NASH-GUIDED COMPATIBLE OPTIMAL TRANSPORT FOR MULTIMODAL KNOWLEDGE GRAPH FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Knowledge Graphs (MMKGs) integrate heterogeneous modalities to enhance knowledge representation. Yet, a fundamental challenge lies in multimodal representation fusion, where modality inputs often contain substantial target-irrelevant semantics (*i.e.*, low-target relevance), and different modalities contribute unevenly during fusion (*i.e.*, modality imbalance). Existing approaches typically align all modality representations indiscriminately and rely on predefined fusion strategies, making them sensitive to noisy or dominant modalities. To address these limitations, this paper proposes a **Na**sh-guided **C**ompatible **O**ptimal **T**ransport algorithm (NaCOT), which formulates multimodal fusion as an Optimal Transport (OT) problem. Specifically, NaCOTfirst constructs cascaded representation candidates and then employs sliced Wasserstein distance (SWD)-based selection to identify compatible source–target scales for OT alignment. Importantly, NaCOTintroduces a Nash bargaining-based fusion mechanism that refines transport plans under fusion and OT constraints, jointly mitigating low-target relevance and modality imbalance. Experiments on four MMKG benchmarks show that NaCOTachieves state-of-the-art performance, with average relative improvements of 24.68% in MRR and 23.51% in Hit@1 over the strongest baselines.

1 INTRODUCTION

Multi-modal Knowledge Graphs (MMKGs) (Liu et al., 2019) enrich traditional knowledge graphs by integrating structural relations with heterogeneous modalities, such as textual descriptions and visual content. This integration enables MMKGs to capture both relational and modality-specific semantics, leading to more informative and expressive entity representations. As a result, multimodal representation fusion has become a fundamental problem in MMKGs, with broad implications for downstream tasks such as question answering (Liu et al., 2022), recommendation (Meng et al., 2025), and information retrieval (Wang et al., 2025).

Existing approaches can be broadly grouped into two major lines. The first line adopts *structure-guided fusion*, which exploits the relational topology of the knowledge graph to regulate multimodal interactions. Representative approaches include methods that incorporate graph structural constraints (Zhang et al., 2025a), relational message passing (Cheng et al., 2025), or topology-aware regularization (Lee et al., 2023), *etc*. The second line relies on *feature-level fusion*, which leverages attention mechanisms (Chen et al., 2023; Li et al., 2023a;b), contrastive learning (Lu et al., 2024; Liu et al., 2024), or co-training strategies (Radford et al., 2021) to align heterogeneous modality representations in a shared embedding space. However, modality-specific embeddings often reside in heterogeneous representation spaces, and direct fusion may distort their modality-specific distributions (Liang et al., 2022; Hemker et al., 2024). Accordingly, Optimal Transport (OT)-based fusion has emerged as a promising direction (Cao et al., 2022; Zhang et al., 2024d; Tang et al., 2023). By formulating cross-modal alignment as distribution matching, OT softly transports modality representations into a shared space while preserving their distributional structures.

Nevertheless, existing OT-based methods often overlook two critical limitations in fusion: *low target relevance* and *modality imbalance*. The former refers to the presence of target-irrelevant semantics and modality-specific noise in textual or visual inputs, while the latter denotes the uneven contribution of different modalities during fusion. These limitations can lead to unreliable cross-modal

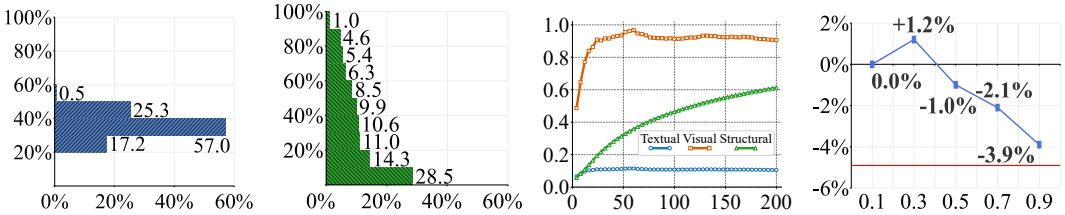
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping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

transport and biased fused representations, thereby degrading downstream performance. A detailed analysis is provided in Section 2 to illustrate their impact on MMKG representation fusion.

To address these two limitations, this paper introduces NaCOT, a **Na**sh-guided **C**ompatible **O**ptimal **T**ransport algorithm. NaCOT first constructs cascaded candidates to represent entity-specific evidence from multi-scale multimodal inputs. It then uses sliced Wasserstein distance to efficiently identify compatible transport pairs, avoiding indiscriminate cross-modal alignment. Importantly, NaCOT introduces a Nash bargaining-based adaptive fusion mechanism to refine the transport matrix under both fusion and OT constraints, jointly mitigating low target relevance and modality imbalance during fusion. In summary, our main contributions are[1]:

- We propose NaCOT, which, *to the best of our knowledge*, is the first Nash bargaining-guided compatible OT fusion framework to jointly address low-target relevance and modality imbalance in MMKGs.
- We design a cascaded representation construction strategy with sliced Wasserstein distance-based selection to represent entity-specific evidence and identify compatible cross-modal transport pairs efficiently.
- We introduce a Nash bargaining-based adaptive mechanism to refine the transport matrix under both fusion and OT constraints, with correctness justified by theoretical analysis.
- We conduct extensive experiments on four competitive MMKG benchmarks, showing that NaCOT achieves state-of-the-art performance with average relative improvements of 24.68% in MRR and 23.51% in Hit@1.

2 Revisiting OT-based MMKG Fusion



(a) Target-related mentions (b) Target-region ratio (c) Wasserstein distance $\mathcal{W}$ (d) Impact of λ

Figure 1: Empirical analysis using MKG-W. (a)–(b) Target relevance in textual and visual modalities, measured by target-related mentions and target-entity regions, respectively. The x-axis denotes the proportion of samples, while the y-axis denotes the proportion of target-relevant semantics. (c) Modality imbalance measured by Wasserstein distance $\mathcal{W}$ between modality-specific and fused representation. The x- and y-axes denote training epochs and $\mathcal{W}$, respectively. (d) MRR sensitivity to λ, with OTKGE (black line) and TransE (red line, single-modality) as references.

OT-based fusion formulates multi-modal representation alignment as a distribution matching problem, with *the core idea of learning a transport plan* that moves different modality embeddings into a unified space (Cao et al., 2022; Gong et al., 2025). Formally, let $E_{(m)} \in \mathbb{R}^{n\times d}$ denote the embedding matrix of modality m, where $m \in \mathcal{M}$ and $\mathcal{M}$ denotes the set of available modalities. OT-based fusion computes a transport-plan matrix T to produce the target-aligned embedding $E_{(m)}$:

$$\hat{E}_{(m)} = \mathrm{diag}(1/\mathbf{o})(T^\top + \Delta_T)E_{(m)}, \tag{1}$$

where $\mathbf{o}$ denotes the target marginal distribution, and Δ_T is a learnable weight. The final fusion then seeks a unified representation E_F that minimizes the weighted Wasserstein distance $\mathcal{W}(\cdot,\cdot)$ between E_F and each aligned modality:

$$E_F = \arg\min_{E} \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \lambda_i \mathcal{W}(\hat{E}_{(i)}, E), \tag{2}$$

where λ_i is a manually specified modality weight.

[1]Source codes are available on `https://anonymous.4open.science/r/NaCOT-66B7`

Despite their strong empirical performance, existing OT-based fusion methods leave two issues insufficiently addressed: **low-target relevance** and **modality imbalance**. Low-target relevance arises when modality inputs contain substantial off-target semantics (Zhang et al., 2025d; Cai et al., 2025). Since existing OT-based fusion lacks explicit mechanisms to filter unreliable inputs, noisy semantics can be transported into the shared representation space, not to mention the computation cost. In addition, modality imbalance occurs when one modality dominates the fusion process and suppresses the contribution of others (Hu et al., 2025a; Wei et al., 2025; Fan et al., 2023). Existing fusion methods typically rely on manually specified λ, making them sensitive to hyperparameter selection and modality dominance.

To better characterize these two issues, a preliminary analysis is performed via applying OTKGE (Cao et al., 2022) on the MKG-W benchmark (Xu et al., 2022), where each entity is associated with structural, textual, and visual modalities. Specifically, low-target relevance is measured by the proportion of entity-relevant semantics in textual and visual inputs, while modality imbalance is studied through the correlation between fused and modality-specific representations and the sensitivity to modality-weighting hyperparameters[2]. As shown in Fig. 1(a)–(b), 74.2% of textual samples contain less than 40% target-related mentions while 42.8% of visual samples contain less than 20% target-entity regions, indicating clear low target relevance in MMKG inputs. Fig. 1(c) further shows that the fused representation is much closer to the textual modality than to others, suggesting modality bias. Fig. 1(d) further confirms the model sensitivity to λ, where $\lambda = 0.3$ improves MRR by 1.2%, while $\lambda = 0.9$ reduces MRR by 3.9%.

We further argue that these two issues are closely coupled. That is, a more target-relevant modality can trigger the model bias toward modality-specific shortcuts, which can improve immediate alignment but also amplify modality imbalance by suppressing weaker modalities. Conversely, low-relevance modalities introduce noisy signals into fusion, further destabilizing the balance among modalities. Therefore, *low-target relevance and modality imbalance should be addressed jointly rather than treated as independent problems*. Accordingly, the following section introduces NaCOT that explicitly alleviates both two issues during multimodal representation fusion.

3 METHODOLOGY

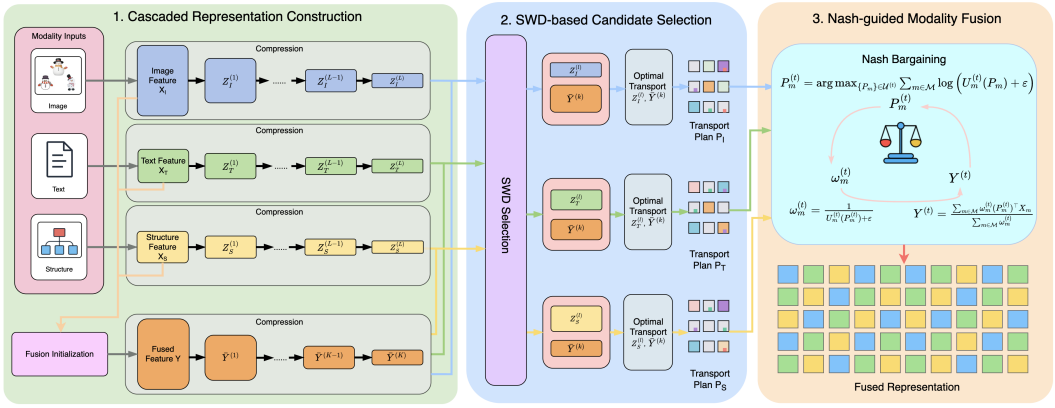


Figure 2: Illustration of NaCOT. The proposed method first progressively constructs cascaded representation candidates, selects compatible cross-modal transport pairs, and then adaptively coordinates modality fusion through Nash-guided transport refinement.

This section introduces NaCOT, a **Na**sh-guided **C**ompatible **O**ptimal **T**ransport algorithm. As shown in Fig. 2, NaCOT mainly consists of three components: Cascaded Representation Construction (CRC), SWD-based Candidate Selection (SCS), and Nash-guided Modality Fusion (NMF).

[2]Detailed settings are provided in Appendix B (for low-target relevance) and C (for modality imbalance).

3.1 Problem Formulation

Let $\mathcal{E} = \{e_i\}_{i=1}^{N}$ denote the set of entities, and $\mathcal{M}$ denote the set of available modalities. For each entity e_i and modality $m \in \mathcal{M}$, there may exist multiple modality-specific samples, such as several images or multiple textual mentions associated with the same entity. Let $\{\boldsymbol{e}_{i,m}^{(s)}\}_{s=1}^{S_{i,m}}$ denote their raw feature representations, where $S_{i,m}$ is the sample number of e_i with modality m. These features can be extracted by pre-trained encoders such as VGG (Simonyan & Zisserman, 2015) and BERT (Devlin et al., 2019) for visual and textual modalities, respectively. The goal of MMKG representation fusion is to learn a mapping function $\phi : \{\{\boldsymbol{e}_{i,m}^{(s)}\}_{s=1}^{S_{i,m}}\}_{m\in\mathcal{M}} \rightarrow \boldsymbol{y}_i$, where $\boldsymbol{y}_i$ denotes the unified representation of e_i.

3.2 Cascaded Representation Construction

Since OT alignment is highly sensitive to the representations being transported, directly using raw modality features can be unreliable when modality samples contain only partial entity-specific information. To identify suitable inputs for OT, the Cascaded Representation Construction (CRC) component generates representation candidates at different scales, which serve as candidate inputs for subsequent compatibility-aware OT alignment.

Specifically, let $\widetilde{X}_m \in \mathbb{R}^{n_m \times d_m}$ denote the original feature matrix of modality m, where each row corresponds to one sample representation $\boldsymbol{e}_{i,m}^{(s)}$ with $s \in [1, S_{i,m}]$, and $n_m = \sum_i S_{i,m}$. Let $Y \in \mathbb{R}^{N\times d}$ denote the fused entity representation matrix, whose i-th row is $\boldsymbol{y}_i$. To make modality features comparable with Y, each modality is projected into the shared d-dimensional space:

$$X_m = \widetilde{X}_m W_m \in \mathbb{R}^{n_m \times d}, \tag{3}$$

where $W_m \in \mathbb{R}^{d_m \times d}$ is a trainable projection matrix. Subsequently, CRC constructs representation candidates through a cascaded transformation chain:

$$X_m = Z_m^{(0)} \rightarrow Z_m^{(1)} \rightarrow \cdots \rightarrow Z_m^{(L)}, \tag{4}$$

where L denotes the number of construction layers. Each layer derives a representation candidate from its immediate predecessor through a trainable transformation:

$$Z_m^{(\ell+1)} = A_m^{(\ell)} Z_m^{(\ell)}, \tag{5}$$

where $A_m^{(\ell)}$ denotes a trainable matrix at layer ℓ. The dimensionality is designed to be gradually reduced to encourage compact candidate representations that preserve shared, target-consistent semantics while suppressing weakly relevant or sample-specific noise. As such, a compression budget is introduced via $\eta_Z \in [0, 1]$, where $\frac{\dim(Z_m^{(\ell+1)})}{\dim(Z_m^{(\ell)})} = \eta_Z$.

Similarly, a K-step cascaded factorization process is applied to Y to construct multi-scale target representations:

$$\overline{Y}^{(0)} = Y, \quad \overline{Y}^{(k-1)} = B^{(k)}\overline{Y}^{(k)}, \quad k \in [1, K]. \tag{6}$$

At each step, a rank-r_k truncated SVD is performed on $\overline{Y}^{(k-1)}$, where $B^{(k)} = U_{r_k}$ denotes the retained left singular basis and $\overline{Y}^{(k)} = \Sigma_{r_k} V_{r_k}^{\top}$ denotes the compact low-rank representation, and r_k is controlled via a compression budget η_Y, *i.e.*, $r_k = \lfloor \eta_Y \cdot \mathrm{rz}(\overline{Y}^{(k-1)}) \rfloor$, and $\mathrm{rz}(\cdot)$ is the row size.

As a result, for each modality, a *source*-side candidate set $\mathcal{Z}_m$ from compressed modality representations and a *target*-side candidate set $\mathcal{Y}$ from compressed fused representations are constructed, respectively:

$$\begin{cases} \mathcal{Z}_m = \{Z_m^{(\ell)} \mid 1 \le \ell \le L\}, \\ \mathcal{Y} = \{\overline{Y}^{(k)} \mid 1 \le k \le K\}. \end{cases} \tag{7}$$

This cascaded multi-scale construction allows the subsequent OT-based alignment to compare modality samples at different representation granularities, aiming for reducing the influence of sample-specific noise and/or weakly-relevant semantics[3].

[3]This study focuses on sample-level candidate construction and does not explicitly perform feature-level refinement. A possible extension is to introduce feature-side transformations, such as applying a learnable

3.3 SWD-BASED CANDIDATE SELECTION

The proposed CRC provides multiple source- ($\mathcal{Z}_m$) and target-side ($\mathcal{Y}$) representation candidates for OT-based fusion. However, not all scales are suitable for OT alignment, and exhaustive OT over all combinations is also computationally expensive. To address this, sliced Wasserstein distance (SWD)-based Candidate Selection (SCS) is introduced as a lightweight pre-selection step, where SWD (Rabin et al., 2011) estimates distributional compatibility:

$$(Z_m^*, Y_m^*) = \arg \min_{\substack{Z_m^{(\ell)} \in \mathcal{Z}_m \\ \overline{Y}^{(k)} \in \mathcal{Y}}} \mathrm{SWD}\Big(Z_m^{(\ell)}, \overline{Y}^{(k)}\Big). \tag{8}$$

Notably, the selected Y_m^* is modality-specific, since different modalities can be aligned with different target scales. Given selected (Z_m^*, Y_m^*), their OT cost matrix can be estimated as $C_m[i,j] = c_m(Z_m^*[i], Y_m^*[j])$, where $c_m(\cdot,\cdot)$ denotes a distance metric. Accordingly, the transport matrix $\hat{P}_m$ for modality m can be computed using the traditional entropic Sinkhorn algorithm:

$$\hat{P}_m = \arg \min_{P_m \in \hat{\mathcal{P}}_m} \langle P_m, C_m \rangle + \tau \Omega(P_m), \tag{9}$$

where $\Omega(P_m)$ is the entropic regularizer with $\tau > 0$, $\hat{\mathcal{P}}_m := \{P | P\mathbf{1} = \mathbf{p}_m,\ P^\top \mathbf{1} = \mathbf{q}_m,\ P \geq 0\}$, and $\mathbf{p}_m$ and $\mathbf{q}_m$ are uniform marginal distributions. Furthermore, applying the barycentric projection, Z_m^* can be transported to the target representation as:

$$\widehat{Y}_m^* = \mathrm{diag}(\mathbf{q}_m)^{-1} \hat{P}_m^\top Z_m^* = \mathrm{diag}(\mathbf{q}_m)^{-1} \hat{P}_m^\top A_m^{(l)^*} \cdots A_m^{(0)^*} X_m \tag{10}$$

where $A_m^{(\cdot)^*}$ denote the cascaded refinement matrices corresponding to Z_m^*.

Since Y_m^* may lie in a compressed target scale, it can be further lifted back to the final fused representation space through the corresponding reconstruction matrices $B_m^{(\cdot)^*}$, *i.e.*, $Y_m = B_m^{(1)^*} \cdots B_m^{(k)*} \widehat{Y}_m^{k^*}$. To simplify notation we write

$$P_m^\top = B_m^{(1)^*} \cdots B_m^{(k)^*} \mathrm{diag}(\mathbf{q}_m)^{-1} \hat{P}_m^\top A_m^{(l)^*} \cdots A_m^{(0)^*} \tag{11}$$

and therefore obtain $Y_m = P_m^\top X_m$.

3.4 NASH-GUIDED MODALITY FUSION

Although P_m provides a target-aligned projection for each modality, it does not guarantee reliable fusion under varying target relevance. As discussed in Section 2, low-target relevance and modality imbalance are closely coupled during multimodal fusion. A modality with stronger target-relevant evidence may dominate the shared representation space, improving local alignment while reducing the contribution of other modalities. In contrast, modalities with lower target relevance may introduce noisy semantics into fusion, which further destabilizes the fused representation. This creates a trade-off in multimodal fusion: reliable evidence should be sufficiently exploited, while weaker but complementary modalities should also not be overwhelmed. To address this issue, modality fusion is formulated as a cooperative Nash bargaining game, where *each modality m acts as a player, the transport plan serves as its strategy, and the utility measures its fusion benefit*. This formulation encourages modalities to reach a proportionally fair balance, thereby mitigating modality dominance and improving fusion robustness under low-target relevance.

Specifically, for each modality $m \in \mathcal{M}$, we have representation X_m and a target-aligned transport plan $P_m^{(0)} = P_m$, obtained using the traditional OT as Eq. equation 11. The fused representation is thereby initialized as a weighted aggregation:

$$Y^{(0)} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} (P_m^{(0)})^\top X_m.$$

Hereafter, we use $Y^{(t)}$ to denote the fused representation that is iteratively updated by solving the Nash bargaining problem (NBP).

projection $H_m^{(\ell)}$ to obtain $Z_m^{(\ell)} H_m^{(\ell)}$, or imposing sparse self-expression on the fused representation, *i.e.*, $Y = YP$ with an ℓ_1-regularization on P. We leave this direction for future work.

Utility. The expense of transport plan P_m is defined as

$$\exp_m(P_m, Y) = \frac{1}{2}\left\|P_m^\top X_m - Y\right\|_F^2 + \gamma R(P_m), \tag{12}$$

where Y is an auxiliary variable denoting one fused-representation input, $R(P_m) = \langle P_m, C_m\rangle + \tau\Omega(P_m)$ preserves the original OT transport cost and entropic regularization, and $\gamma > 0$ is a hyper-parameter. The first term measures the reconstruction error between the transport-aligned modality-specific representation and the final fused representation, thereby quantifying the contribution of each modality to the fused representation after OT alignment. Additionally, the second term regularizes the transport plan to remain consistent with the original OT objective. At iteration t, the utility of P_m is accordingly defined as the reduction in expense relative to the previous plan:

$$U_m^{(t)}(P_m) = \exp_m(P_m^{(t-1)}, Y^{(t-1)}) - \exp_m(P_m, Y^{(t-1)}). \tag{13}$$

Nash bargaining update. The transport plans are then updated by maximizing the Nash product that is the equilibrium of the Nash Bargaining Problem:

$$P_m^{(t)} = \arg\max_{\{P_m\}\in\mathcal{U}^{(t)}} \sum_{m\in\mathcal{M}} \log\Big(U_m^{(t)}(P_m) + \varepsilon\Big), \tag{14}$$

where $\varepsilon > 0$ is used for numerical stability. The feasible set or the so-called strategy space is defined as

$$\mathcal{U}^{(t)} = \Big\{P_m \;\Big|\; \frac{\sum_{m\in\mathcal{M}} \omega_m^{(t-1)} P_m^\top X_m}{\sum_{m\in\mathcal{M}} \omega_m^{(t-1)}} = Y^{(t-1)}, P_m \in \mathcal{P}_m, \forall m\Big\}, \tag{15}$$

where the bargaining weight for modality m is computed by

$$\omega_m^{(t-1)} = \frac{1}{U_m^{(t-1)}(P_m^{(t-1)}) + \varepsilon}. \tag{16}$$

This inverse-utility weighting assigns larger corrective emphasis to modalities with lower previous utility, preventing dominant modalities from overwhelming weaker but complementary ones. Finally, the fused representation $Y^{(t)}$ can be simply estimated as:

$$Y^{(t)} = \frac{\sum_{m\in\mathcal{M}} \omega_m^{(t)} (P_m^{(t)})^\top X_m}{\sum_{m\in\mathcal{M}} \omega_m^{(t)}}. \tag{17}$$

After T iterations, the resulting representation $Y^{(T)}$ is used as the final fused output. Theoretical analysis of the proposed Nash equilibrium formulation and optimization procedure is provided in Appendix E.

4 Experiments

4.1 Experimental Setup

Datasets. Evaluation is conducted on four widely-used MMKG benchmarks, including MKG-W (Xu et al., 2022), MKG-Y (Xu et al., 2022), DB15K (Liu et al., 2019), and KVC16K (Zhang et al., 2024a). Their detailed statistics are provided in Table 3 (see Appendix G.1).

Baselines. NaCOT is compared against four baseline categories. *Uni-modal Methods* include TransE (Bordes et al., 2013) and ComplEx (Trouillon et al., 2016). *Non-OT Methods* include IMF (Li et al., 2023b), HFR-MKGC (Wang et al., 2026), AdaMF (Zhang et al., 2024c), MOMOK (Zhang et al., 2025b), and MyGO (Zhang et al., 2025c). *LLM-based Methods* include SAT (Liu et al., 2025b), KoPA (Zhang et al., 2024b), and LBMKGC (Guo et al., 2025). *OT-based Methods* include OTKGE (Cao et al., 2022), OTMEL (Zhang et al., 2024d), SLOTAlign (Tang et al., 2023), and DecAlign (Qian et al., 2026). More baseline details are provided in Appendix A.

Implementation details. All experiments are conducted on a single NVIDIA H100 GPU. NaCOT employs VGG (Simonyan & Zisserman, 2015) and BERT (Devlin et al., 2019) as default visual and

Table 1: Performance comparison of various methods across four MMKGC benchmarks. The **best** results are bold and the second best results are underlined. Green superscripts indicate absolute gains over the second best results, while the relative gain row reports percentage improvements over them. The symbol * denotes reproduced results, while all others are directly reported from the original papers.

Model	MKG-W			MKG-Y			DB15K			KVC16K		
	MRR↑	Hit@1↑	Hit@3↑	MRR↑	Hit@1↑	Hit@3↑	MRR↑	Hit@1↑	Hit@3↑	MRR↑	Hit@1↑	Hit@3↑
TransE	29.19	21.06	33.20	30.73	23.45	35.81	24.86	12.78	31.48	8.54	0.64	10.97
ComplEx	24.93	19.09	26.69	28.71	22.26	32.12	27.48	18.37	31.57	12.85	7.48	13.79
IMF	34.50	28.77	36.62	35.79	32.95	37.14	32.25	24.20	36.00	12.01	7.42	12.82
HFR-MKGC	38.62	31.37	42.12	39.00	34.41	42.12	35.20	23.04	43.25	13.16	7.24	11.96
AdaMF	34.27	27.21	37.86	38.06	33.49	40.44	32.51	21.31	39.67	15.26	8.56	16.71
MOMOK	35.89	30.38	37.54	37.91	35.09	39.20	39.57	32.38	43.45	16.87	10.53	18.26
MyGO	36.09	30.02	38.26	38.66	34.85	39.96	37.17	29.62	41.66	16.78	10.45	17.92
SAT*	34.61	28.96	36.75	35.91	33.00	37.32	33.00	24.56	36.85	11.78	7.42	12.62
KoPA*	32.57	25.76	34.89	33.59	31.15	35.97	30.35	22.10	34.17	9.80	2.51	10.91
LBMKGC	38.46	31.20	41.78	40.03	33.89	43.11	37.23	27.78	42.75	16.15	9.34	17.90
OTKGE	34.36	28.85	36.25	35.51	31.97	37.18	23.86	18.45	25.89	8.77	5.01	9.31
OTMEL*	34.38	28.79	36.80	34.17	30.98	36.81	38.54	30.83	40.68	16.53	10.31	17.61
SLOTAlign*	33.78	27.95	37.89	34.26	31.24	39.28	26.29	18.45	30.11	13.45	7.97	14.94
DecAlign*	34.54	29.19	37.32	34.89	30.94	36.24	35.70	31.33	39.54	16.17	10.00	17.10
NaCOT	**41.49**$^{\uparrow 2.87}$	**32.16**$^{\uparrow 0.79}$	**45.07**$^{\uparrow 2.95}$	**45.40**$^{\uparrow 5.37}$	**41.59**$^{\uparrow 6.50}$	**46.64**$^{\uparrow 3.53}$	**45.33**$^{\uparrow 5.76}$	**35.10**$^{\uparrow 2.72}$	**50.46**$^{\uparrow 7.01}$	**27.55**$^{\uparrow 10.68}$	**17.33**$^{\uparrow 6.80}$	**30.38**$^{\uparrow 12.12}$
Rel. Gain (%)	+7.43%	+2.52%	+7.00%	+13.41%	+18.52%	+8.19%	+14.56%	+8.40%	+16.13%	+63.31%	+64.58%	+66.37%
± Std.	±0.10	±0.86	±0.82	±0.09	±0.84	±0.89	±0.13	±0.92	±0.73	±0.12	±0.88	±0.77

textual encoders for a fair comparison with existing work. We set the batch size to 2000, embedding dimension to 300, learning rate (from AdamW) to 1×10^{-3}, $\eta_Z = \eta_Y = 0.8$, $L = K = 5$, $T = 3$, and $\tau = \gamma = 0.1$. Hyperparameters are selected by grid search, with sensitivity analyses presented in Appendix G.3. *Baseline results are taken from original papers when available, or reproduced using official code*. Evaluation uses MRR and Hit@K ($K = 1, 3$) for MMKG link prediction.

4.2 Main Results

Table 1 reports the overall results on four MMKG benchmarks over five runs. NaCOT consistently achieves the best performance across all datasets and metrics, with average relative improvements of 24.68%, 23.51%, and 24.42% in MRR, Hit@1, and Hit@3 over the second-best baselines, respectively. The gains are particularly notable on `KVC16K`, which contains larger numbers of entities and relations, suggesting that NaCOT is effective in large-scale scenarios. Meanwhile, the larger entity space may also introduce more target-irrelevant semantics, making fusion more challenging. By constructing representation candidates, selecting compatible transport pairs, and dynamically coordinating modality contributions, NaCOT suppresses noisy modality input and maintains stable alignment. We further conduct significance tests using paired t-tests against the corresponding strongest baselines on four benchmarks, yielding $p < 1 \times 10^{-5}$ in all cases. These results confirm the statistical significance and robustness of NaCOT.

4.3 Ablation Study

To analyze the effectiveness of NaCOT, we conduct ablation studies on `MKG-W` and `KVC16K`, which represent noisy multimodal input and large-scale MMKGC scenarios, respectively.

Model Breakdown. NaCOT comprises components of Cascaded Representation Construction (CRC), SWD-based Candidate Selection (SCS), and Nash-guided Modality Fusion (NMF). Their contributions are evaluated by incrementally adding each to the standard OT (*i.e.*, `Base`), leading to the following variants: (i) `+CRC` uses one cascade step ($L = K = 1$) with $\eta_Z = \eta_Y = 0.8$, without SCS and NMF; (ii) `+SCS` identifies the compatible candidate using $\eta_Z = \eta_Y = 0.8$ and $L = K = 5$, without further Nash fusion; (iii) `+NMF` applies Nash-guided fusion to `Base` without CRC and SCS.

As shown in Fig. 3, each component consistently improves performance, indicating that CRC, SCS, and NMF provide complementary benefits for fusion. Among them, `+CRC` achieves performance close to `Base`, suggesting that candidate construction alone is insufficient without compatibility selection and adaptive fusion. In contrast, `+NMF` yields larger average gains, especially on `KVC16K`, highlighting the Nash-guided coordination at scale. On

MKG-W, +SCS improves more than +NMF, which may be attributed to the fact that reliable transport inputs are more critical for the subsequent fusion under low-target relevance.

Impact of CRC. This section evaluates the impact of the compression budgets η_Z and η_Y in the CRC component. Specifically, both η_Z and η_Y are jointly varied from 0.2 to 1.0 with an interval of 0.2, while all other components remain unchanged.

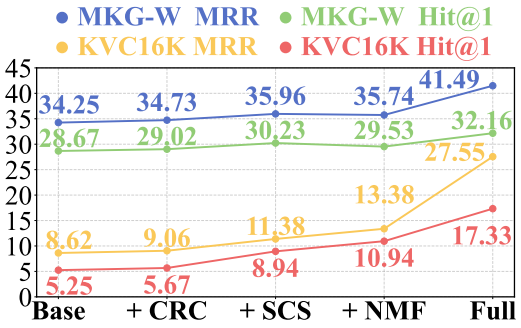
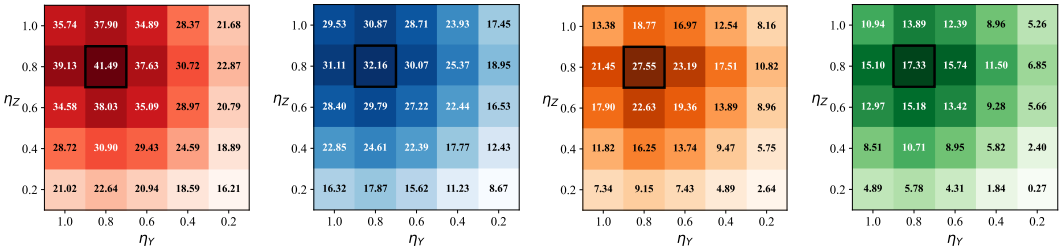


Figure 3: Performance gains from individual components in NaCOT.

As shown in Fig. 4, the best performance is generally achieved when η_Z and η_Y fall within the range of 0.6–0.8, with the peak at $\eta_Z = \eta_Y = 0.8$. Overly strong compression ($\eta_{Z/Y} \leq 0.4$) reduces MRR and Hit@1 due to information loss, while insufficient compression ($\eta_{Z/Y} \rightarrow 1.0$) fails to suppress noisy or weakly relevant modality semantics. Notably, CRC does not rely on a single compression level; instead, it provides multi-level representation candidates, from which SCS selects the compatible pair with the shortest sliced Wasserstein distance. These results support the CRC design, showing that appropriate compression ranges provide reliable candidate representations for the subsequent OT-based fusion.



(a) MKG-W (MRR) (b) MKG-W (Hit@1) (c) KVC16K (MRR) (d) KVC16K (Hit@1)

Figure 4: Sensitivity analysis of compression budgets η_Z and η_Y. Darker colors indicate better performance.

Alternative Candidate Selection. SWD is used in NaCOT to estimate distributional compatibility for candidate selection. To evaluate its effectiveness, SWD is replaced with the following alternatives while keeping all other components unchanged: (i) Random samples transport candidates uniformly as a baseline; (ii) Greedy enumerates all candidate pairs and selects the one with the lowest transport cost; (iii) MMD selects candidates based on Maximum Mean Discrepancy; and (iv) Chamfer selects candidates using Chamfer distance (Barrow et al., 1977; Fan et al., 2017). For all settings, we set $\eta_Z = \eta_Y = 0.8$ and $L = K = 5$.

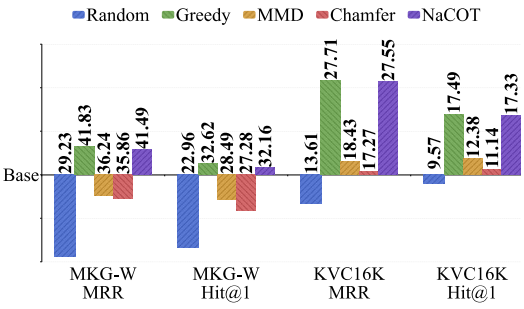
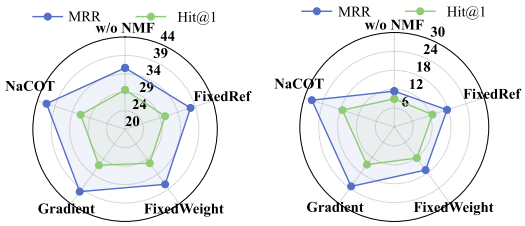


Figure 5: Effect of different candidate selection strategies on model performance, where the *Base* is set to HFR-MKGC on MKG-W and MOMOK on KVC16K.

As illustrated in Fig. 5, Random performs the worst and serves as the lower-bound reference. Greedy achieves performance comparable to SWD, suggesting that SWD provides a reliable and more efficient surrogate for candidate selection. Notably, our goal is not to solve an exact transport matching problem, but to identify compatible candidates that benefit downstream MMKG

fusion. Additionally, the MMD and Chamfer variants generally underperform SWD, indicating that kernel mean matching and nearest-neighbour set matching are less reliable than Wasserstein-based distributional comparison in terms of cross-domain transport selection. Nevertheless, they still remain competitive in some cases; for instance, on KVC16K, they still outperform the strongest baseline.

Effect of Nash Bargaining. Nash-guided Modality Fusion (NMF) is designed to adaptively coordinate modality contributions during fusion. To evaluate its effectiveness, we compare NaCOT with four variants: (i) w/o NMF removes NMF and directly fuses modalities after SCS with the standard OTKGE strategy; (ii) FixedRef fixes the reference term $\exp_m(P_m^{(t-1)}, Y^{(t-1)})$ in Eq. equation 13 to its initial value $\exp_m(P_m^{(0)}, Y^{(0)})$; (iii) FixedWeight replaces adaptive bargaining weights in Eq. equation 16 by setting $\omega_m^{(t)} = 1/|\mathcal{M}|$ for all modalities and iterations; and (iv) Gradient treats multimodal fusion as a multi-task optimization problem and adopts gradient-based Nash optimization (Navon et al., 2022).



(a) MKG-W (b) KVC16K

Figure 6: Effect of different Nash-based variants on the final fusion performance.

As shown in Fig. 6, removing NMF leads to the worst performance, again indicating that adaptive Nash-guided coordination is essential for balancing modality contributions. FixedRef and FixedWeight also degrade performance, with MRR drops of 10.65 and 10.03 points on KVC16K, respectively. This shows that both dynamic utility references and adaptive bargaining weights are important for modeling the evolving fusion process. Finally, replacing NMF with gradient-based bargaining yields weaker results than NaCOT, underperforming by 1.41 MRR on MKG-W and 4.23 MRR on KVC16K. This suggests that transport-level Nash refinement is more suitable for MMKG fusion. Overall, these results demonstrate the necessity of NMF and confirm the benefits of dynamic utility estimation, adaptive modality weighting, and transport-level Nash-guided fusion.

Other Analyses. We further analyze NaCOT from multiple perspectives. First, complexity analysis in Appendix F shows that NaCOT achieves a balanced trade-off between computational cost and fusion performance compared with existing MMKGC baselines. Second, robustness experiments under modality noise, modality missing, and link sparsity demonstrate that NaCOT maintains stable performance under corrupted, incomplete, and sparse multimodal inputs (Appendix G.2). Third, hyperparameter sensitivity analysis on the regularization terms and fused representation dimension shows that NaCOT remains stable across different parameter settings (Appendix G.3). We also investigate modality contribution towards the proposed method by analyzing different combinations of structural, textual, and visual modalities, confirming that the proposed method benefits from complementary information (Appendix G.4). At last, experiments with different visual and textual backbone combinations further demonstrate the generalizability of NaCOT across diverse encoder architectures (Appendix G.5). Cross-task experiments further demonstrate the transferability of NaCOT to multimodal entity linking (Appendix G.6).

5 Conclusion

This paper introduces NaCOT, the *first* Nash-guided compatible Optimal Transport algorithm for MMKG fusion. NaCOT addresses low-target relevance and modality imbalance jointly by constructing cascaded representation candidates, selecting compatible scales, and refining transport plans through Nash-guided modality fusion. More importantly, it formulates modality coordination as a Nash bargaining process and provides theoretical justification for the resulting Nash equilibrium. Extensive experiments demonstrate its superiority over existing baselines. Future work may explore more expressive utility functions and fine-grained feature-level alignment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with semantic classification to conduct a modality correlation analysis. We don't use generative AI to do any other research work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the source code, dataset resources, and instructions for running the experiments in the anonymous repository at `https://anonymous.4open.science/r/NaCOT-66B7`. The main paper describes the proposed method, experimental setup, hyperparameters, evaluation protocol, and main results. The appendix provides the complete pseudocode in Appendix D, theoretical analysis and proofs in Appendix E, dataset statistics and additional experimental details in Appendix G, and further sensitivity and robustness analyses.

REFERENCES

H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'77, pp. 659–663, 1977. URL https://www.ijcai.org/Proceedings/77-2/Papers/024.pdf.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, and Prof Javen Qinfeng Shi. On the value of cross-modal misalignment in multimodal representation learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 145859–145911, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d6c06b4cab132aaade78d4d4d930b9c8-Paper-Conference.pdf.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Otkge: Multi-modal knowledge graph embeddings via optimal transport. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 39090–39102, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ffdb280e7c7b4c4af30e04daf5a84b98-Paper-Conference.pdf.

Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. Mmea: Entity alignment for multi-modal knowledge graph. In *Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020, Hangzhou, China, August 28–30, 2020, Proceedings, Part I*, pp. 134–147, 2020. ISBN 978-3-030-55129-2. doi: 10.1007/978-3-030-55130-8_12. URL https://doi.org/10.1007/978-3-030-55130-8_12.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z. Pan, Wenting Song, and Huajun Chen. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 3317–3327, 2023. ISBN 9798400701085. doi: 10.1145/3581783.3611786. URL https://doi.org/10.1145/3581783.3611786.

Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen, and Wen Zhang. Noise-powered multi-modal knowledge graph representation framework. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 141–155, 2025. URL https://aclanthology.org/2025.coling-main.11.pdf.

Jingwei Cheng, Mingxiao Guo, and Fu Zhang. SGMEA: Structure-guided multimodal entity alignment. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7851–7861, 2025. URL https://aclanthology.org/2025.coling-main.525/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, June 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2463–2471, July 2017. doi: 10.1109/CVPR.2017.264. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8099747.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, 2023. doi: 10.1109/CVPR52729.2023.01918. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10204495.

Min Gao, Haifeng Zheng, Xinxin Feng, and Ran Tao. Multimodal fusion using multi-view domains for data heterogeneity in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16736–16744, Apr. 2025. doi: 10.1609/aaai.v39i16.33839. URL https://ojs.aaai.org/index.php/AAAI/article/view/33839.

Fengjiiao Gong, Angxiao Yue, and Hongteng Xu. An optimal transport-based latent mixer for robust multi-modal learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (16):16826–16834, Apr. 2025. doi: 10.1609/aaai.v39i16.33849. URL https://ojs.aaai.org/index.php/AAAI/article/view/33849.

Yuan Guo, Qian Ma, Hui Li, Qiao Ning, Furui Zhan, Yu Gu, Ge Yu, and Shikai Guo. Lbmkgc: Large model-driven balanced multimodal knowledge graph completion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 111874–111896, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a254abbfdd029d388fc35fd850bbc551-Paper-Conference.pdf.

Kushagra Gupta, Surya Murthy, Mustafa Karabag, Ufuk Topcu, and David Fridovich-Keil. Cooperative bargaining games without utilities: Mediated solutions from direction oracles. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 3780–3805, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/05c1c6b0ecfed4c29e39adb226da0d5e-Paper-Conference.pdf.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: Multimodal fusion for heterogeneous biomedical data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 64479–64498, 2024. doi: 10.52202/079017-2057. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/765871e77d2ca65126d3d64d31aa6908-Paper-Conference.pdf.

Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. Adaptive multimodal fusion: dynamic attention allocation for intent recognition. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25, 2025a. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i16.33898. URL https://doi.org/10.1609/aaai.v39i16.33898.

Zhiwei Hu, Víctor Gutiérrez-Basulto, Ru Li, and Jeff Z. Pan. Multi-level matching network for multimodal entity linking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 508–519, 2025b. doi: 10.1145/3690624.3709306. URL https://dl.acm.org/doi/epdf/10.1145/3690624.3709306.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10378323.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. VISTA: Visual-textual knowledge graph representation learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7314–7328, 2023. doi: 10.18653/v1/2023.findings-emnlp.488. URL https://aclanthology.org/2023.findings-emnlp.488/.

Qian Li, Cheng Ji, Shu Guo, Zhaoji Liang, Lihong Wang, and Jianxin Li. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 987–999, December 2023a. doi: 10.18653/v1/2023.findings-emnlp.70. URL https://aclanthology.org/2023.findings-emnlp.70/.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. Imf: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 2572–2580, 2023b. ISBN 9781450394161. doi: 10.1145/3543507.3583554. URL https://doi.org/10.1145/3543507.3583554.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. Multi-modal contrastive representation learning for entity alignment. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2572–2584, 2022. URL https://aclanthology.org/2022.coling-1.227/.

Junping Liu, Shijie Mei, Xinrong Hu, Xun Yao, Jie Yang, and Yi Guo. Seeing the wood for the trees: a contrastive regularization method for the low-resource knowledge base question answering. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1085–1094, July 2022. doi: 10.18653/v1/2022.findings-naacl.82. URL https://aclanthology.org/2022.findings-naacl.82/.

Ruizhou Liu, Zongsheng Cao, Zhe Wu, Qianqian Xu, and Qingming Huang. Multimodal knowledge graph embeddings via lorentz-based contrastive learning. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024. doi: 10.1109/ICME57554.2024.10687608. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10687608.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 38–55, 2025a. ISBN 978-3-031-72970-6. doi: 10.1007/978-3-031-72970-6_3. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06319.pdf.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. Mmkg: Multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings*, pp. 459–474, 2019. ISBN 978-3-030-21347-3. doi: 10.1007/978-3-030-21348-0_30. URL https://doi.org/10.1007/978-3-030-21348-0_30.

Yu Liu, Yanan Cao, Xixun Lin, Yanmin Shang, Shi Wang, and Shirui Pan. Enhancing large language model for knowledge graph completion via structure-aware alignment-tuning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20970–20984, November 2025b. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1061. URL https://aclanthology.org/2025.emnlp-main.1061/.

Yuxing Lu, Weichen Zhao, Nan Sun, and Jinzhuo Wang. Enhancing multimodal knowledge graph representation learning through triple contrastive learning. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5963–5971, 8 2024. doi: 10.24963/ijcai.2024/659. URL https://doi.org/10.24963/ijcai.2024/659. Main Track.

Pablo Mendes, Max Jakob, and Christian Bizer. DBpedia: A multilingual cross-domain knowledge base. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1813–1817, May 2012. URL https://aclanthology.org/L12-1323/.

Fanshen Meng, Zhenhua Meng, Ru Jin, Rongheng Lin, and Budan Wu. Doge: Llms-enhanced hyper-knowledge graph recommender for multimodal recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12399–12407, Apr. 2025. doi: 10.1609/aaai.v39i12.33351. URL https://ojs.aaai.org/index.php/AAAI/article/view/33351.

John F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. URL https://doi.org/10.2307/1907266.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Dung V. Nguyen, Anh T. Nguyen, Minh H. Nguyen, Luc Q. Nguyen, Shiqi Jiang, Ethan Fetaya, Linh Duy Tran, Gal Chechik, and Tan M. Nguyen. Expert merging in sparse mixture of experts with nash bargaining. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ef9f16cd0cb8d7797ea35c0ee7964cdf-Paper-Conference.pdf.

Haojie Pan, Zepeng Zhai, Yuzhou Zhang, Ruiji Fu, Ming Liu, Yangqiu Song, Zhongyuan Wang, and Bing Qin. Kuaipedia: a large-scale multi-modal short-video encyclopedia, 2022. URL https://arxiv.org/abs/2211.00732.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f7f5f501282771c96bb3fedcc96bedfe-Paper-Conference.pdf.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Proceedings of the Third International Conference on Scale Space and Variational Methods in Computer Vision*, SSVM'11, pp. 435–446, 2011. ISBN 9783642247842. doi: 10.1007/978-3-642-24785-9_37. URL https://doi.org/10.1007/978-3-642-24785-9_37.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Bin Shang, Yinliang Zhao, Jun Liu, and Di Wang. LAFA: Multimodal knowledge graph completion with link aware fusion and aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8957–8965, 2024. doi: 10.1609/aaai.v38i8.28744. URL https://ojs.aaai.org/index.php/AAAI/article/view/28744.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 697–706, 2007. ISBN 9781595936547. doi: 10.1145/1242572.1242667. URL https://doi.org/10.1145/1242572.1242667.

Jianheng Tang, Weiqi Zhang, Jiajin Li, Kangfei Zhao, Fugee Tsung, and Jia Li. Robust attributed graph alignment via joint structure learning and optimal transport. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 1638–1651, 2023. doi: 10.1109/ICDE55515.2023.00129. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10184815.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2071–2080, 2016. URL https://proceedings.mlr.press/v48/trouillon16.html.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Di Wang, Junping Du, Zhe Xue, Meiyu Liang, Guanhua Ye, Yingxia Shao, and Haisheng Li. Hfr-mkgc: Hierarchical fusion reasoning with mllms for multi-modal knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15815–15823, Mar. 2026. doi: 10.1609/aaai.v40i18.38613. URL https://ojs.aaai.org/index.php/AAAI/article/view/38613.

Junyao Wang, Tongle Duan, Sen Yao, and Yasong Wang. Adamkg: Context-aware multimodal knowledge graph fusion for cross-domain retrieval, reasoning and analytics in enterprise information governance. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Digital Finance*, AIDF '25, pp. 239–254, 2025. ISBN 9798400714955. doi: 10.1145/3764727.3764767. URL https://doi.org/10.1145/3764727.3764767.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2250–2259, 2025. doi: 10.1109/ICCV51701.2025.00217. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Wei_Improving_Multimodal_Learning_via_Imbalanced_Learning_ICCV_2025_paper.pdf.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27328–27337, 2024. doi: 10.1109/CVPR52733.2024.02581. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Wei_Enhancing_Multimodal_Cooperation_via_Sample-level_Modality_Valuation_CVPR_2024_paper.pdf.

Jing Wu and Mehrtash Harandi. Munba: Machine unlearning via nash bargaining. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–12, 2025. doi: 10.1109/ICCV51701.2025.00452. URL https://openaccess.thecvf.com/content/ICCV2025/supplemental/Wu_MUNBa_Machine_Unlearning_ICCV_2025_supplemental.pdf.

Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 3857–3866, 2022. ISBN 9781450392037. doi: 10.1145/3503161.3548388. URL https://doi.org/10.1145/3503161.3548388.

Haochen You and Baojing Liu. Mover: Multimodal optimal transport with volume-based embedding regularization. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 5444–5448, 2025. ISBN 9798400720406. doi: 10.1145/3746252.3760958. URL https://doi.org/10.1145/3746252.3760958.

Yi Zeng, Xuelin Yang, Li Chen, Cristian Canton Ferrer, Ming Jin, Michael I. Jordan, and Ruoxi Jia. Fairness-aware meta-learning via nash bargaining. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83235–83267, 2024. doi: 10.52202/079017-2648. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/97a0f7d7fc2c26732943860241af1c7a-Paper-Conference.pdf.

Yan Zhang, Xiangyu Luo, Jing Hu, Miao Zhang, Kui Xiao, and Zhifei Li. Graph structure prefix injection transformer for multi-modal entity alignment. *Information Processing & Management*, 62(3):104048, 2025a. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2024.104048. URL https://www.sciencedirect.com/science/article/pii/S0306457324004072.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 91–101, 2024a. ISBN 9798400704314. doi: 10.1145/3626772.3657800. URL https://doi.org/10.1145/3626772.3657800.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Wen Zhang, and Huajun Chen. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 233–242, 2024b. ISBN 9798400706868. doi: 10.1145/3664647.3681327. URL https://doi.org/10.1145/3664647.3681327.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17120–17130, 2024c. URL https://aclanthology.org/2024.lrec-main.1487/.

Yichi Zhang, Zhuo Chen, Lingbing Guo, yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 54811–54828, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/88593e16b09104fb6010d370c081d7bc-Paper-Conference.pdf.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):13322–13330, Apr. 2025c. doi: 10.1609/aaai.v39i12.33454. URL https://ojs.aaai.org/index.php/AAAI/article/view/33454.

Zefeng Zhang, Jiawei Sheng, Chuang Zhang, Yunzhi Liang, Wenyuan Zhang, Siqi Wang, and Tingwen Liu. Optimal transport guided correlation assignment for multimodal entity linking. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4103–4117, 2024d. doi: 10.18653/v1/2024.findings-acl.243. URL https://aclanthology.org/2024.findings-acl.243/.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77233–77253, 13–19 Jul 2025d. URL https://proceedings.mlr.press/v267/zhang25dq.html.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. MoSE: Modality split and ensemble for multimodal knowledge graph completion. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10527–10536, 2022. doi: 10.18653/v1/2022.emnlp-main.719. URL https://aclanthology.org/2022.emnlp-main.719/.

APPENDIX CONTENTS

A Related Work **19**

A.1 Multi-Modal Knowledge Graph 19

A.2 Optimal Transport 19

A.3 Nash Bargaining 19

B Analysis on Low-Target Relevance **20**

B.1 Text Modality Analysis 20

B.2 Visual Modality Analysis 21

C Analysis on Modality Imbalance **22**

C.1 Representation Correlation Analysis 22

C.2 Hyperparameter Sensitivity Analysis 23

D Pseudocode for NaCOT **23**

E Theoretical Analysis **23**

F Computational Complexity **24**

G Additional Experiments **26**

G.1 Statistics of Datasets 26

G.2 Model Robustness 26

G.3 Hyperparameter Sensitivity Analysis 27

G.3.1 Sensitivity to Construction Depths and Nash Iterations 27

G.4 Modality Contribution 28

G.5 Impact of Different Backbones 28

G.6 Transferability to Multimodal Entity Linking 29

H Case Study **30**

H.1 Case Study on MKG-W 30

H.2 Improvement analysis on KVC16K 30

I Further Discussion **31**

I.1 Task-level Distinction 31

I.2 Methodological Distinction 32

J Limitations **32**

A Related Work

A.1 Multi-Modal Knowledge Graph

The study of Multi-Modal Knowledge Graphs (MMKGs) attracts increasing attention, as real-world entities are naturally described by heterogeneous information from multiple modalities, such as relational structure, textual descriptions, and visual signals. A central challenge in MMKGs, accordingly, is effective multimodal fusion, which aims to integrate complementary yet heterogeneous information into a unified and semantically coherent representation.

Existing approaches can be broadly cast into five representative categories. (i) **Cross-modal interaction methods** leverage attention-based mechanisms to explicitly model dependencies and information exchange across modalities. (Chen et al., 2020; 2023; Li et al., 2023a;b; Zhang et al., 2025c). (ii) **Consistency learning approaches** employ contrastive objectives to align modality-specific embeddings within a shared latent space, thereby enforcing semantic consistency across modalities (Lin et al., 2022; Lu et al., 2024; Liu et al., 2024). (iii) **Structure-guided methods** incorporate relational biases from the underlying knowledge-graph structure to regulate multimodal interactions and preserve structural semantics (Lee et al., 2023; Zhang et al., 2025a; Cheng et al., 2025). (iv) **Reweighting strategies** dynamically adjust modality contributions to mitigate modality imbalance and dominance during training (Zhang et al., 2024c; 2025b). (v) **LLM-based KGC methods** exploit large language or foundation models to encode structural and multimodal knowledge, thereby enhancing semantic reasoning and knowledge graph completion (Zhang et al., 2024b; Liu et al., 2025b; Guo et al., 2025).

Despite promising empirical performance, existing MMKG fusion methods often overlook a fundamental issue, namely that embeddings from different modalities typically reside in heterogeneous spaces with distinct geometric properties (Gao et al., 2025). Direct fusion may distort modality-specific distributions, leading to semantic inconsistency and incomplete representations (Liang et al., 2022; Hemker et al., 2024; Wei et al., 2024). A recent line of work explores **Optimal Transport (OT)-based fusion**, which aligns heterogeneous distributions by minimizing the transport cost across modalities, to preserve intrinsic modality-specific distributions.

A.2 Optimal Transport

Optimal Transport (OT) theory is proposed to measure the distance between probability distributions by computing the minimal cost required to transport mass from one distribution to another. Formally, given two discrete probability measures $\mathbf{a} \in \Delta^n$ and $\mathbf{b} \in \Delta^m$, and a cost matrix $C \in \mathbb{R}^{n \times m}$, OT seeks a transport plan $P \in \mathbb{R}_+^{n \times m}$ that solves:

$$\min_{P \in \Pi(\mathbf{a}, \mathbf{b})} \langle P, C \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} C_{ij}, \tag{18}$$

where $\Pi(\mathbf{a}, \mathbf{b}) = \{P \geq 0 \mid P\mathbf{1}_m = \mathbf{a}, P^\top \mathbf{1}_n = \mathbf{b}\}$ denotes the set of feasible transport plans that preserve marginal distributions. Due to its ability to align distributions while respecting their intrinsic geometry, OT is widely adopted in multimodal fusion. Representative approaches, including OTKGE (Cao et al., 2022), OTMEL (Zhang et al., 2024d), MOVER (You & Liu, 2025), SLOTAlign (Tang et al., 2023), and DecAlign (Qian et al., 2026), leverage OT to establish soft correspondences across heterogeneous modalities, leading to geometry-aware cross-modal alignment.

Yet, existing OT-based fusion strategies typically rely on a barycenter-style objective, where a unified representation is obtained by aggregating multiple modality-specific embeddings through weighted transport distances. *Such formulations are prone to modality dominance, where one modality disproportionately influences the fused representation, especially when the corresponding weights or distribution scales are not properly calibrated.* To address this limitation, this work introduces a **Nash bargaining-based fusion**, which formulates multi-modal fusion as a multi-objective game.

A.3 Nash Bargaining

Nash bargaining theory aims to resolve conflicts among multiple players by seeking a Pareto-efficient solution that maximizes their collective gains relative to a disagreement baseline (Nash,

1950). In this context, each player is associated with a utility function that quantifies its preference over candidate solutions, and these utilities jointly define the feasible outcome space. Formally, the Nash bargaining solution is obtained via:

$$\max \sum_{i=1}^{K} \log\left(u_i - d_i\right), \quad \text{s.t. } \forall i:\ u_i > d_i, \tag{19}$$

where u_i denotes the utility of player i and d_i represents its disagreement point. This objective ensures balanced improvement across players and prevents any single player from dominating the solution. Nash bargaining-based approaches have since been widely adopted in various tasks, including multi-task learning, fairness-aware optimization, and expert model routing (Nash, 1950; Navon et al., 2022; Zeng et al., 2024; Nguyen et al., 2026; Gupta et al., 2025; Wu & Harandi, 2025). These applications consistently demonstrate strong empirical performance in balancing competing objectives and improving overall performance.

Motivated by this perspective, multimodal fusion can be naturally viewed as a negotiation process among competing modalities, where each modality seeks to maximize its own utility while contributing to a shared representation. This view motivates the use of Nash bargaining to dynamically regulate modality contributions, enabling a more balanced and adaptive fusion process.

B Analysis on Low-Target Relevance

This section provides detailed analyses of low-target relevance in MMKGs. Specifically, 14,463 entities are randomly sampled from MKG-W to examine their modality inputs, including textual descriptions and corresponding visual images. For each modality, the input content is categorized according to its semantic relevance to the target entity. The proportion of each semantic category is then calculated and divided into five intervals, *i.e.*, $[0, 20\%)$, $[20, 40\%)$, $[40, 60\%)$, $[60, 80\%)$, and $[80, 100\%]$, to visualize their statistical distributions.

B.1 Text Modality Analysis

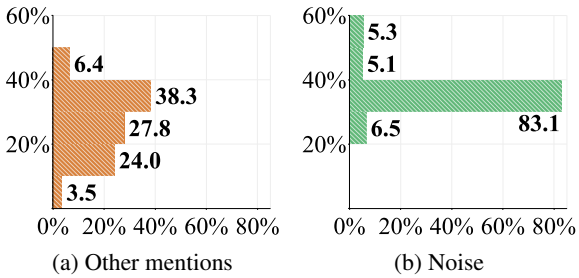
Prompt for segmenting entity-related contents.

Role: You are an expert in Multi-modal Knowledge Graph (MMKG) and Named Entity Recognition (NER).
Task: Given a target entity and its textual description (see below), identify tokens belonging to the following semantic categories:
1. **Target**: tokens that directly describe the target entity, including its name, aliases, abbreviations, attributes, or entity-specific facts.
2. **Other**: tokens associated with other valid entities appearing in the textual description, such as persons, organizations, locations, or concepts. Use NER to identify these entity mentions.
3. **Noise**: all remaining tokens that provide limited entity-specific evidence, including generic background statements, function words, prepositions, or formatting artifacts.
Rules: Each token from the input Textual Description can ONLY belong to exactly one category. Do not infer any entities outside the textual description. Output valid JSON only, without explanations.
Input:
Target Entity: [[target entity]]
Textual Description: [[textual content]]
Output Format:
{
"Target": {"tokens": [], "ratio": []}, "Other": {"tokens": [], "ratio": []}, "Noise": {"tokens": [], "ratio": []}, "total_tokens": []
}

Semantic segmentation. To examine low-target relevance in textual descriptions, a fine-grained *Named Entity Recognition* (NER)-based analysis is conducted to segment entity-related contents. Specifically, given an entity e, its textual description d_e is decomposed into three semantic categories: **Target**, **Other**, and **Noise**. The *Target* category denotes tokens that directly describe e, such as its name, attributes, aliases, abbreviations, or entity-specific facts. The *Other* category refers

to tokens associated with other valid entities appearing in d_e, such as different persons, organizations, locations, or concepts. The *Noise* category contains all remaining tokens that provide limited entity-specific evidence, such as generic background statements, function words, prepositions, or formatting artifacts. Accordingly, GPT-5.4-mini is prompted to identify these semantic categories from d_e, as shown in *Prompt for segmenting entity-related contents*. The ratio of each category with respect to the total number of tokens in d_e is then computed.

Statistical observations. The token ratios are then aggregated and their statistical distributions are visualized in Fig. 1(a) and Fig. 7(a)–(b). To begin with, target-related semantics are not dominant in textual descriptions, accounting for only 35.8% of tokens on average, suggesting that MMKG textual descriptions contain only partial target-specific evidence. Second, other-entity semantics also occupy a substantial proportion, reaching 27.1%, indicating that textual descriptions frequently mix target entities with information from other entities. Third, noise semantics further account for a considerable proportion, reaching up to 37.1%. These observations reveal a clear *low-target relevance* issue in MMKG inputs: the text modality does not always provide clean target-specific context, but instead mixes target semantics with off-target entity information and noisy tokens. Such inconsistent semantics may be directly transported into the shared representation space through OT-based fusion methods, thereby degrading downstream performance.



(a) Other mentions (b) Noise

Figure 7: Empirical analysis using MKG-W. (a)–(b) characterize target relevance in textual modalities by reporting the distributions of other-entity mentions and noise. The x-axis indicates the proportion occupied by each category, and the y-axis indicates the proportion of samples.

B.2 Visual Modality Analysis

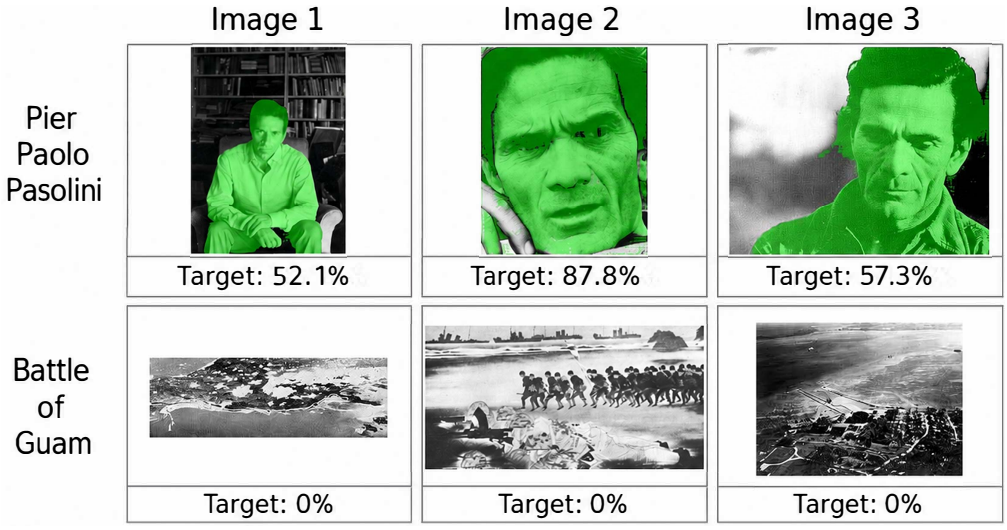


Figure 8: Examples of highlighted target-related regions in visual-modality inputs.

Object detection. Similar to the text modality analysis, a statistical analysis is also conducted on the image modality to examine whether low-target relevance also exists in visual inputs. Notably, images in the MKG-W benchmark are collected from Wiki-based sources, where each image is usually centered on a single target entity rather than multiple co-occurring entities. Therefore, unlike textual descriptions, the image modality is mainly analyzed from the perspective of target-related

and noise (*i.e.*, non-target) semantics, *measured by the pixel area occupied by detected objects*. For entities whose images do not contain clear visual evidence of the target semantics (such as historical events: battle of Guam), the detected target region is set to zero, causing the image modality to be dominated by noisy semantics (see Fig. 8 as examples). Specifically, Grounding DINO (Liu et al., 2025a) is employed with the entity name as the grounding query to detect target-related visual regions, and the detected regions are further annotated using SAM (Kirillov et al., 2023). Based on the segmentation masks, image semantics are accordingly divided into two categories: **Target**, corresponding to regions of the target entity, and **Noise**, corresponding to all remaining non-target visual regions.

Statistical observations. As shown in Fig. 1(d), the proportion of target-related semantics varies substantially across image samples. Only 5.6% of images are dominated by target-related regions ($\geq 80\%$), while the majority of images contain limited target semantics. Surprisingly, 42.8% of images contain approximately 20% target-related regions, equivalently indicating that 80% of visual content is composed of noisy regions. This suggests that the visual modality is often dominated by semantically weak content, such as backgrounds or irrelevant objects, rather than the target entity itself. These observations again reflect the low target relevance issue in MMKG inputs, where visual samples exhibit highly heterogeneous semantic relevance to the target entity.

C Analysis on Modality Imbalance

This section investigates modality imbalance from two perspectives: the correlation between the fused representation and modality-specific representations, and the sensitivity to modality weighting hyperparameters. The former examines whether the fused representation is biased toward a particular modality, while the latter evaluates whether the fusion process heavily depends on manually specified modality weights. Together, these analyses provide a direct characterization of modality imbalance in OT-based MMKG fusion. Specifically, the experimental setting of OTKGE (Cao et al., 2022) is strictly reproduced on the MKG-W dataset using its officially released code.

C.1 Representation Correlation Analysis

To analyze modality contributions in OT-based fusion, the Wasserstein distance (as defined in Eq. equation 2) between the fused representation and each aligned modality representation in OTKGE is examined. Following the original setting from OTKGE, the modality weighting hyperparameters are set to $\lambda_s = 0.1$, $\lambda_v = 0.1$, and $\lambda_t = 0.8$ for the structural, visual, and textual modalities, respectively, and the model is trained for 200 epochs. Based on the obtained fused representations, the distance between each entity's fused representation and its modality-specific embeddings is then computed. The experiment is repeated five times with different random seeds, and the averaged results of distance distributions are visualized in Fig. 1(c).

As observed, the fused representation does not maintain comparable Wasserstein distances to different modalities. Instead, a clear bias toward specific modalities is exhibited: the distance to the textual modality remains consistently lower, staying around 0.1, while the distances to the visual and structural modalities are relatively higher. Specifically, the distance to the visual modality rapidly increases from around 0.5 to above 0.9 and then remains at a high level, whereas the distance to the structural modality gradually increases from around 0.05 to around 0.6 during training. Notably, rather than moving closer to the fused representation during training, the visual and structural modalities become increasingly distant from it, indicating an evident modality imbalance in the fusion process. Overall, this trend suggests that the fused representation is unevenly attracted by different modalities, staying closer to the dominant modality (*e.g.*, the textual modality in this case) while remaining farther from others. Thus, optimizing distribution-level OT alignment alone does not necessarily guarantee balanced multimodal utilization. That is, the learned fused representation may still encode disproportionate modality contributions, leading to modality imbalance during fusion.

C.2 Hyperparameter Sensitivity Analysis

The above analysis reveals uneven modality bias in OT-based fusion. A natural question is whether such imbalance can be mitigated by tuning the modality weighting hyperparameter λ. To examine the impact of modality weighting, for simplicity, a sensitivity analysis is conducted by setting $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, where λ denotes the weighting coefficient of the *structural* modality, and the weighting coefficients of the *textual* and *visual* modalities from OTKGE are both set to $(1-\lambda)/2$, respectively:

$$\lambda_{\text{text}} = \lambda_{\text{image}} = \frac{1-\lambda}{2}.$$

Therefore, we can observe model performance by controlling one single hyperparameter. Then, the model is still trained for 200 epochs, and the MRR curves are plotted over training epochs to evaluate performance sensitivity, as shown in Fig. 1(d). For comparison, the result originally reported by OTKGE (the black reference line) and TransE (the red reference line, where only the structural modality is used) are also included.

It is observed that selecting an appropriate λ value can substantially improve model performance. For instance, when $\lambda = 0.3$, the model outperforms the default setting by 1.2%. In this setting, the fusion process places more emphasis on the `structural` modality, rather than following the original configuration where the `text` modality is preferred. This suggests that improving cross-modal balance helps alleviate modality imbalance, allowing the model to benefit from complementary multimodal information. **However**, the performance also exhibits strong sensitivity to λ. When unsuitable values (say $\lambda = 0.9$, where the textual and visual modalities are barely used) are used, performance (not surprisingly) drops considerably and approaches the structural-only result indicated by the red curve. This shows that poor cross-modal balance can significantly weaken multimodal fusion and amplify modality imbalance.

Notably, existing OT-based methods typically adopt a static λ and rely on extensive validation to identify a suitable value for it. Such a strategy is label-intensive, time-consuming, and lacks generalizability. Overall, these results suggest that OT-based fusion remains highly sensitive to cross-modal balance, indicating a clear modality imbalance issue that existing fusion strategies cannot adaptively address.

D Pseudocode for NaCOT

Algorithm 1 summarizes the full procedure of the proposed NaCOT. *Line 1* and *Lines 3–4* perform Cascaded Representation Construction (CRC), where the fused representation is decomposed into multi-scale target-side candidates and each modality representation is projected and transformed into source-side candidates, respectively. *Lines 5–10* and *Line 12* conduct SWD-based Candidate Selection (SCS) and estimate the initial OT alignment, weight and fused representation. *Lines 13–22* perform Nash-guided Modality Fusion (NMF). Each iteration in NMF evaluates the expense and utility of each modality, solves the Nash bargaining problem under the feasible consensus constraint, updates the bargaining weights, and then refines the fused representation in *Line 21*. The fused representation $Y^{(T)}$ is taken as the output of NaCOT, where T is the Nash iteration.

E Theoretical Analysis

This section provides theoretical analysis of the proposed Nash-guided Modality Fusion (NMF). Since NMF formulates modality coordination as a bargaining process among modality-specific utilities, its optimization is closely related to the Nash Bargaining Problem (NBP). NBP relies on four standard axioms to justify the existence and uniqueness of the equilibrium solution, which can be boiled down to convexity/concavity of the utility function and strategy space (Nash, 1950). In our iterative optimization, each iteration maximizes the Nash product in Eq. equation 14, corresponding to a Nash equilibrium solution. Therefore, we *analyze the convexity/concavity of two key components: i.e., the utility function and the strategy space*.

First, the utility function defined in Eq. equation 13 is *concave*. That is, the F-norm in Eq. equation 12 is convex, the inner product $\langle P_m, C_m \rangle$ in $R(P_m)$ is linear and the entropy in $R(P_m)$ is

convex. Adding convex functions to linear functions gives a convex function, which becomes concave in our final utility function. This implies *a unique maximum if the feasible set is compact*.

Second, the strategy space in Eq. equation 15, *i.e.*, the feasible set, is defined by linear equality constraints over all transport matrices. Since each $P_m \in \mathcal{P}_m$ is a probability matrix satisfying marginal constraints, this leads to the fact that *the feasible set is convex and compact*, we simply need to *ensure that the feasible set is non-empty*.

To make this part independent, we redefine the dimensions of the matrices in Eq. equation 15. Let $P_m \in \mathbb{R}^{n_m \times N}$, $X_m \in \mathbb{R}^{n_m \times d}$, and $Y \in \mathbb{R}^{N \times d}$. The total degree of freedom is $N \sum_m n_m$ and the reduction of degree of freedom from the equality is Nd and $(N + n_m - 1)M$ from the probability matrices constraint, where M is the total number of modalities, *i.e.* $|\mathcal{M}|$. The condition for $\mathcal{U}^{(t)}$ being non-empty is to have nonzero degree of freedom in the system, *i.e.*

$$N \sum_m n_m - Nd - (N + n_m - 1)M > 0.$$

If $n_m = n,\ \forall m$, then we simplify to

$$NMn - Nd - (N + n - 1)M > 0.$$

This is easy to satisfy in practice since in general M is small, while N and n are quite large. Since d is the model dimension, and apparently a choice we have to make. From above, we have the following requirement

$$M \frac{(N-1)(n-1)}{N} > d.$$

With some approximation, it can be further simplified as

$$Mn > d.$$

This coincides with the standard requirement of observational data analysis, *i.e.* the number of data points observed must be higher than the dimensionality. Under this condition, *the feasible set is non-empty and compact*. Therefore, this proposed NBP will have a unique equilibrium.

F Computational Complexity

In this section, we briefly analyze the computational complexity of NaCOTand existing OT-based baselines, including OTKGE (Cao et al., 2022), OTMEL (Zhang et al., 2024d), and DecAlign (Qian et al., 2026). Again, for modality m, let $P_m \in \mathbb{R}^{n_m \times N}$ denote the transport matrix, $X_m \in \mathbb{R}^{n_m \times d}$ the modality-specific representation, and $Y \in \mathbb{R}^{N \times d}$ the fused representation. Let $M = |\mathcal{M}|$ denote the number of modalities, and let L denote the number of Sinkhorn iterations used by the traditional OT solver.

OTKGE. For modality m, OTKGE first constructs the transport cost matrix between n_m modality-side supports and N target-side supports, costing $\mathcal{O}(n_m N d)$. The Sinkhorn algorithm is then applied to compute the OT plan, introducing the cost of $\mathcal{O}(L n_m N)$. After obtaining the transport plan, barycentric projection computes $P_m^\top X_m$, where $P_m \in \mathbb{R}^{n_m \times N}$ and $X_m \in \mathbb{R}^{n_m \times d}$, resulting in a complexity of $\mathcal{O}(n_m N d)$. Therefore, the total complexity is:

$$\boxed{\text{OTKGE}: \quad \mathcal{O}\left(\sum_{m \in \mathcal{M}} (2 n_m N d + L n_m N)\right).}$$

OTMEL. For modality m, OTMEL follows a similar OT-based alignment procedure as OTKGE. It constructs a cost matrix between n_m modality-side representations and N target-side representations, which costs $\mathcal{O}(n_m N d)$. The Sinkhorn solver introduces an additional cost of $\mathcal{O}(L n_m N)$, and the barycentric projection costs $\mathcal{O}(n_m N d)$. Summing over all modalities, the total complexity is also

$$\boxed{\text{OTMEL}: \quad \mathcal{O}\left(\sum_{m \in \mathcal{M}} (2 n_m N d + L n_m N)\right).}$$

Yet, to improve efficiency, OTMEL further employs knowledge distillation to transfer global OT assignment knowledge into a one-pass attention mechanism. In the accelerated version, the iterative Sinkhorn solver is replaced at inference by a query–key attention map, which removes the L Sinkhorn iterations required by the standard OT solver. Thus, the complexity becomes

$$\boxed{\text{OTMEL}: \quad \mathcal{O}\left(\sum_{m\in\mathcal{M}} 2n_m N d\right).}$$

DecAlign. For modality m, DecAlign first maps and decouples modality features, which costs $\mathcal{O}(n_m d^2)$. Then, it performs heterogeneous alignment and homogeneous alignment, which costs $\mathcal{O}\big(n_m d^2 + d^3 + M n_m^2 d\big)$ in total. After alignment, computing the attention map over n_m modality-side supports costs $\mathcal{O}(n_m^2 d)$, while the linear projections and feed-forward layers cost $\mathcal{O}(n_m d^2)$. Therefore, the time complexity of DecAlign is

$$\boxed{\begin{array}{l}\text{DecAlign}: \\ \mathcal{O}\left(\sum_{m\in\mathcal{M}} \big((M+1)n_m^2 d + 3n_m d^2\big) + M d^3\right).\end{array}}$$

NaCOT. In NaCOT, for modality m, CRC first transforms modality features into cascaded representations, which costs $\mathcal{O}(n_m^2 d)$. The target-side compression performed on Y costs $\mathcal{O}(N^2 d)$. SCS computes SWD scores to select the transport pair, and its cost is lower-order compared with the CRC costs. After the selection, for modality m, computing pairwise costs between n_m modality-side supports and N target-side supports costs $\mathcal{O}(n_m N d)$, and solving the initial entropic OT problem contributes at most $\mathcal{O}(L n_m N)$. During NMF, all modalities jointly participate in the Nash bargaining update. For modality m, the dominant computation is evaluating the transport-aligned representation $P_m^\top X_m$ and its corresponding utility, which costs $\mathcal{O}(n_m N d)$. Hence, the total time complexity of NaCOT is

$$\boxed{\begin{array}{l}NaCOT: \\ \mathcal{O}\left(\sum_{m\in\mathcal{M}} \big(n_m^2 d + N^2 d + 2n_m N d + L n_m N\big)\right)\end{array}}$$

Notably, $\mathcal{O}(L n_m N)$ is a worst-case upper bound. Since in NaCOT, OT is performed on the compressed representation pair selected by SCS, the actual OT cost is often lower in practice (such as $\mathcal{O}(L \times \eta_Z \times n_m \times \eta_Y \times N)$, with $\eta_{Z/Y} < 1$).

Overall, the additional components introduced by NaCOT do not change the dominant complexity term compared with standard OTKGE. To complement the theoretical analysis, we further report empirical training time and GPU memory usage in Table 2. The results show that NaCOT maintains time efficiency comparable to OTKGE while achieving stronger empirical performance. The accelerated version of OTMEL runs faster by replacing iterative OT with attention maps, but maintaining these maps substantially increases GPU memory usage compared with NaCOT. Compared with DecAlign, NaCOT achieves a more favorable efficiency–performance trade-off, requiring less training time and GPU memory.

Table 2: Comparison of the model training time (second) and GPU memory usage (MB) of different OT-based methods.

Method	Time	GPU Memory
OTMEL	20	9229
DecAlign	129	3207
OTKGE	76	2543
NaCOT	74	2321

Although NaCOT introduces additional CRC and NMF computations, the training time it measures is slightly shorter than that of OTKGE. This is primarily because OTKGE applies the Sinkhorn

solver directly to the full-scale transfer matrix $P_m \in \mathbb{R}^{n_m \times N}$, whereas CRC and SCS allow NaCOT to solve the OT problem only for selected compressed source-target pairs. With the default parameters $\eta_Z = \eta_Y = 0.8$, even with just one compression step, the number of paired transmission elements can be reduced to approximately $0.8 \times 0.8 = 64\%$ of the original size. If a deeper compression layer is chosen, the compression ratio may be higher. Furthermore, with the default settings ($T = 3$, Appendix $G.3$), NMF converges in three iterations, so its additional overhead remains less than the time saved by reducing the OT problem. This both explains the observed time difference and is consistent with the worst-case complexity analysis discussed above.

G Additional Experiments

G.1 Statistics of Datasets

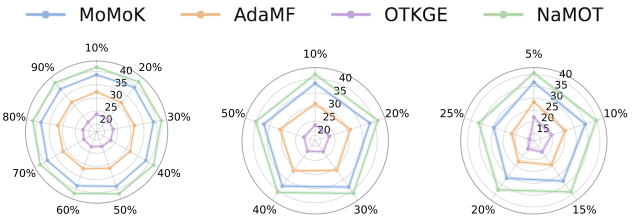
Four MMKG datasets are used in our experiments: MKG-W, MKG-Y, DB15K, and KVC16K. Among them, MKG-W, MKG-Y, and DB15K are derived from Wikidata (Vrandečić & Krötzsch, 2014), YAGO (Suchanek et al., 2007), and DBpedia (Mendes et al., 2012), respectively, while KVC16K is adapted from KuaiPedia (Pan et al., 2022), a micro-video encyclopedia. These datasets are all real-world multimodal knowledge graphs enriched with visual and textual modalities. Table 3 summarizes the dataset statistics, including the numbers of entities, relations, and train/validation/test triples, together with the availability and dimensionality of visual and textual features.

Table 3: Statistics of the four datasets.

Statistic	MKG-W	MKG-Y	DB15K	KVC16K
#.Entity	15,000	15,000	12,842	16,015
#.Relation	169	28	279	4
#.Train	34,196	21,310	79,222	180,190
#.Valid	4,276	2,665	9,902	22,523
#.Test	4,274	2,663	9,904	22,525

G.2 Model Robustness

This section evaluates the robustness of NaCOT under three challenging multimodal conditions: modality noise, modality missing, and sparse links, corresponding to corrupted modality features, incomplete multimodal observations, and insufficient structural supervision, respectively. More precisely, modality noise is simulated by adding Gaussian noise to a proportion of selected input features, with the noise ratio ranging from 10% to 90%. Modality missing is created by randomly removing single-modal features from a proportion of entities, with the missing ratio ranging from 10% to 50%. Sparse links are generated by randomly discarding training triples, with the sparsity ratio ranging from 5% to 25%. For fair comparison with prior studies, all robustness experiments are conducted on `DB15K`. We specifically compare NaCOT with the following baselines: OTKGE (Cao et al., 2022), AdaMF (Zhang et al., 2024c), and MoMoK (Zhang et al., 2025b), covering both OT and non-OT based alignment methods.



(a) Modality Noise (b) Modality Missing (c) Sparse Links

Figure 9: MRR results on DB15K under three challenging scenarios: modality noise, modality missing, and link sparsity.

As shown in Fig. 9, NaCOT consistently achieves the best MRR across all three scenarios. Under modality noise, NaCOT remains stable across all noise ratios and changes more smoothly than the baselines, indicating stronger tolerance to corrupted modality signals. Under modality missing, NaCOT maintains a clear advantage, suggesting that it better preserves useful cross-modal evidence when part of the multimodal information is unavailable. Under sparse links, all methods suffer per-

formance degradation, but NaCOT still retains the best performance across different sparsity ratios. Overall, these results support the robustness of NaCOT: compatible OT helps select more reliable alignment candidates, while Nash-guided fusion adaptively coordinates modality contributions under noisy, incomplete, and sparse conditions.

G.3 Hyperparameter Sensitivity Analysis

To evaluate the stability of NaCOT under different hyperparameter settings, we analyze the final fusion dimension d, the OT penalty τ, and the Nash regularization coefficient γ on MKG-W and KVC16K. We additionally study the source- and target-side construction depths L and K, together with the Nash iteration number T, on MKG-W. The effects of the compression budgets η_Z and η_Y have been analyzed in the *"Impact of CRC"* subsection of Section 4.3.

Table 4: Sensitivity analysis of NaCOThyperparameters, including the final fusion dimension d, the OT penalty hyperparameter τ and the Nash regularization coefficient γ.

Setting	MKG-W		KVC16K	
	MRR	**Hit@1**	**MRR**	**Hit@1**
$d = 100$	35.50	27.61	22.62	15.00
$d = 200$	39.86	30.77	26.11	16.69
$d = 300$	41.49	32.16	27.55	17.33
$d = 400$	41.61	**32.29**	27.72	17.35
$d = 500$	**41.70**	32.27	**27.73**	**17.40**
$\gamma = 0.01$	37.71	29.05	24.74	15.59
$\gamma = 0.05$	40.52	31.27	26.87	17.06
$\gamma = 0.1$	**41.49**	**32.16**	**27.55**	**17.33**
$\gamma = 0.2$	41.22	31.89	26.71	16.99
$\gamma = 0.5$	39.18	31.37	25.20	16.61
$\tau = 0.01$	37.70	29.56	24.83	16.32
$\tau = 0.05$	40.96	31.74	26.93	17.18
$\tau = 0.1$	**41.49**	**32.16**	**27.55**	**17.33**
$\tau = 0.2$	41.00	31.93	27.21	17.08
$\tau = 0.5$	40.59	31.41	26.98	16.71

As observed from Table 4, NaCOT remains relatively stable across different settings of d, τ, and γ. In general, moderate hyperparameter values consistently achieve the best performance, while excessively small or large values lead to noticeable degradation. Moreover, increasing the representation dimension beyond $d = 300$ only brings marginal improvements, suggesting that the performance gains mainly originate from the proposed fusion mechanism rather than simply enlarging the embedding size. Compared with MKG-W, KVC16K exhibits slightly larger fluctuations under extreme hyperparameter settings, likely due to its larger entity number and stronger multimodal ambiguity. Overall, the relatively limited performance variation across different settings demonstrates that NaCOT is not highly sensitive to hyperparameter selection and maintains stable performance over a broad parameter range.

G.3.1 Sensitivity to Construction Depths and Nash Iterations

The construction depths L and K determine the sizes of the source- and target-side candidate sets rather than directly specifying the final transported representations. Since SCS subsequently selects a compatible representation pair from these sets, NaCOT is expected to be less sensitive to the individual choices of L and K. We evaluate this behavior by jointly varying both depths over $\{3, 5, 7, 9\}$. We further vary the Nash iteration number T over $\{2, 3, 5, 7, 9\}$ while fixing $L = K = 5$.

Table 5: Sensitivity of NaCOT to the construction depths L and K, and the Nash iteration number T, on MKG-W in terms of MRR.

(a) Joint impact of L and K

L/K	**3**	**5**	**7**	**9**
3	36.28	38.67	38.67	38.68
5	39.06	**41.49**	**41.49**	**41.46**
7	39.05	**41.49**	**41.47**	**41.51**
9	39.08	**41.50**	**41.49**	**41.50**

(b) Impact of T with $L = K = 5$

T	**2**	**3**	**5**	**7**	**9**
MRR	39.35	**41.49**	41.53	41.52	41.53

As shown in Table 5(a), performance stabilizes once both L and K reach 5. When $L = K = 3$, the limited candidate sets may not contain a sufficiently compatible representation pair, resulting in an MRR of 36.28. Increasing both depths to 5 improves MRR to 41.49, while further increasing them produces nearly identical results. With $\eta_Z = \eta_Y = 0.8$, deeper cascade levels generate increasingly compressed candidates, whose dimensions decrease approximately with η_Z^ℓ and η_Y^k. Such overly compressed candidates tend to have larger sliced Wasserstein distances and are therefore less likely to be selected by SCS.

Table 5(b) shows that increasing T from 2 to 3 improves MRR from 39.35 to 41.49. Additional iterations produce variations of at most 0.04 MRR points, indicating that the Nash equilibrium can be effectively reached within three iterations. Together, these results show that the default setting $L = K = 5$ and $T = 3$ provides stable, near-best performance without requiring large candidate sets or many bargaining iterations.

G.4 Modality Contribution

This section investigates the impact of different modality inputs on NaCOT. The full-modality setting uses structural, textual, and visual modalities, where the structural modality provides essential relational topology for knowledge graph reasoning. Therefore, the structural modality is retained. Accordingly, two variants are considered: *Structural–Text* and *Structural–Visual*, while keeping all other components unchanged. This design allows us to examine whether textual and visual modalities provide complementary information beyond structural representations. The MOMOK baseline is employed hereafter for the comparison purposes.

Table 6: Modality contribution on MKG-W and KVC16K.

Method	**Variant**	**MKG-W**		**KVC16K**	
		MRR	**Hit@1**	**MRR**	**Hit@1**
MOMOK	**Str–Txt**	34.65	28.54	15.12	9.92
	Str–Img	33.62	28.13	14.51	9.30
	Full	35.89	30.38	16.87	10.53
NaCOT	**Str–Txt**	39.87	30.48	25.74	15.76
	Str–Img	39.56	30.12	25.28	15.34
	Full	41.49	32.16	27.55	17.33

As shown in Table 6, all reduced-modality variants underperform the full-modality setting, indicating that structural, textual, and visual modalities all provide useful signals for MMKGC. For instance, with NaCOT, the *Str–Txt* variant achieves MRR scores of 39.87 on MKG-W and 25.74 on KVC16K, which remain lower than the full three-modality results of 41.49 and 27.55. Nevertheless, NaCOT still outperforms MOMOK under reduced-modality settings. For instance, the *Str–Img* variant of NaCOT achieves an MRR of 39.56 on MKG-W, substantially higher than MOMOK at 33.62. These results demonstrate that textual and visual modalities provide complementary semantics beyond structural representations, and verify that NaCOT effectively exploits and coordinates information from all modalities through Nash-guided fusion.

G.5 Impact of Different Backbones

To further evaluate the generalizability of NaCOT across different backbones, additional experiments are conducted on MKG-W using combinations of visual encoders (VGG (Simonyan & Zisserman, 2015) and CLIP (Radford et al., 2021)) and textual encoders (BERT (Devlin et al., 2019) and Llama2-7B (Touvron et al., 2023)). For all settings, the embedding dimension is set to 300, the compression budgets η_Z and η_Y are both set to 0.8, and the construction depths are set to $L = K = 5$. AdamW is used for optimization, with learning rates of 2×10^{-5} for CILP and 1×10^{-4} for Llama2-7B-based settings, respectively.

As observed from Fig. 10, stable performance is consistently achieved across different backbone settings, demonstrating the robustness and compatibility of NaCOT with diverse backbone architectures. In addition, stronger language-model backbones such as Llama2 generally yield better performance, indicating the potential of integrating NaCOT with LLM-based encoders for further gains.

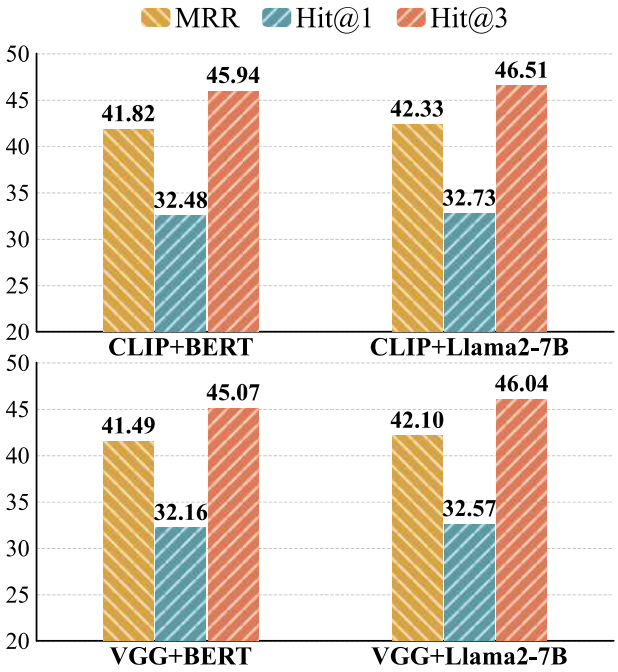


Figure 10: Performance results with different visual and textual backbone combinations on MKG-W.

G.6 Transferability to Multimodal Entity Linking

Although NaCOT is developed for MMKGC, its three components operate on modality-specific and fused representations without relying on a particular KG prediction head. We therefore further evaluate its transferability to multimodal entity linking (MEL), which links a multimodal mention to its corresponding entity in a knowledge base. MEL provides a suitable test bed because its textual and visual inputs can also contain noisy semantics and exhibit uneven modality quality.

For this adaptation, the MMKGC scoring function is replaced with a mention–entity similarity head, while CRC, SCS, and NMF remain unchanged. The hidden representation dimension is set to 96, the batch size is 256, and the learning rate is 5×10^{-6}. We conduct experiments on WikiMEL, RichpediaMEL, and WikiDiverse, and compare against CLIP (Radford et al., 2021), OT-MEL (Zhang et al., 2024d), and M^{3}EL (Hu et al., 2025b).

Table 7: MRR results (%) on three multimodal entity linking benchmarks.

Method	WikiMEL	RichpediaMEL	WikiDiverse
CLIP (Radford et al., 2021)	88.23	77.57	71.69
OT-MEL (Zhang et al., 2024d)	92.59	88.27	75.43
M^{3}EL (Hu et al., 2025b)	92.30	88.26	81.29
NaCOT	**92.87**	**89.83**	**84.29**

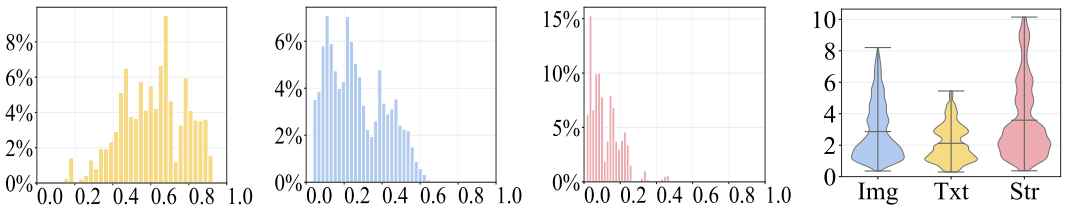
As shown in Table 7, NaCOT consistently achieves the best MRR across all three datasets. Compared with the strongest baseline on each dataset, it improves MRR by 0.28, 1.56, and 3.00 points on WikiMEL, RichpediaMEL, and WikiDiverse, respectively. The larger improvement on WikiDiverse suggests that compatible transport and Nash-guided coordination are particularly

helpful when multimodal mentions exhibit greater semantic diversity. These results demonstrate that NaCOT is transferable beyond MMKGC and remains effective on the MEL task.

H Case Study

H.1 Case Study on MKG-W

To better understand how NaCOT adaptively coordinates multimodal information, we visualize the learned modality weights on MKG-W.



(a) Textual (b) Visual (c) Structural (d) Wasserstein Distance($\mathcal{W}$)

Figure 11: Case analysis on MKG-W. (a)–(c) Distribution of normalized modality weights for textual, visual, and structural modalities learned by NaCOT, respectively. (d) Distribution of Wasserstein distances $\mathcal{W}$ between each modality-specific representation and the fused representation.

Specifically, Fig. 11(a)–(c) show the distributions of modality weights in the fused representations. The learned weights vary substantially across entities, indicating that NaCOT adaptively adjusts modality contributions according to entity-specific multimodal semantics, rather than relying on static global fusion weights as in traditional methods such as OTKGE. This result clearly demonstrates the effectiveness of Nash-guided adaptive coordination in handling heterogeneous multimodal evidence.

Fig. 11(d) further presents the distributions of Wasserstein distances $\mathcal{W}$ between each modality-specific representation and the fused representation. As observed, the distances across different modalities become noticeably more balanced after applying NaCOT. This indicates that the fused representation achieves more consistent cross-modal semantic alignment, suggesting that NaCOT effectively alleviates modality imbalance during multimodal fusion.

H.2 Improvement analysis on KVC16K

The performance improvement of NaCOT is particularly pronounced on KVC16K. Unlike the other benchmarks, KVC16K is adapted from a short-video encyclopedia, where visual and textual inputs often contain cluttered backgrounds, transient content, and descriptions that are only partially relevant to the target entity. Moreover, KVC16K contains only four relation types, compared with 169, 28, and 279 relations in MKG-W, MKG-Y, and DB15K, respectively. Its weaker relation diversity provides less structural supervision for distinguishing entities, making the model more dependent on potentially noisy multimodal evidence. These characteristics can amplify both low-target relevance and modality imbalance.

This setting directly matches the motivation of NaCOT. CRC constructs representation candidates at multiple compression scales, while SCS selects a compatible source–target pair before OT alignment, reducing the influence of target-irrelevant modality semantics. NMF then adaptively coordinates the transported modality representations, preventing one modality from dominating the fused representation. Consequently, the advantage of jointly handling these two challenges becomes more visible on KVC16K.

To provide an entity-level view of this behavior, Table 8 reports the Wasserstein distances between each modality-specific representation and the fused representation for four representative entities during training. The initially large discrepancies among textual, visual, and structural modalities generally decrease and become more balanced as training proceeds.

Table 8: Entity-level Wasserstein distances between modality-specific and fused representations on KVC16K at different training epochs.

Entity	Epoch	Textual	Visual	Structural
Dapeng Bird	1	13.71	6.07	4.32
	50	11.46	7.36	5.65
	100	9.00	7.17	6.13
	150	7.87	7.23	7.43
	200	6.96	6.07	6.99
Wu Song	1	8.60	28.82	10.37
	50	11.75	20.33	13.36
	100	13.31	18.06	14.61
	150	11.77	13.63	12.17
	200	9.85	11.89	10.10
Elm Bonsai	1	11.15	10.27	8.13
	50	10.01	15.02	8.99
	100	9.78	12.73	7.77
	150	7.98	8.92	6.04
	200	6.41	6.73	5.94
Jinghong	1	29.97	17.93	5.12
	50	21.36	13.28	7.44
	100	15.73	11.91	8.81
	150	9.03	8.80	7.93
	200	6.83	6.80	5.79

For example, the visual modality of *Wu Song* initially has a substantially larger distance (28.82) than its textual (8.60) and structural (10.37) modalities. By epoch 200, the three distances converge to 9.85, 11.89, and 10.10, respectively. A similar convergence pattern is observed for the other entities, particularly *Jinghong*, whose initial cross-modal distance gap decreases from 24.85 to 1.04. These cases indicate that NaCOT progressively improves cross-modal consistency while retaining entity-specific modality contributions, helping explain its larger gains on the more ambiguous KVC16K benchmark.

In conclusion, NaCOT differs from existing methods by jointly addressing LTR and MI. Its advantage is therefore more significant on KVC16K, where irrelevant multimodal semantics and modality imbalance are both particularly severe.

I Further Discussion

I.1 Task-level Distinction

NaCOT explicitly identifies and jointly addresses two closely coupled challenges in MMKG fusion: low-target relevance and modality imbalance. A modality with stronger target-relevant evidence may dominate the fusion process, whereas a low-relevance modality may introduce noise and destabilize fusion. Existing approaches generally focus on either filtering unreliable modality evidence or balancing modality contributions, while their interaction is rarely modeled explicitly. Table 9 summarizes the distinction between representative methods. To the best of our knowledge, NaCOT is the first framework to characterize these two issues jointly and address them within a unified MMKGC fusion process.

Table 9: Comparison of representative MMKGC methods in addressing low-target relevance and modality imbalance. "Not explicitly modeled" indicates that the corresponding issue is not a primary design target.

Method	Low-target relevance	Modality imbalance
OTKGE (Cao et al., 2022)	Not explicitly modeled	Not explicitly modeled
AdaMF (Zhang et al., 2024c)	Not explicitly modeled	Adaptive weight reassignment
MOMOK (Zhang et al., 2025b)	Not explicitly modeled	Modality-level expert weighting
HFR-MKGC (Wang et al., 2026)	Not explicitly modeled	Hierarchical fusion
LBMKGC (Guo et al., 2025)	Not explicitly modeled	LLM-based balancing
MoSE (Zhao et al., 2022)	Modality-split relation learning	Not explicitly modeled
LAFA (Shang et al., 2024)	Link-aware visual-noise filtering	Not explicitly modeled
SNAG (Chen et al., 2025)	Modality-level noise masking	Not explicitly modeled
NaCOT	**CRC + SCS**	**NMF**

I.2 Methodological Distinction

Methodologically, NaCOT introduces three components, i.e., CRC, SCS, and NMF, that extend beyond standard OT or Nash methods. Existing OT-based fusion methods, such as OTKGE, directly transport raw, full-scale modality representations. This can be viewed as a special case of CRC in which no cascaded compression is applied. In contrast, CRC constructs multi-scale representation candidates, and SCS selects a compatible source–target pair before solving OT. Therefore, existing OT-based methods overlap only with a special case of CRC in which the full-scale representation is used without cascaded compression. The selected representations therefore determine the support on which the initial transport plan is estimated.

In addition, NMF differs from existing Nash-based approaches by defining modality-specific utilities directly over the transport plan (in Eq.(12)) produced by the preceding CRC. Rather than using traditional Nash bargaining as a standalone weighting scheme, NMF optimizes from CRC according to cross-modal utility, which in turn refines the alignment and forms a closed-loop interaction between three components. Moreover, we also provide additional theoretical analysis in Appendix E by examining the concavity/convexity of the proposed utility function and strategy space, thereby supporting the validity and solvability of the resulting bargaining solution.

In conclusion, NaCOT goes beyond a combination or incremental extension of existing OT- and Nash-based methods. Motivated by a quantitative analysis of the closely coupled LTR and MI issues, it integrates CRC, SCS, and NMF into a unified framework to improve transport alignment and balance modality fusion.

J Limitations

This work is the first attempt to introduce Nash Bargaining theory into multimodal fusion for MMKGs. Despite its strong empirical performance, the proposed framework still has several limitations. First, although the proposed Nash-guided fusion mechanism consistently outperforms existing baselines, the current utility formulation mainly considers reconstruction-based modality coordination. More advanced utility designs, such as contrastive or task-adaptive utilities, may further improve the flexibility and robustness of Nash-guided fusion. Second, while NaCOTemploys SWD-based candidate selection to avoid exhaustive OT over all scale combinations, computational efficiency remains another important consideration. Future work may incorporate sparse transport or mini-batch OT solvers to further improve the scalability of Nash-guided compatible OT.

Algorithm 1 Pseudocode of NaCOT.

Require: Modality feature matrices $\{\widetilde{X}_m\}_{m\in\mathcal{M}}$, initial fused representation Y, compression budgets η_Z, η_Y, construction depths L, K, distance functions $\{c_m\}_{m\in\mathcal{M}}$, hyperparameters $\tau, \gamma, \varepsilon$, Nash iteration T.

Ensure: Final fused representation $Y^{(T)}$.

Construct target-side candidates $\mathcal{Y} = \{Y^{(k)}\}_{k=1}^{K}$ by truncated SVD.

for $m \in \mathcal{M}$ **do**

Project modality features: $X_m = \widetilde{X}_m W_m$.

Construct source-side candidates $\mathcal{Z}_m = \{Z_m^{(\ell)} \mid 1 \le \ell \le L\}$ by

$$Z_m^{(0)} = X_m, \qquad Z_m^{(\ell+1)} = A_m^{(\ell)} Z_m^{(\ell)}.$$

Select the compatible source–target pair:

$$(Z_m^*, Y_m^*) = \arg \min_{Z_m^{(\ell)} \in \mathcal{Z}_m,\, \overline{Y}^{(k)} \in \mathcal{Y}} \mathrm{SWD}\Big(Z_m^{(\ell)}, Y^{(k)}\Big).$$

Construct cost matrix C_m with $C_m[i,j] = c_m(Z_m^*[i], Y_m^*[j])$.

Compute the initial entropic OT plan:

$$\hat{P}_m = \arg \min_{P_m \in \hat{\mathcal{P}}_m} \langle P_m, C_m \rangle + \tau \Omega(P_m).$$

Calculate the target-aligned transport plan:

$$P_m^\top = B_m^{(1)*} \cdots B_m^{(k)*} \operatorname{diag}(\mathbf{q}_m)^{-1} \hat{P}_m^\top A_m^*.$$

Set $P_m^{(0)} = P_m$ and $\omega_m^{(0)} = 1/|\mathcal{M}|$.

end for

Initialize

$$Y^{(0)} = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} (P_m^{(0)})^\top X_m.$$

for $t = 1$ to T **do**

Calculate expense

$$\exp_m(P_m, Y^{(t-1)}) = \frac{1}{2} \left\| P_m^\top X_m - Y^{(t-1)} \right\|_F^2 + \gamma R(P_m),$$

Calculate utility:

$$U_m^{(t)}(P_m) = \exp_m(P_m^{(t-1)}, Y^{(t-1)}) - \exp_m(P_m, Y^{(t-1)}).$$

Calculate the feasible bargaining set:

$$\mathcal{U}^{(t)} = \left\{ \{P_m\}_{m\in\mathcal{M}} \;\middle|\; \sum_{m\in\mathcal{M}} \omega_m^{(t-1)} P_m^\top X_m = Y^{(t-1)},\ P_m \in \mathcal{P}_m, \forall m \right\}.$$

Update transport plans by Nash bargaining:

$$\{P_m^{(t)}\}_{m\in\mathcal{M}} = \arg \max_{\{P_m\}\in\mathcal{U}^{(t)}} \sum_{m\in\mathcal{M}} \log\Big(U_m^{(t)}(P_m) + \varepsilon\Big).$$

for $m \in \mathcal{M}$ **do**

Update

$$\omega_m^{(t)} = \frac{1}{U_m^{(t)}(P_m^{(t)}) + \varepsilon}.$$

end for

Update the fused representation:

$$Y^{(t)} = \frac{\sum_{m\in\mathcal{M}} \omega_m^{(t)} (P_m^{(t)})^\top X_m}{\sum_{m\in\mathcal{M}} \omega_m^{(t)}}.$$

end for

return $Y^{(T)}$.