) that is ***universal and transferable to any MMKG.*** However, inductive reasoning for MMKG presents a non-trivial challenge. To begin with, each MMKG encapsulates multimodal contexts that reflect its unique characteristics and domain-specific information. Furthermore, these graphs are shaped by various construction preferences and approaches to multimodal knowledge acquisition. As a result, there are significant differences in multimodal features among different MMKG. Consequently, a notable challenge emerges for models engaged in cross-graph induction: how to effectively address and leverage entirely unseen multimodal information that may differ significantly from the modalities observed in the training graph.

To address the challenge of multimodal feature disparity in MMKG inductive learning, we propose generating adaptive embeddings conditioned on the any graph through prototypes. Specifically, we propose a modality class-adaptive prototype learning strategy that dynamically models prototypes tailored to the given graph, allowing for the expression of its multimodal features rather than relying on fixed patterns. Furthermore, we implement prototype regularization and alignment to ensure compact embeddings within each class while preserving inter-class distinctiveness. For graph structure learning, we extract meta-relations across various relative positional contexts to produce structural embeddings. Finally, our model follows a dual-cue prediction that incorporates both structural cues and multimodal cues for the final predictions. Fig. 2 provides the framework comparison between the existing MMKG model and our proposed inductive MMKG model. Unlike transductive models that rely on learning specific embeddings, our model focuses on generating adaptive embeddings conditioned on the given graph, enabling transferable learning and facilitating effective cross-graph induction within MMKG.

To the best of our knowledge, we are the first to shift the MMKG paradigm from a traditional transductive setting to an inductive framework. We introduce a universal inductive MMKG model that avoids learning graph-specific embeddings, enabling zero-shot generalization across diverse MMKG. Additionally, our model is compatible with transductive link prediction, offering efficiency, flexibility, and ease of use. Our contributions are as follows:

- Our IndMKG is the first model for multimodal inductive link prediction, offering universal and transferable capabilities beyond the transductive setting of existing MMKG models.
- The proposed class-adaptive prototype learning addresses the challenge of multimodal feature discrepancies in cross-graph induction, enabling effective leverage of multimodal features to enhance the inductive performance of MMKG.
- We validated IndMKG in >20 scenarios with nodes ranging from 14,541-411,05 and edges from 26,638-3,101,16. In Zero-shot settings, IndMKG outperformed SOTA inductive model (non-multimodal) by up to 216%, while the existing best-performing MMKG models reached only 43% of our performance. Additionally, IndMKG exceeded SOTA transductive models with >10x fewer parameters.

2 RELATED WORK

2.1 Multimodal Knowledge Graph

MMKG integrates diverse modalities such as text and image data to enhance knowledge representation and link prediction. Models like MKGC[20], IKRL[31], AdaMF-MAT[39], and NativE[38] focus on effectively combining these modalities to enrich entity representations by projecting modality-specific information into unified embedding spaces. Additional models[3, 4, 7, 21, 28, 36] also leverage diverse modalities but often struggle to maintain the unique properties of each one. To address this challenge, IMF[15] emphasizes the independent learning of distinct modality-specific features, while MoCi[35] captures inter-entity modality semantics and integrates them to improve link prediction. Despite these advancements, existing MMKG models still operate under a transductive setting, where all entities, relations, and multimodal data in the test graph are observed during training. This restricts their ability to generalize to new entities, relations, or unseen multimodal information, often necessitating costly retraining when new graphs are introduced. Consequently, this limitation hinders their applicability in real-world MMKG scenarios involving unseen entities, relations, and multimodal information.

2.2 Inductive Learning

Although a number of inductive KG models have been proposed, they all have various limitations. The initial inductive KG models[9, 19, 26, 27] requires that unseen entities be associated with new entities. Then there are inductive models that can deal with entities that are completely unseen, but they require all relations to be seen in order to obtain embeddings of unseen entities or inductive inference through relational patterns, such as NBFNet[40], Grail[25], INDIGO[16], and Morse[6]. However, such models cannot deal with unseen relations, which means that they are no longer effective

when unseen relations occur. To the best of our knowledge, the first method that can solve the simultaneous occurrence of unseen entities and unseen relations is Maker[5], which realizes inductive reasoning by meta-learning. However, such methods need to build sub graphs, which is very expensive to calculate and cannot be extended to large datasets. Finally, INGRAM[12], HyRel[34], and Ultra[8] can leverage the shared structural information of KG to address unknown entities and relations for inductive reasoning. Improving the inductive inference performance by structure-shared encoding is limited since there are significant differences between diverse KGs, which makes it difficult to accurately capture fine-grained semantic differences cross modality, especially in MMKG. Additionally, these models do not account for multimodal information, which prevents them from fully utilizing the rich data available for inductive reasoning, thereby restricting their performance and application in multimodal contexts.

3 TASK FORMULATION

A knowledge graph is formalized as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}, \mathcal{R}$ denote the entity set and relation set, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ refers to triplets that describe relations between entities. In MMKG, each entity is associated with multi-modal context, including textual, visual, and structural information. We define the set of modalities as $M = \{s, v, t\}$, where s, v, and t denote structural, visual, and textual modalities, respectively.

Transductive Inference: Transductive Inference requires that all entities, relations, and multimodal information be observed during the training phase, with predictions limited to new graphs containing only seen entities and relations. In this setting, the entities and relations in the training and inference sets are the same, expressed as $\mathcal{E}_{train} = \mathcal{E}_{inf}$ and $\mathcal{R}_{train} = \mathcal{R}_{inf}$. Consequently, the model is evaluated on the same entities and relations it was trained on, utilizing the learned embeddings to make predictions.

Inductive Inference: In this paper, we pose a more challenging and universal task: inductive MMKG inference. Under the inductive setting, the test graphs are permitted to encompass entities and relations that are unobserved in the training set, or even entirely novel. To maximize the generalizability of the task, our focus primarily lies on the fully unseen scenario, where the test set is disjoint from the training set, meaning that the entities and relations in the inference graph are entirely unseen during training, i.e., $\mathcal{G}_{train} \cap \mathcal{G}_{inf} = \emptyset$. This implies that all entities and relations in the inference graph are new, expressed as $\mathcal{E}_{train} \cap \mathcal{E}_{inf} = \emptyset$ and $\mathcal{R}_{train} \cap \mathcal{R}_{inf} = \emptyset$. In this setting, our model leverages weight transfer learned during the training phase to generate adaptive embeddings for the given graph, allowing it to generalize and make predictions on these unobserved entities and relations, thereby enabling cross-graph induction within MMKG. Meanwhile, it is noteworthy that our proposed model, not being tailored to learn embeddings for specific graphs, is also compatible with partially unseen cases in the test set and equally applicable to the transductive setting.

4 METHODOLOGY

Existing MMKG models struggle to handle the differences in multimodal features between different graphs, especially when faced with completely unknown multimodal information, resulting in limited generalization capabilities. To address this issue, we extract meta-relations under different relative positional contexts, generating embeddings that are independent of specific entities, relations, and graph structures. Moreover, we propose modality class-adaptive prototype learning to generate embeddings of each modility that are conditioned on the given graph. Furthermore, we apply cross-modality entity and prototype alignment to ensure that the embeddings within each class are more compact while maintaining distinctiveness between classes. Finally, we implement dual cues prediction, which incorporates both structural cues from the graph and fused multimodal information cues for the final predictions. Fig.3 presents the overall architecture of IndMKG.

4.1 Non-specific Graph Structure Learning

Currently, MMKG learning methods all adopt transductive strategy and rely on entity and relation embeddings derived from a specific training graph, which limits their generalization ability to handle unseen entities and relations. To address this limitation, we propose non-specific graph structure learning method inspired by [10, 40], which leverages the naturally intrinsic relative positions of relations (called transferable meta-relations) to capture structural embeddings that are independent of any particular graph.

Specifically, We define four common types of meta-relations, denoted as $\mathcal{P} = \{h{-}h, h{-}t, t{-}h, t{-}t\}$, where $h{-}h$, $h{-}t$, $t{-}h$, and $t{-}t$ represent head-to-head, head-to-tail, tail-to-head, and tail-to-tail meta-relations, respectively. Building on the above, we construct the meta-relation graph $\mathcal{G}_r = (\mathcal{R}, \mathcal{P}, \mathcal{T}_p)$, where $\mathcal{T}_p = \{(r_1, p, r_2) \mid r_1, r_2 \in \mathcal{R}, p \in \mathcal{P}\}$. Given a query $(h, r, ?)$, we can obtain neighbor-aware relation embedding r_N via message passing over $\mathcal{G}_r$:

$$\mathbf{r}_N^{l+1} = A_{gg}\big(M_{sg}(\mathbf{r}_w^l, \mathbf{p})|w \in \mathcal{N}_p(r), p \in \mathcal{P}\big) \tag{1}$$

where $\mathbf{r}_N^{l+1}$ is the $l+1$-layer relation embedding integrated with neighbor information. $\mathcal{N}_p(r)$ indicates the set of neighbors for relation r at meta-relation p. The initial relation embedding of the 1-st layer is defined as $\mathbf{r}^1 = \mathbf{1}_{\mathbf{r}=r} * \mathbf{1}^d$. $A_{gg}()$ represents sum aggregation, and $M_{sg}()$ is a non-parametric DistMult function[8, 33]. Then, the relation embedding at layer $l+1$ is obtained by $\mathbf{r}^{l+1} = W_r[\mathbf{r}^l; \mathbf{r}_N^{l+1}]$, where $W_r \in \mathbb{R}^{2d \times d}$ is learnable parameter matrix used to aggregate neighbor-aware relation embeding and original relation embedding. Similarly, we can obtain the entity neighbor-aware embedding e_N via relation-assisted message passing over original $\mathcal{G}$:

$$\mathbf{e}_N^{l+1} = A_{gg}\big(M_{sg}(\mathbf{e}_w^l, Mlp(\mathbf{r}))|w \in \mathcal{N}_r(e), r \in \mathcal{R}\big) \tag{2}$$

where $\mathcal{N}_r(e)$ indicates the set of neighbors for entity node e with relation r. The initial entity embedding of the 1-st layer is defined as $\mathbf{e}^1 = \mathbf{1}_{\mathbf{e}=e} * \mathbf{r}$, where $\mathbf{r}$ corresponds to the relation embedding of the last layer obtained above. Then, the entity embedding at layer $l+1$ is obtained by $\mathbf{e}^{l+1} = W_e[\mathbf{e}^l; \mathbf{e}_N^{l+1}]$, where $W_e \in \mathbb{R}^{2d \times d}$ is learnable parameter matrix. Here, we define the learned entity embeddings of the last layer as structure features and denote as e^s. That above learning strategies allow IndMKG to take advantage of the meta-relations objectively present in any graph to generate flexible non-specific graph structure embeddings. Additionally, by observing these complex interaction patterns, it facilitates the understanding of unseen entities and relations in new graphs.

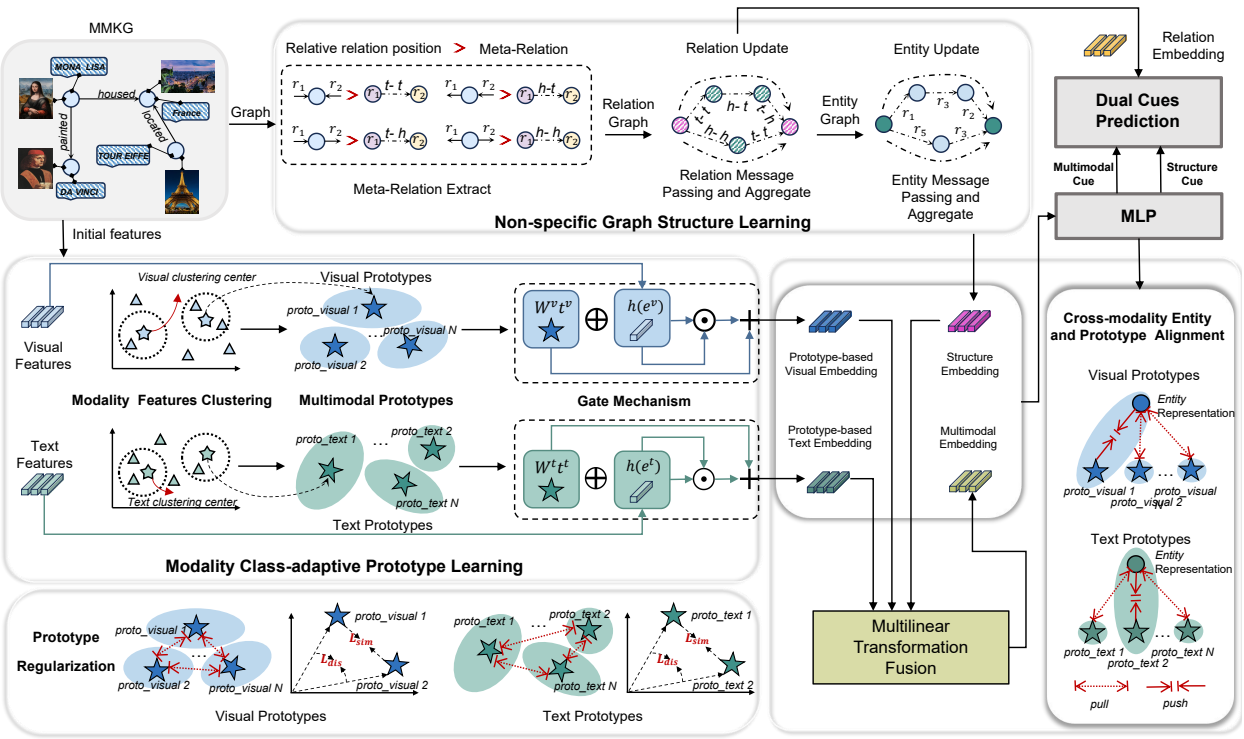


Figure 3: The overall architecture of the IndMKG model. IndMKG generates relation and entity structure embeddings using Non-specific Graph Structure Learning(Sec. 4.1). Modality Class-adaptive Prototype Learning(Sec. 4.2) produces graph-specific multimodal embeddings, while Cross-modality Entity and Prototype Alignment(Sec. 4.3) enhances prototype robustness and embedding expressiveness, followed by predictions using Dual Cues Prediction(Sec. 4.4).

4.2 Modality Class-adaptive Prototype Learning

Although multimodal information can provide rich context and details for entity representation, intra-modality variations and inter-modality semantic gaps, particularly in cross-graph scenarios, increase reasoning challenges. To effectively capture the multimodal distribution of the given graph-level features of all entities, and accurately describe the entity-level multimodal details of each entity, thereby improving the model's robust adaptability to multimodal information across MMKG, we propose the modality class-adaptive prototype learning strategy.

Specifically, IndMKG first performs K-means clustering on multimodal inputs of all entities to obtain K clustering centers of different modalities, defined as $Z^m \in \mathbb{R}^{K\times d}, m \in \{t, v\}$. Then, visual prototypes and textual prototypes are obtained by $C^v = Z^v W^v, C^t = Z^t W^t$, respectively, where W^v and $W^t \in \mathbb{R}^{d\times d}$ are learnable parameter matrices. For multimodal features of each entity $e^m \in \mathbb{R}^d, m \in \{t, v\}$, the corresponding i prototypes $c_i^m \in \mathbb{R}^d$, namely the i-th row of C^m, entities and prototypes are matched according to the modality-specific clustering centers. To accurately generate the entity-level features for each modality, we construct a feature gating unit by leveraging prototypes with global information in conjunction with entity-specific local features, which dynamically adjusts the contribution of different components. The prototype-based embedding is defined as follows:

$$\tilde{e}^m = c_i^m + \sigma\left(f((c_i^m) \oplus h(e^m))\right) \odot h(e^m) \tag{3}$$

where c_i^m is i-th prototype of modality m. $f()$, $h()$ are modality-shared single-layer MLPs used to transform the modality features into shared semantic spaces, and $\oplus$ is the concatenation operation used to fuse global prototype and local modality-specific feature. Moreover, $\sigma()$ and $\odot$ are sigmoid activation function and element-wise product for gating probability generation and feature filtering, respectively. Utilizing the gate mechanism in Eq. (3), we eliminate class-irrelevant information from the original modality feature e^m, resulting in consistent embeddings within each cluster.

To further enhance the distinctiveness of prototypes and alleviate semantic overlap among modality embeddings, we introduce prototype regularization strategy to encourage inter-class separation by constraining the similarity of different prototypes. We calculate the cosine similarity between prototypes as follows:

$$X^m = \overline{C^m} \cdot \overline{C^m}^\top \in \mathbb{R}^{K\times K} \tag{4}$$

where $\overline{C^m} \in \mathbb{R}^{K\times d}$ is normalized form of prototypes $C^m = [c_1^m; ...c_K^m]^\top$ of modality $m \in \{t, v\}$. Hence, x_{ij}^m denotes the cosine similarity between prototype c_i^m and c_j^m. To enhance the spatial distinction between prototypes, we aim to decrease the cosine similarities among them, thereby minimizing the $L_{2,1}$ norm of X^m.

$$\mathcal{L}_{sim} = \frac{1}{|m|} \sum_{m\in\{t,v\}} (\|X^m\|_{2,1}) \tag{5}$$

Considering the advantage of cosine similarity in capturing directivity and the intuitiveness of Euclidean distance in measuring

spatial distance, we further supplement the spatial distance perception of $\mathcal{L}_{sim}$ by introducing Euclidean distance. The contrastive strategy between different prototypes of modality m is defined as:

$$D^m_{ij,i\neq j} = \left\| c^m_i - c^m_j \right\|^2_2 \tag{6}$$

where $D^m = (d^m_{ij}) \in \mathbb{R}^{K\times K}$, represents the Euclidean distance between the prototypes c^m_i and c^m_k. To keep the distance between prototypes, we sort the elements of each row in the matrix D^m in increasing order to get $D^{m'} = (d^{m'}_{ik}) \in \mathbb{R}^{N\times N}$, and choose the top-$k$ minimum of each row to broaden them:

$$\mathcal{L}_{dis} = \frac{1}{2}\sum_{m\in\{t,v\}} \Big(\max(0, -d^{m'} + \gamma)\Big), d^{m'} = \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k} d^{m'}_{ij} \tag{7}$$

where γ is a hyperparameter that adjusts the distance margin. The loss for modality class-adaptive prototype learning is defined as:

$$\mathcal{L}_{pr} = \mathcal{L}_{sim} + \mathcal{L}_{dis} \tag{8}$$

The combined regularization loss ensures robust prototype learning.

4.3 Cross-modality Entity and Prototype Alignment

Aiming to enhance prototype robustness and entity embedding expressiveness, we design a loss function for fine-grained matching between entities and their corresponding prototypes, which compacts embeddings within each class while preserving inter-class distinctiveness. The alignment loss between entity and prototype is defined as follows:

$$\mathcal{L}_{aep} = -\frac{1}{2}\sum_{m\in\{v,t\}} \log \frac{\exp(\langle \overline{e}, \overline{c^m_i}\rangle/\tau)}{\sum_{i=0}^{K}\exp(\langle \overline{e}, \overline{c^m_j}\rangle/\tau)} \tag{9}$$

where $e \in \{e^s, e^f\}$, $\bar{\cdot}$ denotes the normalization operation, τ is a learnable temperature hyperparameter, and N is the number of modality prototype categories. Eq. (9) associates the multimodal features of each entity, namely e^s, $\tilde{e}^t$, and $\tilde{e}^v$ obtained by Eq.(2) and Eq.(3), with its matching prototypes with global semantics $c^m_i, m \in \{v, t\}$. It is worth noting that $e^f \in \mathbb{R}^d$ denotes the fused multimodal feature of given entity, which is obtained through the multilinear transformation of individual modality embeddings, and the specific operation is provided as below.

Multilinear Transformation Fusion: To achieve efficient multimodal context interaction between the structure embedding (e^s) and the prototype-based modal embedding ($\tilde{e}^t, \tilde{e}^v$), we employ a multilinear transformation fusion strategy [35]:

$$\tilde{\mathcal{E}}' = \tilde{\mathcal{E}} * \tilde{\mathcal{W}} = fold(bcirc(\tilde{\mathcal{E}}) \cdot unfold(\tilde{\mathcal{W}})) \tag{10}$$

where $\tilde{\mathcal{E}} \in \mathbb{R}^{|E|\times d\times|M|}$ is composed of structural, visual, textual features e^s, $\tilde{e}^t$ and $\tilde{e}^v \in \mathbb{R}^d$. Here $|E|$, and $|M|$ defaults to 3, denote the number of entity and the number of modality, respectively. $\tilde{\mathcal{W}} \in \mathbb{R}^{d\times d'\times|M|}$ is learnable parameters for multilinear feature transformation and fusion, and $*$ indicates the Tensor product operator. $bcirc(\tilde{\mathcal{E}}) \in \mathbb{R}^{|M||E|\times|M|d}$ means that performing block circulant unfolding on $\tilde{\mathcal{E}}$. The $unfold()$ operator flattens tensor $\tilde{\mathcal{E}}$ to matrix with size of $|M||E|\times d$, and $fold()$ corresponds to its inverse. More details about the Tensor product operation can be referred to [11, 30]. Subsequently, $\tilde{\mathcal{E}}' \in \mathbb{R}^{|E|\times d\times|M|}$ is summed along the 3-rd dimension to obtain the joint features after full multimodal context interaction and fusion, namely $E^f = \sum_m \tilde{\mathcal{E}}'_{:,:,m} \in \mathbb{R}^{|E|\times d}$.

After multilinear transformation fusion operation, the entity embedding $e^f \in \mathbb{R}^d$, and prototype embeddings $c^m, m \in \{t, v\} \in \mathbb{R}^d$, also pass through a parameter-shared MLP layer before computing alignment loss between entity and prototype according to Eq.(9).

4.4 Dual Cues Prediction

IndMKG follows a dual-cue prediction that incorporates both structural cues and multimodal cues for the final predictions. Specifically, we obtain a precise entity-relation composite representation through convolutional relational context, represented as $\mathbf{e}'^m = [\mathbf{e}^m; \mathbf{r}^m] * \omega$, where $m \in \{s, f\}$ and ω is the convolutional filter. Subsequently, $\mathbf{e}'^m$ will pass through the final Mlp: $\mathbb{R}^d \to \mathbb{R}^1$, which maps the node states to logits $pb(h, r, t)$, representing the score of node t as a potential tail of the initial query $(h, r, ?)$. We train by minimizing the binary cross entropy loss on positive and negative triplets:

$$\mathcal{L}_{bce} = \frac{1}{|2|}\sum_{m\in\{s,j\}} \Bigg(-\log pb(h^m, r, t'^m) - \sum_{i=1}^{n}\frac{1}{n}\log(1 - pb(h'^{m*}_i, r, t'^{m*}_i))\Bigg) \tag{11}$$

where negative samples generated by corrupting either the head h or the tail t of the positive sample. To optimize the model, we define a total loss function:

$$\mathcal{L} = \mathcal{L}_{pr} + \mathcal{L}_{aep} + \mathcal{L}_{bce} \tag{12}$$

In the subsequent experiments, we will discuss the effectiveness of each loss term adopted here in detail.

Table 1: Overview of Datasets.

Datasets	#Ent	#Rel	#Triplets		
			Train	Valid	Test
YAGO15K [17]	15,283	32	86,020	12,289	24,577
DB15K [17]	14,777	279	69,319	9,903	19,806
FB15K-237 [17]	14,541	237	272,115	17,535	20,466
WN18RR++ [14]	41,105	11	86,835	3,034	3,134
MKG-W [39]	15,000	169	34,196	4,276	4,274
MKG-Y [39]	15,000	28	21,310	2,665	2,663

5 EXPERIMENTAL SETUP

5.1 Datasets

In this paper, we extensively gathered nearly all existing MMKG datasets, including YAGO15K (based on YAGO[24]), DB15K (based on DBpedia[1]), and FB15K-237 (based on Freebase[2]) from MMKG [17], as well as WN18RR++ (based on WordNet) from VISTA[14], and MKG-W (based on Wikidata) and MKG-Y (based on YAGO) from AdaMF-MAT[39]. Table 1 provides a detailed statistical overview. These datasets encompass a wide range of information collected from various domains, such as film, sports, and education, demonstrating significant differences in both domains and content, which presents a substantial challenge for cross-graph inductive reasoning.

Table 2: The performance of IndMKG and baselines in inductive link prediction. $*$ represents the results obtained by retraining the model after zero-shot inductive inference.

Trained on DB15K

Model	DB15K to MKG-Y			DB15K to YAGO15K			DB15K to MKG-W			DB15K to WN18RR++			DB15K to FB15K237		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
INGRAM[12]$_{ICML'23}$	0.0260	0.0120	0.0750	0.0360	0.0190	0.0610	0.0860	0.0420	0.1060	0.0210	0.0130	0.0610	0.0980	0.0580	0.1750
*IMF[15]$_{WWW'23}$	0.0042	0.0017	0.0075	0.0585	0.0482	0.0723	0.0840	0.0811	0.0867	0.0223	0.0197	0.0272	0.0187	0.0148	0.0259
*MoCi[35]$_{ACMMM'24}$	0.0181	0.0142	0.0250	0.0919	0.0491	0.1804	0.1317	0.0991	0.1915	0.0265	0.0203	0.0383	0.0257	0.0187	0.0390
ULTRA[8]$_{ICLR'24}$	0.3272	0.2904	0.3919	0.3582	0.2790	0.5087	0.3048	0.2421	0.4169	0.2505	0.1529	0.4322	0.2068	0.1252	0.3773
HyRel[34]$_{ACMMM'24}$	0.0750	0.0310	0.0980	0.0960	0.0410	0.1210	0.1040	0.0710	0.1360	0.0430	0.0210	0.0970	0.1240	0.0670	0.1810
OUR zero-shot	**0.3692**	**0.3331**	**0.4354**	**0.4027**	**0.3280**	**0.5342**	**0.3382**	**0.2755**	**0.4559**	**0.2535**	**0.1570**	**0.4392**	**0.2510**	**0.1723**	**0.4090**
OUR fine-tuned	0.3940	0.3570	0.4550	0.4443	0.3820	0.5620	0.3779	0.3184	0.4943	0.5495	0.5020	0.6430	0.3624	0.2630	0.5520

Trained on YAGO15K

Model	YAGO15K to MKG-Y			YAGO15K to MKG-W			YAGO15K to WN18RR++			YAGO15K to DB15K			YAGO15K to FB15K237		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
INGRAM[12]$_{ICML'23}$	0.0830	0.0510	0.1210	0.0060	0.0020	0.0130	0.0310	0.0190	0.0430	0.0870	0.0430	0.1530	0.0130	0.0040	0.0260
*IMF[15]$_{WWW'23}$	0.0046	0.0021	0.0071	0.0959	0.0887	0.1035	0.0215	0.0188	0.0258	0.0725	0.0630	0.0865	0.0210	0.0165	0.0297
*MoCi[35]$_{ACMMM'24}$	0.0206	0.0151	0.0263	0.1427	0.1023	0.2030	0.0255	0.0201	0.0378	0.0796	0.0554	0.1363	0.0243	0.0179	0.0367
ULTRA[8]$_{ICLR'24}$	0.3513	0.3113	0.4299	0.2991	0.2319	0.4310	0.2753	0.1919	0.4294	0.2048	0.1333	0.3493	0.1596	0.0916	0.2983
HyRel[34]$_{ACMMM'24}$	0.1040	0.0630	0.1340	0.0230	0.0050	0.0410	0.0430	0.0210	0.0840	0.1140	0.0830	0.1670	0.0610	0.0230	0.0930
OUR zero-shot	**0.3604**	**0.3300**	**0.4150**	**0.3338**	**0.2763**	**0.4472**	**0.3194**	**0.2368**	**0.4708**	**0.3433**	**0.2654**	**0.4932**	**0.2870**	**0.1979**	**0.4689**
OUR fine-tuned	0.3958	0.3620	0.4583	0.3715	0.3130	0.4880	0.5351	0.4990	0.6070	0.4324	0.3720	0.5530	0.3514	0.2710	0.5180

Trained on WN18RR++

Model	WN18RR++ to MKG-W			WN18RR++ to YAGO15K			WN18RR++ to DB15K			WN18RR++ to MKG-Y			WN18RR++ to FB15K237		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
INGRAM[12]$_{ICML'23}$	0.0650	0.0240	0.0960	0.0570	0.0370	0.0910	0.0360	0.0210	0.0850	0.1030	0.0860	0.1350	0.0750	0.0350	0.1130
*IMF[15]$_{WWW'23}$	0.0875	0.0817	0.0941	0.0650	0.0522	0.0878	0.0649	0.0587	0.0705	0.0050	0.0026	0.0081	0.0213	0.0166	0.0300
*MoCi[35]$_{ACMMM'24}$	0.1356	0.1001	0.1985	0.0933	0.0478	0.1784	0.0803	0.0566	0.1375	0.0214	0.0163	0.0266	0.0228	0.0165	0.0359
ULTRA[8]$_{ICLR'24}$	0.1768	0.1319	0.2612	0.1776	0.1352	0.2487	0.1329	0.0843	0.2375	0.1352	0.0929	0.2119	0.1089	0.0560	0.2107
HyRel[34]$_{ACMMM'24}$	0.0790	0.0350	0.1130	0.0890	0.0390	0.1110	0.0460	0.0260	0.0890	0.1060	0.0890	0.1400	0.0910	0.0670	0.1440
OUR zero-shot	**0.2363**	**0.1821**	**0.3370**	**0.1804**	**0.1122**	**0.3170**	**0.1601**	**0.1063**	**0.2730**	**0.1804**	**0.1267**	**0.3005**	**0.1346**	**0.0731**	**0.2578**
OUR fine-tuned	0.3729	0.3140	0.4860	0.4413	0.3900	0.5410	0.4483	0.3820	0.5740	0.3946	0.3570	0.4600	0.3687	0.2790	0.5510

Trained on MKG-Y

Model	MKG-Y to MKG-W			MKG-Y to YAGO15K			MKG-Y to DB15K			MKG-Y to WN18RR++			MKG-Y to FB15K237		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
INGRAM[12]$_{ICML'23}$	0.0640	0.0310	0.1050	0.0120	0.0030	0.0180	0.0310	0.0190	0.0830	0.0040	0.0010	0.0130	0.0640	0.0230	0.0950
*IMF[15]$_{WWW'23}$	0.0961	0.0889	0.1054	0.0491	0.0243	0.0737	0.0724	0.0628	0.0857	0.0215	0.0188	0.0258	0.0208	0.0162	0.0288
*MoCi[35]$_{ACMMM'24}$	0.1377	0.1021	0.1975	0.0908	0.0487	0.1796	0.0852	0.0574	0.1384	0.0245	0.0197	0.0381	0.0254	0.0179	0.0412
ULTRA[8]$_{ICLR'24}$	0.2446	0.1753	0.3851	0.3413	0.2875	0.4303	0.2290	0.1554	0.3783	0.2701	0.1786	0.4215	0.1158	0.0719	0.2035
HyRel[34]$_{ACMMM'24}$	0.0870	0.0470	0.1090	0.0650	0.0140	0.0840	0.0930	0.0450	0.1540	0.0510	0.0160	0.0770	0.1070	0.0510	0.1350
OUR zero-shot	**0.3033**	**0.2381**	**0.4254**	**0.3688**	**0.3070**	**0.4884**	**0.2663**	**0.1921**	**0.4113**	**0.3133**	**0.2291**	**0.4695**	**0.1772**	**0.1075**	**0.3294**
OUR fine-tuned	0.3679	0.3040	0.4910	0.4472	0.3860	0.5590	0.4290	0.3670	0.5460	0.5373	0.4970	0.6230	0.3680	0.2710	0.5620

To validate the model's cross-domain performance across different scenarios, we conducted cross-graph inductive experiments.

5.2 Baselines

To demonstrate our model's performance in the multimodal inductive setting, we selected the most classic and recent SOTA inductive inference models: INGRAM[34], HyRel[34], and Ultra[8], which can handle unknown entities and relations with scalability, leveraging the shared structural information of KG for inductive reasoning. In the inductive setting, MMKG models do not fix the number of entities and relations, enabling inductive learning, with final results obtained through retraining after zero-shot inductive inference.

Similarly, to demonstrate our model's performance in the MMKG transductive setting, we selected several classic and SOTA MMKG models: MKGC[20], IKRL[31], AdaMF-MAT[39], VISTA[14], and NativE[38], which focus on effectively integrating various modalities to enrich entity representations by projecting modality-specific information into unified embedding spaces. IMF[15], in particular, emphasizes the independent learning of distinct modality-specific features, while MoCi[35] captures inter-entity modality semantics and integrates them.

5.3 Implementation Details

Our experiments were conducted on NVIDIA RTX L20 GPUs with 48GB of RAM. We configured the training process for 20 epochs, using a batch size of 64, with modality embedding dimensions set to 128 and the number of negative samples set to 512. We employed the Adam optimizer for parameter learning, setting its learning

Table 3: Transductive Results on MKG-W and WN18RR++.

Model	MKG-W			WN18RR++		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
IKRL[31]$_{IJCAI'17}$	0.323	0.261	0.440	0.381	0.302	0.474
MKGC[20]$_{SEMEAVL'18}$	0.312	0.239	0.438	0.369	0.290	0.469
INGRAM[12]$_{ICML'23}$	0.099	0.064	0.166	0.066	0.044	0.098
IMF[15]$_{WWW'23}$	0.345	0.288	0.454	0.474	0.439	0.543
HyRel[34]$_{ACMMM'24}$	0.126	0.079	0.216	0.077	0.049	0.124
AdaMF-MAT[39]$_{COLING'24}$	0.358	0.290	0.484	-	-	-
VISTA[14]$_{EMNLP'24}$	-	-	-	0.552	0.487	0.675
MoCi[35]$_{ACMMM'24}$	0.358	0.307	0.459	0.514	0.468	0.611
NativE[38]$_{SIGIR'24}$	0.365	0.295	-	-	-	-
ULTRA[8]$_{ICLR'24}$	0.338	0.276	0.457	0.495	0.451	0.595
OUR	**0.380**	**0.319**	**0.499**	**0.579**	**0.534**	**0.683**

Table 4: Transductive Results on MKG-Y and FB15K_237.

Model	MKG-Y			FB15K_237		
	MRR	Hit@1	Hit@10	MRR	Hit@1	Hit@10
IKRL[31]$_{IJCAI'23}$	0.332	0.303	0.382	0.309	0.232	0.493
MKGC[20]$_{SEMEAVL'18}$	0.312	0.281	0.363	0.297	0.229	0.494
INGRAM[12]$_{ICML'23}$	0.064	0.042	0.117	0.093	0.061	0.157
IMF[15]$_{WWW'23}$	0.358	0.330	0.406	0.367	0.273	0.557
HyRel[34]$_{ACMMM'24}$	0.164	0.098	0.297	0.122	0.086	0.175
VISTA[14]$_{EMNLP'23}$	-	-	-	0.380	0.287	0.571
MoCi[35]$_{ACMMM'23}$	0.388	0.356	0.449	0.369	0.276	0.554
NativE[38]$_{SIGIR'23}$	0.390	0.347	-	-	-	-
AdaMF-MAT[39]$_{COLING'23}$	0.385	0.343	0.457	-	-	-
ULTRA[8]$_{ICLR'23}$	0.351	0.314	0.416	0.358	0.250	0.565
OUR	**0.397**	**0.359**	**0.464**	**0.399**	**0.304**	**0.592**

rate to 5×10^{-4}. For the baseline methods, we utilized both their originally reported results and our reproduced results.

6 EXPERIMENTAL RESULTS

6.1 Inductive Link Prediction Performance

Zero-Shot Inference: IndMKG consistently surpasses all baseline models across various evaluation metrics, establishing new state-of-the-art inductive inference performance, as shown in Table 2.

Experimental results show that even SOTA transductive models, such as IMF[15] and MoCi[35], experience a decline in performance under inductive scenarios. Even at their best performance, (as seen on the MKG-W dataset), MoCi achieves only about **43%** of the performance of IndMKG. The reason lies in the fact that the reasoning capabilities of existing MMKG models depend on learning specific features and patterns from the training data. Yet, upon generalizing these models to novel, unseen graphs, the previously acquired specific embeddings fail to retain their efficacy. In contrast, our model does not rely on pre-trained embeddings; instead, it generates embeddings conditioned on the given graph for reasoning when faced with new graphs. Consequently, it demonstrates strong generalization and adaptability across all inductive scenarios.

Additionally, the SOTA inductive model (non-multimodal) ULTRA shows some generalization ability in MMKG inductive scenarios, mainly due to its transferable learning of graph structures. Table 2 shows that IndMKG exceeds ULTRA by up to **216%** (hit@1 for YAGO15 zero-shot inference to FB15K-237). IndMKG consistently outperforms in 20 inductive scenarios. For instance, compared to the best existing model, it exceeds MRR by 14.33%, Hit@1 by 13.21%, and Hit@10 by 14.39% in YAGO15 zero-shot inference to DB15K. This indicates that the class-adaptive prototypes modeled by IndMKG effectively characterize multimodal features, exhibiting robustness and high feature utilization efficacy, thus tackling the challenge posed by significant multimodal feature discrepancies between inductive graphs.

Fine-tuned: Table 2 also presents the performance of the short fine-tuned IndMKG (with one additional epoch), indicating that fine-tuning further enhances performance. Notably, in most cases, the fine-tuned performance is comparable to traditional transductive link prediction methods. This suggests that when handling a completely new graph, a simple fine-tuning of IndMKG can achieve results on par with retraining from scratch, while significantly reducing computational overhead. Therefore, IndMKG can achieve comparable results while minimizing computational overhead with only minor accuracy trade-offs.

6.2 Transductive Link Prediction Performance

In addition to achieving SOTA performance in *inductive* settings, IndMKG also surpasses all current *transductive* MMKG models across six diverse datasets, as shown in Table 3 and 4, IndMKG achieves an average advancement of about 4% in Hit@10. Despite not training specific embeddings for testing, as transductive models do, IndMKG demonstrates superior reasoning performance, validating its capability to generate adaptive embeddings conditioned on any graph. Notably, transductive models parameterize embeddings for all entities, relationships, and modalities, resulting in a linear increase in the number of parameters with the graph's size. In contrast, IndMKG does not require the parameterization of all embeddings, making it significantly more lightweight. For instance, the SOTA transductive model MoCi[35] has about 39 million and 232 million parameters for the DB15K and FB15K-237 datasets, respectively (IMF [15] DB15K: 71 million; FB15K-237: 79 million), whereas our model comprises only about 3.7 million parameters, reducing the parameter count by over **10**×. This exceptional performance underscores the versatility of the proposed IndMKG model across various tasks and scenarios, as well as its efficiency and applicability for large-scale MMKGs.

Table 5: Module Ablation studies of IndMKG.

A	*B*	*C*	Transductive Setting			0-shot Setting		
			MRR	H@1	H@10	MRR	H@1	H@10
			0.3964	0.3612	0.4733	0.2895	0.2208	0.4523
✓			0.4421	0.3680	0.5820	0.3324	0.2587	0.4811
✓	✓		0.4557	0.3892	0.5851	0.3387	0.2612	0.4843
✓	✓	✓	**0.4688**	**0.3977**	**0.6013**	**0.3433**	**0.2654**	**0.4932**

6.3 Module and Modality Ablation Study

Module ablation: To validate the effectiveness of each module in IndMKG, we conducted ablation studies on three key modules: *A* (Modality Class-adaptive Prototype Learning), *B* (Cross-modality Entity and Prototype Alignment), and *C* (Dual Cues Prediction), which also reflect the contribution of the corresponding loss functions to the overall performance. These experiments were performed on transductive (YAGO15K) and inductive (YAGO15 zero-shot inference to DB15K setting) link prediction, as shown in Table

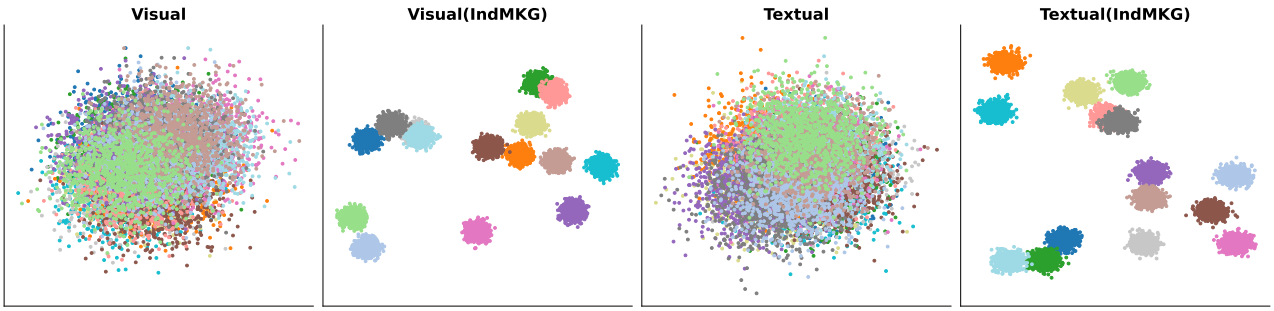


Figure 4: Comparison of the t-SNE visualization results of entity visual and textual feature distributions between the models without prototype processing and IndMKG on the YAGO15K dataset.

5. The ablation results highlight the importance of each component. Removing all three modules leads to a significant performance drop, while gradually reintroducing them yields consistent improvements. Their joint integration achieves the best results, showcasing the synergy and robustness of the IndMKG model.

Table 6: Evaluation Results of Modality Combinations.

Model	Transductive Setting			0-shot Setting		
	MRR	H@1	H@10	MRR	H@1	H@10
S	0.4049	0.3300	0.5550	0.2971	0.2223	0.4577
S+V	0.4328	0.3630	0.5740	0.3045	0.2322	0.4527
S+T	0.4379	0.3660	0.5770	0.3169	0.2456	0.4740
S+V+T	**0.4688**	**0.3977**	**0.6013**	**0.3433**	**0.2654**	**0.4932**

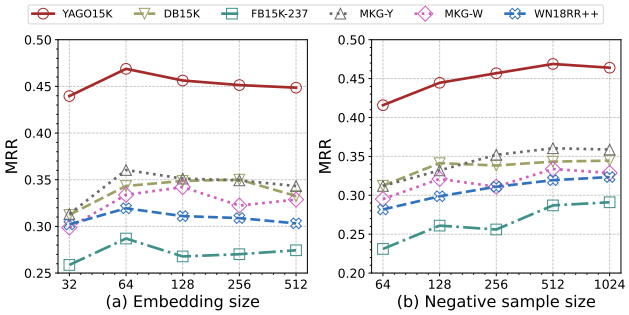


Figure 5: Impact of embedding size and negative sample size.

Modality ablation: To verify the impact of modality information on improving MMKG reasoning, we conducted ablation studies on modality combinations. These experiments were performed on transductive (YAGO15K) and inductive (YAGO15 zero-shot inference to DB15K setting) link prediction, as shown in Table 6. This study involved assessing the contributions of various combinations of modality embeddings, including structural (S), visual (V), and textual (T) embeddings. The results clearly indicate that relying on a single modality leads to the least effective performance, while the integration of multimodal information significantly enhances the results. This emphasizes our model's ability to adeptly leverage multimodal information through prototypes, thereby improving performance in both inductive and transductive tasks.

6.4 Parameter Analysis

Figure 5 illustrates the impact of embedding size on the performance of IndMKG. As shown in the figure, embedding size plays a crucial role in model performance. It is worth noting, however, that larger embedding sizes do not always lead to better performance due to potential overfitting issues. Considering performance, efficiency, and the inductive task setting, the optimal embedding size for IndMKG is 64. At the same time, Figure 5 demonstrates the impact of the number of negative samples on IndMKG's performance. The results suggest that an appropriate selection of negative samples is essential for optimizing the model's performance. Considering the previously discussed factors, the optimal number of negative samples for IndMKG is determined to be 512.

6.5 Case Study

To intuitively illustrate IndMKG's ability to produce compact and distinguishable embeddings, we utilized t-SNE visualization to display the distribution of entity visual and textual embeddings, as shown in Figure 4. The results indicate that IndMKG generates more compact and unique entity embeddings compared to models without prototype processing, visually demonstrating the advantages of modeling in the semantic space via prototypes. In this context, the multimodal embeddings learned by IndMKG exhibit a high level of inter-class distinctiveness and intra-class compactness. This intuitively demonstrates our method's superiority and explains why it is effective for inductive tasks in multimodal knowledge graphs, resulting in strong performance in MMKG reasoning.

7 CONCLUSION

In this study, we introduced the first inductive multimodal knowledge graph infrence model, IndMKG, which shifts the paradigm from traditional transductive approaches to a more flexible inductive framework. IndMKG demonstrates universal and transferable capabilities, effectively addressing the limitations of existing models that rely on specific trained embeddings. Our extensive evaluation across over 20 scenarios demonstrates that IndMKG surpasses existing SOTA models by up to 216% in zero-shot settings. Overall, IndMKG not only advances the state of the art in inductive MMKG link prediction but also broadens its applicability across diverse domains, paving the way for future research and real-world applications in multimodal knowledge representation.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*. Springer, 722–735.

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1247–1250.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Proc. of NeurIPS* 26 (2013).

[4] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems* (2022).

[5] Mingyang Chen, Wen Zhang, Zhen Yao, Xiangnan Chen, Mengxiao Ding, Fei Huang, and Huajun Chen. 2022. Meta-learning based knowledge extrapolation for knowledge graphs in the federated setting. In *International Joint Conferences on Artificial Intelligence Organization*. 1966–1972.

[6] Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. 2022. Meta-knowledge transfer for inductive knowledge graph embedding. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 927–937.

[7] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *ACM SIGIR*. 904–915.

[8] Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. 2024. Towards Foundation Models for Knowledge Graph Reasoning. In *The Twelfth International Conference on Learning Representations*.

[9] Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. 2017. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1802–1808.

[10] Xingyue Huang, Miguel Romero, Ismail Ceylan, and Pablo Barceló. 2024. A theory of link prediction via relational weisfeiler-leman on knowledge graphs. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Zhichao Huang, Xutao Li, Yunming Ye, and Michael K Ng. 2020. MR-GCN: Multi-Relational Graph Convolutional Networks based on Generalized Tensor Product.. In *IJCAI*, Vol. 20. 1258–1264.

[12] Chanyoung Chung Jaejun Lee and Joyce Jiyoung Whang. 2023. InGram: Inductive knowledge graph embedding via relation graphs. In *Proceedings of the 40th International Conference on Machine Learning*. 18796–18809.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* (2017).

[14] Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Whang. 2023. VISTA: Visual-Textual Knowledge Graph Representation Learning. In *EMNLP*.

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

[16] Shuwen Liu, Bernardo Grau, Ian Horrocks, and Egor Kostylev. 2021. INDIGO: GNN-based inductive knowledge graph completion using pair-wise encoding.. In *Proceedings of the 35th Conference on Neural Information Processing Systems*. 2034–2045.

[17] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *ESWC*.

[18] Zhi Liu, Jiaxi Yang, Kui Chen, Tao Yang, Xiaochen Li, Bingjie Lu, Dianzheng Fu, Zeyu Zheng, and Changyong Luo. 2024. TCM-KDIF: An Information Interaction Framework Driven by Knowledge-Data and Its Clinical Application in Traditional Chinese Medicine. *IEEE Internet of Things Journal* (2024).

[19] Sijie Mai, Shuangjia Zheng, Yuedong Yang, and Haifeng Hu. 2021. Communicative message passing for inductive relation reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4294–4302.

[20] Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *SEMEAVL*. 225–234.

[21] Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *AACL*. 225–234.

[22] Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González-Sánchez, and Roberto J López-Sastre. 2019. Answering Visual-Relational Queries in Web-Extracted Knowledge Graphs. (2019).

[23] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *EMNLP*.

[24] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. 697–706.

[25] Komal Teru, Etienne Denis, and Will Hamilton. 2020. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning*. 9448–9457.

[26] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *Proceedings of the 8th International Conference on Learning Representations*. 1–15.

[27] Changjian Wang, Xiaofei Zhou, Shirui Pan, Linhua Dong, Zeliang Song, and Ying Sha. 2022. Exploring Relational Semantics for Inductive Knowledge Graph Completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4184–4192.

[28] 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 MM*. 2391–2399.

[29] Zihan Wei, Ke Wang, Fengxia Li, and Yina Ma. 2024. M3KGR: A Momentum Contrastive Multi-Modal Knowledge Graph Learning Framework for Recommendation. *Information Sciences* (2024), 120812.

[30] Zhebin Wu, Lin Shu, Ziyue Xu, Yaomin Chang, Chuan Chen, and Zibin Zheng. 2022. Robust tensor graph convolutional networks via t-svd based graph augmentation. In *KDD*. 2090–2099.

[31] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *IJCAI*.

[32] Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. 2022. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *ACM MM*. 3857–3866.

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

[34] Jing Yang, Xiaowen Jiang, Yuan Gao, Laurence Tianruo Yang, JieMing Yang, et al. [n. d.]. Generalize to Fully Unseen Graphs: Learn Transferable Hyper-Relation Structures for Inductive Link Prediction. In *ACM Multimedia 2024*.

[35] Jing Yang, ShunDong Yang, Yuan Gao, JieMing Yang, and Laurence Tianruo Yang. 2024. Multimodal Contextual Interactions of Entities: A Modality Circular Fusion Approach for Link Prediction. In *ACM Multimedia 2024*.

[36] Xuan Zhang, Xun Liang, Xiangping Zheng, Bo Wu, and Yuhui Guo. 2022. MULTIFORM: few-shot knowledge graph completion via multi-modal contexts. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

[37] 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*. 91–101.

[38] 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 *SIGIR*. ACM, 91–101.

[39] 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. *LREC-COLING* (2024).

[40] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. In *Proceedings of the 35th Conference on Neural Information Processing System*. 29476–29490.