OTKGE: Multi-modal Knowledge Graph Embeddings via Optimal Transport

Anonymous Author(s)

Abstract

Multi-modal knowledge graph embeddings (KGE) have shown great power in learning representations of entities and relations for downstream tasks. Different from previous uni-modal KGE approaches, multi-modal KGE can leverage a wealth of multi-modal (textual, visual) knowledge and learn more realistic representations of real-world entities. However, the critical challenge along this course lies in that the multi-modal embedding spaces are usually heterogeneous, and direct fusion will destroy the inherent spatial structure of different modal embeddings, which may harm the interaction of multi-modal knowledge. To overcome this challenge, we innovatively revisit multi-modal KGE from a geometric perspective and propose *optimal transport knowledge graph embeddings* (**OTKGE**). Specifically, we model the multi-modal fusion procedure as a transport plan moving different modal embeddings to a unified aligned space by minimizing the Wasserstein distance between multi-modal distributions. Theoretically, we show the distribution differences between source multi-modal spaces and the unified space can be bounded by the Wasserstein distance and demonstrate the advantage of multi-modal KGE in generalization performance over uni-modal KGE. Experimental results on both well-established uni-modal and multi-modal knowledge graph completion benchmarks show that our OTKGE achieves the state-of-the-art performance.

1 Introduction

In the past decades, with the emergent and development of numerous knowledge graphs (KGs) [1, 2], a spectrum of related applications has been widely facilitated, *e.g.*, question answering [3, 4], semantic search [5, 6], and recommendation systems [7, 8]. However, due to the missing facts, real-world knowledge graphs are usually incomplete, which restricts the development of downstream tasks above. To address such an issue, Knowledge Graph Embedding (KGE) [9, 10, 11, 12, 13] emerges as an effective solution to predict missing links by the low-dimensional representations of entities and relations.

Nowadays, most of the KGE methods [11, 12, 13] mainly focus on the uni-modal knowledge graphs represented with pure symbols or concepts (*a.k.a.* structural information), however, there is profuse multi-modal (textual, visual) knowledge that has not been utilized. In recent years, researchers [14, 15, 16, 17] point out that the multi-modal knowledge can significantly improve the quality of learned knowledge graph representations. To that end, the study of the multi-modal KGE raises a new wave, and a great many meaningful work [18, 16, 19, 15, 20] emerges and achieves remarkable success. Specifically, they usually embed the multi-modal knowledge (*e.g.,* texts, pictures) of the entity into different spaces, then fuse these multi-modal embeddings with the help of operations such as concat or mean, hence obtaining a unified representation for the multi-modal entity. Under such circumstances, a dilemma could hinder the use of multi-modal knowledge: the embeddings of different modals are usually in various heterogeneous spaces, and the direct fusion might destroy

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

the intrinsic distribution, which is not conducive to the interaction of multi-modal knowledge.

Taking a step further, can we study a new KGE model, which is capable of learning unified representations of multi-modal entities while overcoming the problem of spatial heterogeneity?

To address this issue, we propose a novel multi-modal KGE method called *Optimal Transport Knowledge Graph Embeddings* (**OTKGE**) from a geometric perspective. Specifically, we formulate the fusion procedure of multi-modal knowledge as an optimal transport (OT) problem [21]. As shown in Figure 1, OT can move different modal embeddings to a unified aligned space by an optimal transport plan while overcoming spatial heterogeneity by minimizing the Wasserstein distance between different distributions. In this way, we can facilitate the information interaction between multi-modal embeddings and obtain a unified representation by the Wasserstein barycenter [22]. On top of this, we theoretically demonstrate that the knowledge representations learned from multi-modal KGE are usually better than those from uni-modal KGE. Moreover, experimental results demonstrate the superiority of OTKGE on both multi-modal and uni-modal knowledge graph completion tasks.

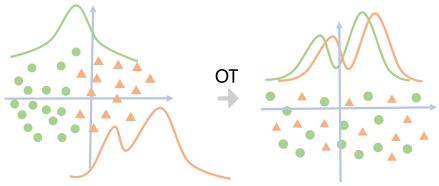


Figure 1: An illustration of the optimal transport (OT) for multi-modal entities. Green and red colors represent the different modals. Lines represent distributions of different modal entities while the circles and triangles represent the embeddings of different modal entities.

In a nutshell, our contributions are summarized as three-fold:

- For the modal fusion problem in multi-modal KGE, we propose a new fusion approach based on optimal transport, which can effectively tackles the modal spatial heterogeneity by reducing the Wasserstein distance between different modal distributions.
- Theoretically, we provide an upper bound for the target errors from source multi-modal spaces to the unified space in OT fusion. Moreover, we show a generalization bound of the multi-modal representation error in multi-modal KGE.
- Extensive experiments show that OTKGE achieves state-of-the-art performance on both uni-modal and multi-modal knowledge graph completion tasks.

2 Related Work

The study of knowledge graphs has attracted widespread attention in recent years. According to the number of used modalities, the previous methods mainly fall into two branches including uni-modal KGEs and multi-modal KGEs.

2.1 Uni-modal Knowledge Greaph Embeddings

In the past decades, there are various embedding approaches proposed to handle binary relational data, such as translational distance models, semantic matching models, and neural network models. Translational distance models are derived from the simple but effective method TransE [9], which models the relation as the translation between entities, and follows the principle $head\ entity + relation = tail\ entity$. Then there are many variants of TransE, *e.g.*, TransH [10], TransR [23], TransD [24], which use different projection vectors for entities and relations, respectively. However, they cannot model some relation patterns (*e.g.,* symmetric, inverse, and composition) [12].

Semantic matching approaches include bilinear models such as RESCAL [25], DistMult [26], ComplEx [11], and HolE [27]. These methods measure plausibility by matching latent semantics of entities and relations. RotatE [12] models the relation as the rotation in two-dimensional space, thus realizing modeling symmetry/antisymmetry, inversion, and composition patterns. Inspired by RotatE, there are also some variants of RotatE, such as QuatE [13], OTE [28], and RotH [29].

Recently there are some models using neural networks to produce KGE with remarkable effects, such as R-GCN [30], ConvE [31], ConvKB [32], KBGAT [33], and A2N [34].

2.2 Multi-modal Knowledge Graph Embeddings

Beyond uni-modal KGE methods, recently a great many multi-modal KGE methods have been proposed, which can be divided into neural network methods and translation methods.

Neural network methods utilize neural networks as the triple plausibility functions to learn structured representations and multi-modal representations. For instance, MMRFAN [17] and MKRL [14] embed the textual knowledge and visual knowledge of entities into a unified space to conduct multi-modal KGE. Unfortunately, these models such as MMRFAN are designed for medical knowledge graphs, and it may be not suitable for other general graph structures [20].

Translation methods follow the modeling principle of TransE and model the relation as the translation for multi-modal entities. For instance, DKRL [18] learns the representations of KGs by exploiting the textual and structural information while IKRL [16] learns representations by integrating visual knowledge and structured knowledge. However, both two models can only use one of textual knowledge and visual knowledge. To address this issue, [19] regards the plausibility of a triple as the sum of sub-energy functions that use textual information, visual information, and structured KGs simultaneously. Then TransAE [15] combines a multi-modal autoencoder with TransE aiming to learn representations of multi-modal KGs. Afterward, MMKRL [20] integrates structured knowledge and multi-modal knowledge in a unified space with the assistance of concat fusion operation. However, these models fuse heterogeneous multi-modal spaces by concat or mean operations, neglecting the differences between distributions. Therefore, learning embeddings for multi-modal knowledge graphs is still an open problem to be tackled.

3 Methodology

In this section, we introduce a new KGE method, termed *Optimal Transport Knowledge Graph Embeddings* (**OTKGE**) for multi-modal knowledge graphs. First of all, we present the definition of the multi-modal KGE problem.

Problem Setup. Given a multi-modal knowledge graph $G = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, and $\mathcal{T} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of KG triples. Here the entity e is equipped with structural (*a.k.a.* concept) embedding $\boldsymbol{e}_S$, linguistic embedding $\boldsymbol{e}_I$ and visual embedding $\boldsymbol{e}_V$. The uni-modal KGE method only focuses on $\boldsymbol{e}_S$ while multi-modal KGE can leverage all these embeddings. Practically, notice that the knowledge graph is usually incomplete, in this paper, our goal is to develop an efficacious multi-modal KGE method for the link prediction task.

Representations for Entities and Relations. In the KGE problem, entities can be represented as vectors in low-dimensional space while the relations can be modeled as geometric transformations. An important procedure in the KGE is to perform semantic matching between entities. Notice that algebraic groups can intuitively reveal the intrinsic geometric relationship between vectors, and embrace rich geometric transformations (*e.g.*, translations, rotations, and reflections), hence it can facilitate compact interactions between entities. Inspired by this, we model the relation r as the geometric group transformations as follows:

$$\left\{ \boldsymbol{R} \mid \boldsymbol{R} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{p} \\ \boldsymbol{q} & 1 \end{bmatrix}, \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^3 \right\} \tag{1}$$

where $\boldsymbol{A}$ is an orthogonal matrix representing the distance-preserving mapping. $\boldsymbol{p}$ and $\boldsymbol{q}$ can represent translations and projections, respectively. The more details for $\boldsymbol{R}$ can refer to Appendix B.

On top of the aforementioned representation, we study the multi-modal KGE. Supposing we have obtained the structural embedding (by uni-modal KGE or initialization) and multi-modal embeddings (by pretrained models) of the entity in KGs, then here our goal is to learn the unified embeddings. Due

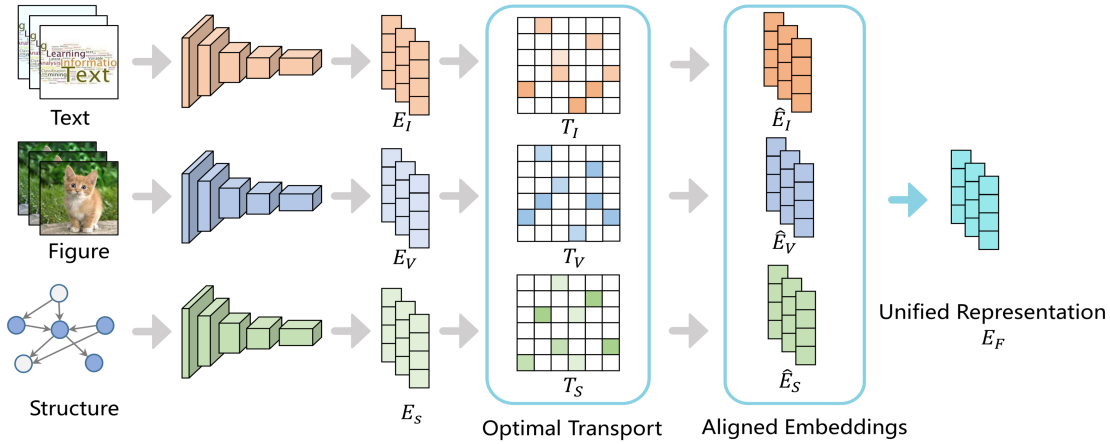


Figure 2: The overall framework for OT fusion. $\boldsymbol{E}_S, \boldsymbol{E}_I, \boldsymbol{E}_V, \boldsymbol{E}_F$ are structural, linguistic, visual, and unified embeddings, respectively. $\boldsymbol{T}_I, \boldsymbol{T}_S, \boldsymbol{T}_V$ are optimal transport plans. $\hat{\boldsymbol{E}}_S, \hat{\boldsymbol{E}}_I, \hat{\boldsymbol{E}}_V$ are the according aligned embeddings, respectively. In our method, $\boldsymbol{T}_S$ is an identity matrix and $\hat{\boldsymbol{E}}_S = \boldsymbol{E}_S$ since the multi-modal knowledge is to align with the structural knowledge.

to the heterogeneity of multi-modal spaces, one can notice that there exists a gap (*e.g.*, distribution discrepancy, dimensional difference) between different multi-modal embeddings. To address such an issue, as shown in Figure 2, we propose to realize the multi-modal alignment by optimal transport, and then perform multi-modal fusion to obtain a unified representation.

Multi-modal Alignment. Considering that multi-modal knowledge is auxiliary information, we let multi-modal embeddings align with the structural embeddings. Here we take the modal alignment between linguistic space $\mathcal{I}$ and structural embeddings $\mathcal{S}$ as an example, and the process between visual space $\mathcal{V}$ and structural embeddings can be achieved accordingly.

From this consideration, our interest is to develop an effective and intuitive approach for multi-modal alignment. Notice that optimal transport (OT) (more details can refer to Appendix A) is a widely used approach to make the source and target distribution closer by seeking a transport plan. Here we adopt the optimal transport as the alignment strategy, which can model the process of modal alignment as a process of transferring different modal embeddings into an alignment space (structural embedding space). Specifically, It can be divided into the following steps: 1) Estimate the distribution μ for linguistic embedding $\boldsymbol{E}_I$ and the distribution ν for structural embedding $\boldsymbol{E}_S$; 2) Find a transport map $\boldsymbol{T}$ from μ to ν; 3) Use $\boldsymbol{T}$ to transport $\boldsymbol{E}_I$ to the unified space $\mathcal{S}$ and form a new embedding $\hat{\boldsymbol{E}}_I$.

Following the aforementioned steps, we denote by $\boldsymbol{u} = \{u_i\}_{i=1}^n \in \Delta_n$ and $\boldsymbol{v} = \{v_j\}_{j=1}^m \in \Delta_m$ for the probabilistic simplexes of linguistic and structural embeddings, respectively. Let $\mu = \sum_{i=1}^n u_i \delta_{\boldsymbol{E}_I^i} \in \mathbb{P}(\mathcal{I})$, $\nu = \sum_{j=1}^m v_j \delta_{\boldsymbol{E}_S^j} \in \mathbb{P}(\mathcal{S})$ represent discrete distributions of linguistic and structural embeddings, respectively, where δ is the Dirac function. Then we denote the joint distribution $\prod(\mu,\nu) : \prod(u,v) = \{\boldsymbol{T} \in \mathbb{R}_+^{n\times m} | \boldsymbol{T}\mathbf{1}_m = \boldsymbol{u}, \boldsymbol{T}^T\mathbf{1}_n = \boldsymbol{v}\}$, where $\mathbf{1}$ denotes an all-one vector. Afterward, we compute the OT map $\boldsymbol{T}$ between the measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, which can be formulated as follows:

$$\mathcal{W}(\mu,\nu) = \inf_{\gamma \in \prod(\mu,\nu)} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\gamma}[c(\boldsymbol{x},\boldsymbol{y})] = \min_{\boldsymbol{T}\in\prod(\mu,\nu)} \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{T}_{ij}\boldsymbol{C}_{ij} \tag{2}$$

where $\boldsymbol{C}_{ij}$ is the cost function evaluating the distance between $\boldsymbol{E}_I^i$ and $\boldsymbol{E}_S^j$. Moreover, $\mathcal{W}(\mu,\nu)$ defines a distance that measures the distribution discrepancy between $\boldsymbol{E}_I$ and $\boldsymbol{E}_S$.

Multi-modal Fusion. After obtaining the transport matrix $\boldsymbol{T}$, the linguistic embedding $\boldsymbol{E}_I$ can be transformed into target aligned embedding $\hat{\boldsymbol{E}}_I$ with some popular alignment strategies [35, 36, 37]:

$$\hat{\boldsymbol{E}}_I = \mathrm{diag}(1/\boldsymbol{v})\boldsymbol{T}^T\boldsymbol{E}_I \tag{3}$$

Once this has been done, the transported linguistic embedding $\hat{\boldsymbol{E}}_I$ is in the same space as structural embedding $\boldsymbol{E}_S$. Moreover, we can also obtain the aligned visual embedding $\hat{\boldsymbol{E}}_V$ in a similar way. Then our next step is to fuse the aligned multi-modal embeddings ($\hat{\boldsymbol{E}}_I$ and $\hat{\boldsymbol{E}}_V$) with the structural embedding $\boldsymbol{E}_S$ utilizing the strategy inspired by the Wasserstein barycenter [22]:

$$\boldsymbol{E}_F = \min_{\boldsymbol{E}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{W}(\boldsymbol{E}_i, \boldsymbol{E}) \tag{4}$$

where $\boldsymbol{E}_F$ is the unified representation, $\mathcal{W}(\cdot,\cdot)$ is the Wasserstein distance, and N is the modal number (here N=3). Specifically, $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3$ are the structural embedding $\hat{\boldsymbol{E}}_S$, linguistic embedding $\hat{\boldsymbol{E}}_I$, and visual embedding $\hat{\boldsymbol{E}}_V$, respectively. Then we can solve Eq.(4) utilizing the algorithm in [38]. Moreover, for ease of use, the whole fused procedure can refer to the Algorithm 1 in Appendix C.

Relational Transformations. On top of the aforementioned process, we have obtained unified representations of entities. Then we transform the head entity h in the triple (h, r, t) with the assistance of r:

$$\boldsymbol{h}_{trans} = \boldsymbol{r} \otimes \boldsymbol{h} \tag{5}$$

where $\boldsymbol{h}$ is the embedding of the head entity h, $\boldsymbol{r}$ is the embedding of the relation r, $\boldsymbol{h}_{trans}$ is the embedding of transformed head entity, and $\otimes$ is the transformations imposed to h by r.

Score Function and Loss. Obtaining the transformed head entity, the score function formulates as follows:

$$f(h, r, t) = d(\boldsymbol{h}_{trans}, \boldsymbol{t}) \tag{6}$$

where $\boldsymbol{t}$ is the embedding of the tail entity t, and $d(\cdot,\cdot)$ denotes a distance metric (*e.g.,* inner product or Euclidean distance). Afterward, we formulate the task as a classification problem, and the model parameters are learned by minimizing the following loss:

$$L = \sum_{(h,r,t)\in\Omega\cup\Omega^-} log(1 + \exp(-Y_{hrt} f(h, r, t))) \tag{7}$$

where $Y_{hrt} \in \{-1, 1\}$ denotes the label of the triple (h, r, t), Ω is the set of observed triples, and $\Omega^- = \mathcal{E} \times \mathcal{R} \times \mathcal{E} - \Omega$ is the set of unobserved triples, which adopts the negative sampling strategies (*e.g.,* uniform sampling and Bernoulli sampling [10]). Noticing that the multi-modal information is actually already pre-trained in advance, here we update the structural embedding in the training. This is also consistent with our original intention of using multi-modal information to assist KGE.

4 Theoretical Analysis

In the section above, we (1) aligned multi-modal information to a unified space via optimal transport; (2) perform modal fusion and learn a unified representation. For this two-step process, one may ask (a) Can the distribution differences between source space and unified space be bounded? (b) Is the learned multi-modal unified representation by OTKGE better than that of the uni-modal KGE?

In face of the problem (a), we introduce the target error [39, 40] as an indicator to measure the distribution relationship between the target unified space and the source space. Here we first give some necessary definitions. Suppose $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{Y}$ are sets of entities, learned representations, and labels of triples, respectively. Notice that both structural embeddings and multi-modal embeddings of the entity are always assumed to obey the same true score function $f^* : \mathcal{Z} \to \mathcal{Y}$ in multi-modal KGE problems [18, 16, 19, 15, 20]. Considering the true scoring function f^* is unknown, one usually chooses a predictor function f from a hypothesis class $\mathcal{F}$ for substitution, $\forall f \in \mathcal{F}, f : \mathcal{Z} \to \mathcal{Y}$. To simplify the notations, here we take the head entity embedding $z \in \mathcal{Z}$ as the example to analysis while fixing the relation and the tail entity. The conclusion for the tail entity can be analyzed similarly.

Afterward, the probability that the hypothesis f disagrees with the score function f^* (in the distribution μ is defined as $\epsilon_\mu(f, f^*) = \mathbb{E}_{z\in\mu}[|f(z) - f^*(z)|]$, which is termed target error. For simplicity

of notation, we use the shorthand $\epsilon_\mu(f) = \epsilon_\mu(f, f^*)$. Moreover, we utilize the Wasserstein distance $\mathcal{W}(\cdot,\cdot)$ to relate the source multi-modal distribution and target distribution, and the target error in the OT fusion can be derived from the following theorem:

Theorem 1 (Upper bound for the target error in OT fusion.) *We denote by μ_S, μ_I, μ_V and μ_F for the distributions of structural, linguistic, visual, and fused unified embeddings. Assume the hypotheses $f \in \mathcal{F}$ are all K-Lipschitz continuous for some K. Then the following statements hold for every hypothesis $f, f^* \in \mathcal{F}$:*

(a) For the structural embeddings, we have:

$$\epsilon_F(f) \leq \epsilon_S(f) + \max\{\mathcal{W}(\mu_S, \mu_V), \mathcal{W}(\mu_S, \mu_I)\} \tag{8}$$

(b) For the multi-modal embeddings (here we take the linguistic mebeddings as an example), we have:

$$\epsilon_F(f) \leq \epsilon_I(f) + 2K\mathcal{W}(\mu_I, \mu_S) + \max\{\mathcal{W}(\mu_I, \mu_V), \mathcal{W}(\mu_I, \mu_S)\} + \lambda. \tag{9}$$

where λ is the combined error of the ideal hypothesis f^* that minimizes the combined error $\epsilon_I(f) + \epsilon_F(f)$.

Remark. The proof can refer to Appendix E. Thus one can see that target error can be bounded by the Wasserstein distance on the source multi-modal space and target unified space. Therefore, the problem (a) is answered. Moreover, one can notice that target errors are difficult to guarantee in other fusion approaches (such as concat and mean) since they cannot effectively reduce the distribution differences between spaces. This illustrates the theoretical advantages of using OT fusion in multi-modal KGE.

To answer the problem (b), here we recall some notions. Suppose $g^* : \mathcal{X} \to \mathcal{Z}$ is the true mapping from the input space (using all modalities) to the latent space *i.e.,* the process of learning unified representation through OT fusion, and $f^* : \mathcal{Z} \to \mathcal{Y}$ is the true task mapping, *i.e.,* the score function in KGE. $f^* \circ g^*(x) = f^*(g^*(x))$ represents the composite function of f^* and g^*. $L(\cdot,\cdot)$ is the loss function in Eq.(7), which is L-smooth with respect to the first coordinate, and can be bounded by a constant P. Our learning objective is to find $f \in \mathcal{F}$ and $g \in \mathcal{G}_K$ ($\mathcal{G}_K$ means using K modalities) to jointly minimize the empirical risk: $\hat{r}(f \circ g_K) := \min_{f\in\mathcal{F}, g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} L(f \circ g_K(\mathbf{x}_i), y_i)$ where $y_i \in \mathcal{Y}$ is the label of triple. Based on the aforementioned definitions, we define $r(h \circ g_K) = \mathbb{E}_{(\mathbf{x}_i, y_i)\sim\Omega}[\hat{r}(h \circ g_K)]$.

Notice that learning latent representations play important role in KGE, and the goodness of the potential representation usually has a critical impact on the performance of the model [9, 12]. In order to measure the goodness of a learned latent representation g, motivated by [41, 42, 43], we introduce the following definition of latent representation quality.

Definition 1 (Latent representation quality.) *For any learned latent representation mapping $g \in \mathcal{G}$, the latent representation quality is defined as*

$$\eta(g) = \inf_{f\in\mathcal{F}} [r(f \circ g) - r(f^* \circ g^*)], \tag{10}$$

where $\inf_{f\in\mathcal{F}} r(f \circ g)$ is the best achievable population risk with the fixed latent representation g. Thus, to a certain extent, $\eta(g)$ measures the loss incurred by the distance between g and g^*.

Following this, our goal is to estimate how close the learned latent representation $\hat{g}_K$ is to the true latent representation g^* as following theorem:

Theorem 2 (Generalization bound for the latent representation quality in multi-modal KGE.) *Let $\Omega = \{(x_i, y_i)\}_{i=1}^m$ be a dataset of m examples drawn i.i.d in knowledge graphs. Here $x_i \in \mathcal{X}$ is the head entity and $y_i \in \mathcal{Y}$ is the label of the triple while fixing relations and tail entities. Assuming we have produced the empirical risk minimizers $(\hat{f}_K, \hat{g}_K)$, for all $1 > \delta > 0$, the following formula holds with the probability at least $1 - \delta$:*

$$\eta(\hat{g}_K) \leq 4L\mathscr{R}_m(\mathcal{F} \circ \mathcal{G}) + 6P\sqrt{\frac{2ln(2/\delta)}{m}} + B \tag{11}$$

where B is the upper bound for $\hat{r}(\hat{f}_K \circ \hat{g}_K) - \hat{r}(f^* \circ g^*)$, and $\mathscr{R}_m(\cdot)$ is Redemacher complexity [44].

On top of this, we analyze the effect of the multi-modal information on the knowledge graph embedding. Similar to [45, 42, 43], suppose that entities in the knowledge graph have K modalities, and denote $\mathcal{G}_0$ and $\mathcal{G}_K$ as learning the unified representation with 0 and K multi-modalities. We can derive that $\mathscr{R}_m(\mathcal{F} \circ \mathcal{G}_K) \sim \sqrt{C(\mathcal{F} \circ \mathcal{G}_K)}$ and $\mathscr{R}_m(\mathcal{F} \circ \mathcal{G}_0) \sim \sqrt{C(\mathcal{F} \circ \mathcal{G}_0)}$, where $C(\mathcal{F})$ represents the intrinsic property of function class $\mathcal{F}$ [44]. Based on the knowledge of Rademacher complexity, we have $C(\mathcal{F} \circ \mathcal{G}_K) \leq C(\mathcal{F} \circ \mathcal{G}_0)$. In this sense, one can derive that multi-modal KGE outperforms uni-modal KGE if the following formula holds:

$$\hat{L}\left(\hat{f}_0 \circ \hat{g}_0, \Omega\right) - \hat{L}\left(\hat{f}_K \circ \hat{g}_K, \Omega\right) \geq \sqrt{\frac{C\left(\mathcal{F} \circ \mathcal{G}_K\right)}{m}} - \sqrt{\frac{C\left(\mathcal{F} \circ \mathcal{G}_0\right)}{m}} \tag{12}$$

Remark. The proof can refer to Appendix F and the result for the tail entity can be derived accordingly. Empirically, Eq.(12) usually holds, thus the multi-modal KGE usually works better than the uni-modal KGE. Moreover, one can observe that: on one hand, the impact of the intrinsic complexity of function classes will be reduced when the sample size m is larger, and the performance difference between multi-modal KGE and uni-modal KGE will be also reduced. On the other hand, using more modalities can efficiently optimize the empirical risk, in this way it will benefit the model performance by improving the latent representation quality. Then the problem (b) is tackled.

Moreover, for the ability to model key patterns and the connection between OTKGE and other models, please refer to Appendix D.

5 Experiments

5.1 Experimental Setup and Results

Dataset. We evaluate our approach on the link prediction task using five standard competition benchmarks as shown in Table 5 in Appendix G, including multi-modal datasets: WN9-IMG [46] and FB-IMG [19]; uni-modal datasets: WN18RR [31], FB15K-237 [31] and YAGO3-10 [47]. The details for these datasets can refer to Appendix G.

Evaluation metrics. For each case, the test triples are ranked among all triples with masked entities replaced by entities in the knowledge graph. Hit ratio with cut-off values $n = 1, 3, 10$ dented as H@n and the Mean Reciprocal Rank (MRR) are reported in the experiments. We follow the standard evaluation protocol in the filtered setting [9]: all true triples in the KG are filtered out during evaluation, since predicting a low rank for these triples should not be penalized.

Baselines. We compare our method to SOTA models, including multi-modal KGE methods including IKRL [16], TBKGE [19], TransAE [15], MMKRL [20]; uni-modal KGE methods including TransE [9], DistMult [26] TuckER [48], ComplEx [11], RotatE [12], QuatE [13], MuRP [49] and RoTH[29]. The Implementation details can refer to Appendix G.

Results on multi-modal datasets. To demonstrate the superiority of OTKGE, we compare it with other popular models on WN9-IMG and FB-IMG datasets as shown in Table 1. One can see that OTKGE achieves the best performance on both datasets. On one hand, the uni-modal KGE methods such as TransE can only learn the structural knowledge while the rich multi-modal information cannot be used effectively. Therefore, they do not perform as well as OTKGE, which demonstrates the advantage of multi-modal KGE. On the other hand, the multi-modal KGE methods such as IKLR use concat or mean to fuse information from multiple modalities, which will be constrained by spatial heterogeneity. Therefore, their experimental performance is weaker than OTKGE, which demonstrates the effectiveness of OT fusion.

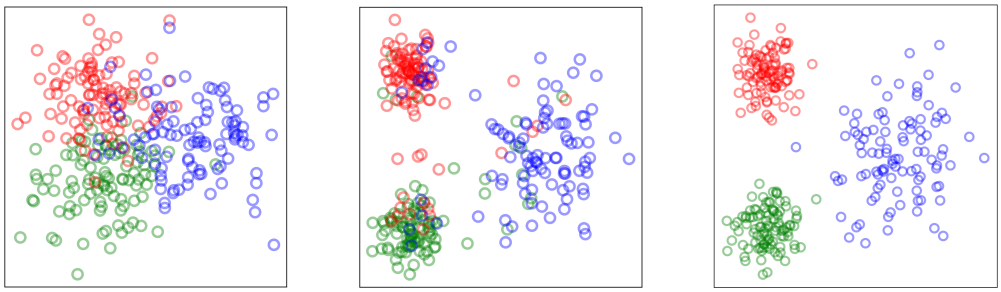
Results on uni-modal datasets. We compare OTKGE with some popular strong baselines in uni-modal datasets as shown in Table 2. Due to page limitations, we report the result on YAGO3-10

Table 1: Link prediction results on the multi-modal datasets (WN9-IMG and FB-IMG). If the indicator Is_M of a model is $\surd$, it means the model is a multi-modal KGE; otherwise, it is a uni-modal KGE. Best results are in **bold**, and second-best results are underlined.

		FB-IMG				WN9-IMG			
Model	Is_M	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
TransE	$\times$	.443	.427	.459	.494	.865	.765	.816	.871
RotatE	$\times$	.569	.538	.595	.654	.932	.841	.913	.945
RoTH	$\times$	.577	.546	.598	.658	.935	.844	.917	.948
QuatE	$\times$	.581	.548	.601	.660	.933	.842	.915	.946
TransAE	$\surd$	.563	.531	.592	.649	.930	.835	.908	.942
IKLR	$\surd$	.427	.414	.432	.479	.929	.815	.896	.938
TBKGE	$\surd$	.562	.521	.582	.645	.931	.838	.914	.948
MMKRL	$\surd$	.584	.567	.604	.661	.938	.853	.923	.952
OTKGE	$\surd$	**.601**	**.583**	**.615**	**.673**	**.945**	**.864**	**.933**	**.961**

datasets in Table 6 in Appendix G, and one can see that OTKGE can achieve the best performance for these datasets. Notice that these datasets embrace rich relation patterns (e.g., symmetry, inverse, composition, transitivity, hierarchy, intersection, and mutual exclusion patterns), which shows the advantage of the ability to model inference patterns in OTKGE.

5.2 Experiment Analysis



(a) Embeddings in TransE. (b) Embeddings in MMKRL. (c) Embeddings in OTKGE.

Figure 3: The embedding comparison Figure 3(a) represents structural embeddings learned by TransE, Figure 3(b) represents multi-modal unified embeddings learned by MMKRL utilizing concat fusion and Figure 3(c) presents the multi-modal unified embeddings learned by OTKGE.

The embedding comparison for OTKGE and other methods. To show that OTKGE can learn a better-unified representation, we choose three kinds of semantic (people, place, and event) entities in the FB-IMG dataset and visualize the entity embeddings using T-SNE following [50] as shown in Figure 3. In these figures, circles in blue (people), green (place), and red (event) represent three kinds of semantic embeddings, respectively. On one hand, one can see that using only structural knowledge in TransE can not separate these three types of embedding well, which demonstrates the necessity of leveraging multi-modal knowledge. On the other hand, there are still many entities that are not clearly separated in MMKRL due to the heterogeneity of modal space. Different from this, OTKGE can effectively reduce the Wasserstein distance between different distributions, facilitate the in-depth fusion of different modal knowledge, and thereby learn a better-unified representation.

The ablation study of components in OTKGE. As shown in Table 3, we conduct the experiment to study the role that different components play in OTKGE. Specifically, we remove the group transformation in OTKGE, which we denote as OTKGE w/o trans; we replace the fusion procedure via optimal transport by the mean and concat operations, which we denote as OTKGE w/ mean and

Table 2: Link prediction results on uni-modal datasets (WN18RR and FB15K-237 datasets). If the indicator Is_M of a model is $\surd$, it means the model is a multi-modal KGE; otherwise, it is a uni-modal KGE. Best results are in **bold**, and second-best results are underlined.

		WN18RR				FB15k-237			
Model	Is_M	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
A2N	$\times$	.430	.410	.440	.510	.317	.232	.348	.486
ComplEx	$\times$	.440	.410	.460	.510	.247	.158	.275	.428
RotatE	$\times$	.476	.428	.492	.570	.338	.241	.375	.533
QuatE	$\times$	.485	.438	.498	.562	.348	.248	.382	.550
TuckER	$\times$	.470	.443	.482	.526	.358	.266	.394	.544
ROTH	$\times$	.484	.441	.504	.570	.344	.246	.380	.535
ATTH	$\times$	.486	.443	.499	**.573**	.348	.252	.384	.540
DualE	$\times$	.482	.440	.497	.561	.330	.237	.363	.518
MuRP	$\times$	.481	.440	.495	.566	.335	.243	.367	.518
TransAE	$\surd$	.430	.407	.443	.522	.320	.225	.341	.503
IKLR	$\surd$	.402	.391	.412	.503	.311	.208	.325	.469
TBKGE	$\surd$	.437	.408	.451	.514	.315	.216	.331	.479
MMKRL	$\surd$	.463	.432	.473	.544	.332	.241	.364	.515
OTKGE	$\surd$	**.495**	**.449**	**.508**	.571	**.371**	**.276**	**.410**	**.560**

OTKGE w/ concat; We only use the visual knowledge, which we denote as OTKGE w/o ling; we only use the linguistic knowledge, which we denote as OTKGE w/o visual. We can see that the performance of these ablation versions is reduced compared to the original version of OTKGE, which shows that group transformations and optimal transformation are all indispensable parts of OTKGE.

Table 3: Link prediction results on WN9-IMG and FB-IMG datasets.

	FB-IMG				WN9-IMG			
Model	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
OTKGE w/o trans	.593	.575	.610	.668.	.941	.858	.927	.955
OTKGE w/ mean	.572	.536.	.588	.659.	.932	.839.	.917	.950
OTKGE w/ concat	.588	.572	.606	.665	.939	.857	.924	.954
OTKGE w/o ling	.585	.571	.605	.663	.938	.855	.924	.951
OTKGE w/o visual	.595	.576	.608	.665	.940	.858	.929	.955

6 Limitations

In this paper, we study the multi-modal KGE based on binary relations. However, in the real world, multi-modal information often exists in the form of beyond-binary relations [51]. Such limitation might be further tackled through learning binary and beyond-binary relations jointly and integrated through mechanisms such as attention, which we will study in future work.

7 Conclusion

In this paper, we design a new KGE model named OTKGE for multi-modal knowledge graphs, where we formulate the fusion procedure by optimal transport theory. Previous work usually neglects the heterogeneity of distributions in the multi-modal fusion, which will do harm to the interaction of multi-modal knowledge. To tackle this problem, we propose to transfer the multi-modal information to a unified space by optimal transport, and fuse the multi-modal information by the Wasserstein barycenter. Theoretically, we prove that OTKGE is advantageous with its capability in learning multi-modal representation and give the representation generalization bound for OTKGE. Moreover, empirical experimental evaluations on both uni-modal and multi-modal well-established datasets show that OTKGE can achieve overall state-of-the-art performance.

References

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

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD*, pages 1247–1250, 2008.

[3] Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. Neural network-based question answering over knowledge graphs on word and character level. In *WWW*, pages 1211–1220, 2017.

[4] Nouha Othman, Rim Faiz, and Kamel Smaïli. Learning english and arabic question similarity with siamese neural networks in community question answering services. *Data Knowl. Eng.*, 138:101962, 2022.

[5] Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *WWW*, pages 1271–1279, 2017.

[6] Duankang Fu, Shufan Zhou, Beijun Shen, and Yuting Chen. Enhancing semantic search of crowdsourcing IT services using knowledge graph. In *SEKE*, pages 307–409, 2019.

[7] Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *WWW*, pages 151–161, 2019.

[8] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *ACM SIGKDD*, pages 353–362. ACM, 2016.

[9] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NeurIPS*, pages 2787–2795, 2013.

[10] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119, 2014.

[11] Theo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *ICML*, pages 2071–2080, 2016.

[12] Zhiqing Sun, Zhihong Deng, Jianyun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *ICLR*, pages 1–18, 2019.

[13] Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In *NeurIPS*, pages 2731–2741, 2019.

[14] Xing Tang, Ling Chen, Jun Cui, and Baogang Wei. Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. *Inf. Process. Manag.*, 56(3):809–822, 2019.

[15] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In *IJCNN*, pages 1–8, 2019.

[16] Ruobing Xie, Stefan Heinrich, Zhiyuan Liu, Cornelius Weber, Yuan Yao, Stefan Wermter, and Maosong Sun. Integrating image-based and knowledge-based representation learning. *IEEE Trans. Cogn. Dev. Syst.*, 12(2):169–178, 2020.

[17] Yingying Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. Multi-modal multi-relational feature aggregation network for medical knowledge representation learning. In *ACM MM*, pages 3956–3965, 2020.

[18] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, pages 2659–2665, 2016.

[19] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. A multimodal translation-based approach for knowledge graph representation learning. In **SEM@NAACL-HLT*, pages 225–234, 2018.

[20] Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. Mmkrl: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence*, pages 1–18, 2021.

[21] Cédric Villani. *Optimal transport: old and new*, volume 338. 2009.

[22] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[23] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187, 2015.

[24] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *ACL*, pages 687–696, 2015.

[25] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, pages 809–816, 2011.

[26] Bishan Yang, Wentau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*, pages 1–13, 2015.

[27] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *AAAI*, pages 1955–1961, 2016.

[28] Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *ACL*, pages 2713–2722, 2020.

[29] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *ACL*, 2020.

[30] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, pages 593–607. Springer, 2018.

[31] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *AAAI*, pages 1811–1818, 2017.

[32] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *NAACL-HLT*, pages 327–333, 2018.

[33] Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In *ACL*, pages 4710–4723, 2019.

[34] Trapit Bansal, Da-Cheng Juan, Sujith Ravi, and Andrew McCallum. A2n: attending to neighbors for knowledge graph inference. In *ACL*, pages 4387–4392, 2019.

[35] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *NeurIPS*, 2020.

[36] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM J. Imaging Sci.*, 7(3):1853–1882, 2014.

[37] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML*, pages 274–289, 2014.

[38] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *ICML*, pages 685–693, 2014.

[39] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, pages 3730–3739, 2017.

[40] Ziyu Liu, Wei Shao, Jie Zhang, Min Zhang, and Kun Huang. Transfer learning via optimal transportation for integrative cancer patient stratification. In *IJCAI*, pages 2760–2766, 2021.

[41] Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *NeurIPS*, 2020.

[42] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *NeurIPS*, 2021.

[43] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020.

[44] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[45] Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *COLT*, pages 403–414, 2008.

[46] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146, 2017.

[47] Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. Yago3: A knowledge base from multilingual wikipedias. 2013.

[48] Ivana Balažević, Carl Allen, and Timothy M Hospedales. Tucker: Tensor factorization for knowledge graph completion. In *EMNLP*, pages 5185–5194, 2019.

[49] Balažević, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. In *NeurIPS*, pages 4465–4475, 2019.

[50] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655, 2014.

[51] Bahare Fatemi, Perouz Taslakian, David Vázquez, and David Poole. Knowledge hypergraphs: Prediction beyond binary relations. In *IJCAI*, pages 2191–2197, 2020.

[52] Koby Crammer, Michael J. Kearns, and Jennifer Wortman. Learning from multiple sources. *J. Mach. Learn. Res.*, 9:1757–1774, 2008.

[53] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR*, 2015.

[54] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[56] Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250, 2008.

[57] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[58] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *ACL*, pages 1532–1543, 2014.

[59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[60] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The experimental results are relatively stable and not sensitive to random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] Assets we use in the experiments are open source.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Assets we use in the experiments are open source.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]