Interactive Cross-modal Learning for Text-3D Scene Retrieval

Yanglin Feng[1], **Yongxiang Li**[1], **Yuan Sun**[2], **Yang Qin**[1], **Dezhong Peng**[1,3], **Peng Hu**[1]*
[1]College of Computer Science, Sichuan University, Chengdu, China.
[2]National Key Laboratory of Fundamental Algorithms and Models
for Engineering Numerical Simulation, Sichuan University, Chengdu, China.
[3]Tianfu Jincheng Laboratory, Chengdu, China.
fcyzfyl@163.com, rhythmli.scu@gmail.com, sunyuan_work@163.com,
qinyang.gm@gmail.com, pengdz@scu.edu.cn, penghu.ml@gmail.com

Abstract

Text-3D Scene Retrieval (T3SR) aims to retrieve relevant scenes using linguistic queries. Although traditional T3SR methods have made significant progress in capturing fine-grained associations, they implicitly assume that query descriptions are information-complete. In practical deployments, however, limited by the capabilities of users and models, it is difficult or even impossible to directly obtain a perfect textual query suiting the entire scene and model, thereby leading to performance degradation. To address this issue, we propose a novel **I**nteractive Text-3**D** Scene Retrieval Method (IDeal), which promotes the enhancement of the alignment between texts and 3D scenes through continuous interaction. To achieve this, we present an Interactive Retrieval Refinement framework (IRR), which employs a questioner to pose contextually relevant questions to an answerer in successive rounds that either promote detailed probing or encourage exploratory divergence within scenes. Upon the iterative responses received from the answerer, IRR adopts a retriever to perform both feature-level and semantic-level information fusion, facilitating scene-level interaction and understanding for more precise re-rankings. To bridge the domain gap between queries and interactive texts, we propose an Interaction Adaptation Tuning strategy (IAT). IAT mitigates the discriminability and diversity risks among augmented text features that approximate the interaction text domain, achieving contrastive domain adaptation for our retriever. Extensive experimental results on three datasets demonstrate the superiority of IDeal. Code is available at https://github.com/Yangl1nFeng/IDeal.

1 Introduction

Recent years have witnessed natural language interfaces to embodied intelligence systems [1, 2, 3, 4, 5] become increasingly prevalent in our daily lives. This opens up further opportunities for natural language-based interaction with intelligent agents, such as a user verbally instructing agents to perform tasks in a specific scene. Before executing any task, an agent must first retrieve the scene relevant to the user's intent. This requirement has spurred recent works on Text-3D Scene Retrieval (T3SR) [6, 7], which enables the retrieval of 3D point-cloud scenes using linguistic queries. Such a language-scene alignment capability lays a critical foundation for enabling agents to generalize across scenes and environments.

Although existing dedicated methods [6] achieve promising performance for T3SR by facilitating fine-grained query text and scene understanding, such success often relies on the assumption that

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

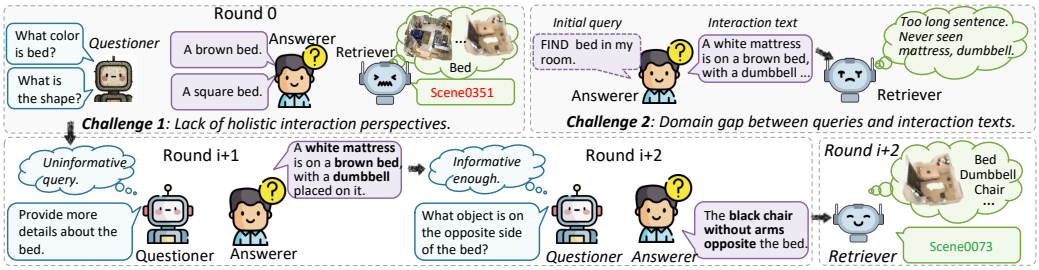


Figure 1: Overview of interactive text-3D scene retrieval. The above gray part illustrates the two specific challenges, while the below white part shows the illustrations of our interactive framework.

the queries provided are information-complete. However, such an assumption is often violated in real-world scenarios due to the inherent limitations of text inputs and models, such as incomplete one-shot descriptions of user intent [8], ambiguous descriptions [6], domain shifts [9], and limited generalization of the models. As a result, the performance and robustness of the models remain persistently constrained, and relying solely on limited internal knowledge is insufficient to overcome this inherent bottleneck.

To break through the bottleneck, recent studies [8, 10, 11] have explored integrating external knowledge from Large Language Models (LLMs) to enhance their understanding and alignment abilities. However, such approaches typically require prohibitively expensive fine-tuning or retraining of offline models [10, 12]. Some other attempts [8, 13, 14] have proposed interactive cross-modal retrieval frameworks that incorporate LLMs and Vision-Language Models (VLMs) to facilitate more fine-grained understanding and alignment, thereby iteratively evolving the retrieval performance. Although these methods have demonstrated remarkable effectiveness in image-text matching [8] and video-text retrieval [15, 16], they still face two tricky challenges in the T3SR setting, as shown in Figure 1. *Firstly*, since the scale and complexity of 3D scenes, these methods lack holistic perspectives beyond a localized focus during interaction, *e.g.*, the LLMs tend to focus on the salient objects in the scene and ignore the fine-grained details at a scene-level perspective, limiting the depth and breadth of LLM interaction, as demonstrated in Table 1. *Secondly*, existing retrieval models exhibit limited generalization ability to biased text domains, limiting their effectiveness in handling realistic interaction texts that exhibit domain gaps.

To address the aforementioned challenges, this paper proposes a novel **I**nteractive Text-3**D** Sc**e**ne Retriev**al** method (IDeal) to conduct continuous interaction between the T3SR models and external users (*e.g.*, LLMs), achieving the active alignment between text queries and 3D scenes, as depicted in Figure 1. Our IDeal consists of two components: an Interactive Retrieval Refinement framework (IRR) and an Interaction Adaptation Tuning strategy (IAT), as illustrated in Figure 2. More specifically, IRR coordinates three specialized agents (*i.e.*, *questioner*, *answerer*, and *retriever*) to perform multi-round interaction. First, the *questioner* adaptively determines whether to continue probing object details or to pursue divergence by exploring the broader scene, based on the assessment of the current round's description. Based on this, it continuously formulates context-relevant questions to the *answerer*. After receiving responses, the *retriever* iteratively integrates information at both the feature and semantic levels, facilitating comprehensive scene-level understanding for progressively precise re-rankings. To mitigate the domain shift between training queries and interactive texts, IAT proposes adapting the *retriever* toward the interaction text domain. Specifically, IAT leverages LLMs to generate more realistic augmented texts that closely resemble the interaction text domain. Subsequently, IAT robustly mitigates the discriminability and diversity theoretical risks in the features of the augmented texts for domain gap bridging, thereby ensuring an unbiased understanding of the interaction texts by the *retriever*. The contributions of this paper are as follows:

- We propose a novel Interactive Text-3D Scene Retrieval Method (IDeal), which actively enhances alignment between text queries and 3D scenes through ongoing interactions.
- An Interactive Retrieval Refinement framework (IRR) is presented to enable a deep interaction for comprehensive scene exploration, leading to progressively improved retrieval.
- An Interaction Adaptation Tuning strategy (IAT) is proposed, which facilitates the transfer of the retriever to the interaction text domain, promoting improved interaction.
- We conduct extensive comparison experiments on text-3D scene datasets. Our IDeal remarkably outperforms the existing methods, demonstrating its superiority.

2 Related Work

Cross-Modal Retrieval. Cross-Modal Retrieval (CMR) [17, 18, 19, 20, 21, 22] aims to match corresponding results across modalities for a given query, bridging the gap caused by modal heterogeneity. In recent years, CMR has garnered significant attention in fields such as Image-Text Retrieval [14, 23, 24, 25], Video-Text Retrieval [26, 27], 2D-3D Retrieval [28, 29], Pointcloud-Text Matching [6]. The primary challenge of CMR lies in effectively aligning multimodal data. To address this issue, most existing works could be broadly categorized into two groups: 1) Coarse-grained retrieval [30, 31, 32] directly maps multimodal data into a shared space, aiming for a more straightforward and computationally efficient alignment. 2) Fine-grained retrieval [33, 34] seeks to establish local associations between fine-grained features across modalities (*e.g.*, regions in images, words in texts). These local associations are then progressively integrated to form precise cross-modal correspondences. This paper focuses on a more challenging CMR task, *i.e.*, Text-3D Scene Retrieval, involving obscure spatial cues and sophisticated 3D scenes (including issues such as viewpoints and occlusions [35, 36]). Although prior work [6] performs well with comprehensive descriptions, it struggles with online and blurry queries. To this end, we propose an Interactive T3SR solution that iteratively incorporates online feedback to achieve more precise and practical scene retrieval.

Interactive Learning. Unlike traditional learning paradigms [6, 37], interactive learning emphasizes the continuous improvement of a model's behavior through ongoing interactions with the environment or users. Specifically, several pioneering works [38, 39] leverage simple forms of user feedback (*e.g.*, preferred sample selection and relevance scoring) to iteratively achieve improved training quality or better satisfy user-specific requirements during testing. With the development of Large Language Models (LLMs), other studies [8, 40, 41] have begun exploring question-answering interactions through free-form text dialogue, closely replicating natural human communication. For example, several methods leverage iterative interactions to continuously refine the retrieval query for better reranking. Recently, more advanced methods such as PlugIR [13], MERLIN [16], ICL [42], and LLaVA-ReID [43] have integrated LLMs for context-aware question generation, mining more visual details. However, these methods cannot be effectively generalized to T3SR due to the differences in tasks and data domains shown in Figure 1. In this paper, we develop an interactive framework tailored for T3SR to help the offline models adapt to complex scene perception.

3 Method

3.1 Problem Formulation

Given a text query set $\mathcal{T} = \{t_i\}_{i=1}^{n_t}$ and a 3D scene gallery $\mathcal{C} = \{c_j\}_{j=1}^{n_c}$, where t_i and c_j represent i-th text and j-th scene, $n_t = |\mathcal{T}|$ and $n_c = |\mathcal{C}|$ means the sample number, and $|\cdot|$ denotes the volume of set. The purpose of T3SR is to use the text query to match the ideal 3D scenes from the gallery, where there exists correspondence $y_{i,j} \in \{0, 1\}$, indicating whether the points are matched (*i.e.*, $y_{ij} = 1$), or unmatched (*i.e.*, $y_{ij} = 0$). Existing methods [6, 37] typically train an offline model to achieve encoding of multimodal data, followed by meticulous single-turn retrieval. They assume that user-provided text queries are information complete, overlooking the practical fact that queries are often partial, ambiguous, or even exhibit domain shift.

To address these issues, an interactive Text-3D scene retrieval method, *i.e.*, IDeal, is proposed to bridge the interaction between the retrieval models and external agents, progressively overcoming the aforementioned query limitations and achieving improved alignment between texts and 3D scenes. More specifically, IDeal asks question q_i^l ($l \in \{1, \cdots, r\}$) about i-th sample based on the users' previous response a_i^{l-1} ($a_i^0 = t_i$), where r is the upper limit of rounds. Subsequently, the external agents recall details and answer a_i^l of the target 3D scene c_j ($y_{ij} = 1$), forming a dialogue context $\mathcal{D}_i = \{a_i^0, (q_i^1, a_i^1), \cdots (q_i^r, a_i^r)\}$ composed of question-answer pairs $(q_i^{\cdot}, a_i^{\cdot})$. Both the j-round response text and 3D scenes are projected into a shared feature space by a trained retrieval model, which can be: $\boldsymbol{u}_i^j = f_r(a_i^j; \boldsymbol{\theta})$ and $\boldsymbol{v}_i = f_r(c_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the learnable parameters.

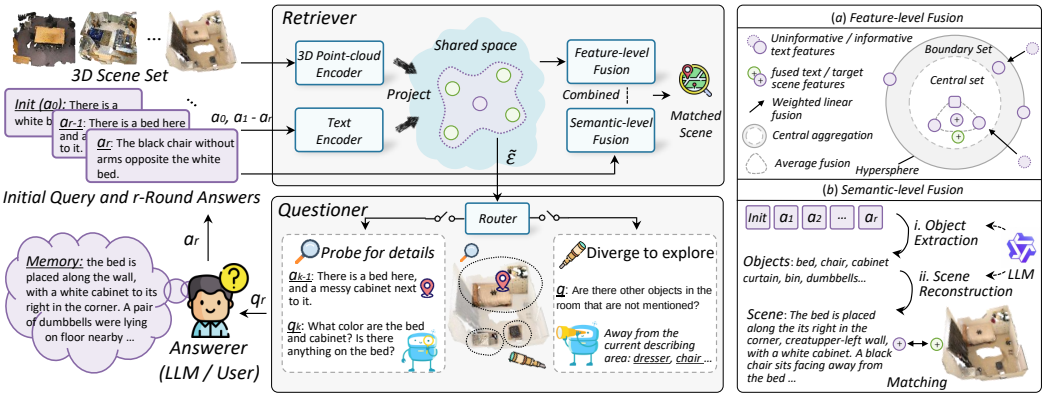


Figure 2: (Left) Pipeline of the proposed IDeal. A *questioner* employs a router to adaptively pose either probe or divergent questions, which require responses from an *answerer*. After receiving iterative responses, a *retriever* projects the multimodal data into a shared feature space, performing feature-/semantic-level fusion (Right) to enable progressively precise scene retrieval.

3.2 Interactive Retrieval Refinement Framework

In this section, we introduce the Interactive Retrieval Refinement framework (IRR), which coordinates three agents (*i.e.*, *questioner*, *answerer*, and *retriever*) to enable iterative interaction. In the following sections, we will elaborate on them by introducing their interaction process.

3.2.1 Adaptive Questioning

To achieve a deep interaction for T3SR, we present an adaptive *questioner*, which enables pertinent questioning to the *answerer* for both detailed and comprehensive exploration of complex 3D scenes. Firstly, a question router is employed, which determines the focus of questioning based on the feature distribution in the shared space. More specifically, the router assesses whether the previous-round description is informative by computing a *Cross-modal Affinity Entropy* ($\mathcal{E}$), formulated as:

$$\mathcal{E}(\boldsymbol{u}_i^{r-1}) = -\sum_{j \in N_k(\boldsymbol{u}_i^{r-1})} p(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_j) \log p(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_j), \tag{1}$$

where $N_k(\boldsymbol{u}_i^{r-1})$ means the k-nearest-neighbor index of $\boldsymbol{u}_i^{r-1}$, and affinity probability $p(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_j)$ is formulated as:

$$p(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_j) = \frac{\exp(\mathcal{S}(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_j)/\tau)}{\sum_{l \in N_k(\boldsymbol{u}_i^{r-1})} \exp(\mathcal{S}(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_l)/\tau)}, \tag{2}$$

where $\mathcal{S}$ represents the computation of similarity between features, and τ is a temperature parameter.

However, the distribution of scene features extracted by the trained retrieval models is fixed and inherently non-uniform, with regions of over-density and under-density introducing a structural density bias [44]. To mitigate the bias, we introduce a *Density Compensated Factor* for each scene feature, which is formulated as: $\rho(\boldsymbol{v}_i) = \frac{1}{(1/k)\sum_{j \in N_k(\boldsymbol{v}_j)} \mathcal{D}(\boldsymbol{v}_i, \boldsymbol{v}_j) + \epsilon}$, where $\mathcal{D}$ represents the distance calculation and ϵ is a minimal constant for numerical stability. Based on this, we try to approximately correct the original similarity score $\mathcal{S}(\boldsymbol{u}_i^{r-1}, \boldsymbol{v}_j)$ by incorporating the *Density Compensated Factor* $\rho(\boldsymbol{v}_i)$, which could be written as: $\tilde{\mathcal{S}}(\boldsymbol{u}_i^{(r-1)}, \boldsymbol{v}_j) = \mathcal{S}(\boldsymbol{u}_i^{(r-1)}, \boldsymbol{v}_j)/\sqrt{\rho(\boldsymbol{v}_j)}$. Subsequently, this corrected similarity is brought back into Equations (1) and (2) to obtain a *Density Compensated Affinity Entropy*, denoted as $\tilde{\mathcal{E}}$. This process approximates a Bayesian Correction leveraging prior density estimation[2], mitigating the impact of the inherent bias in scene feature distribution.

Leveraging the fairer metric $\tilde{\mathcal{E}}$, our *questioner* categorize descriptions with $\tilde{\mathcal{E}} > \beta$ as uninformative, prompting an LLM to generate *questions for detail probe* (*i.e.*, $\mathcal{Q}_1$) for attribute and spatial relationship detail refinements within the described area. Conversely, when $\tilde{\mathcal{E}} \leq \beta$, the descriptions are considered informative, triggering the adopting of *questions for diverge exploration* (*i.e.*, $\mathcal{Q}_2$) that inquire about object arrangements not previously discussed in the dialogue.

[2]Please refer to our Supplemental Material for further discussion.

3.2.2 Iterative Retrieval

After completing the questioning, an LLM is employed to simulate the external user acting as an *answerer* to answer the questions, following existing interactive approaches [8, 13]. It receives multi-round questions and provides responses based on its memory. In this paper, we adopt text modality to simulate the memory, which better approximates how humans recall information in mind.

Upon receiving the response descriptions from the *answerer*, we construct a *retriever* that can be seamlessly integrated with existing cross-modal models [37, 45], enabling iterative scene retrieval. It can obtain the scene retrieval predictions for any given text within the shared feature space. Specifically, given a text feature $\boldsymbol{u}_i$, the prediction can be formulated as:

$$\hat{\boldsymbol{p}}(\boldsymbol{u}_i) = [\hat{p}(\boldsymbol{u}_i, \boldsymbol{v}_1), \hat{p}(\boldsymbol{u}_i, \boldsymbol{v}_2), \ldots, \hat{p}(\boldsymbol{u}_i, \boldsymbol{v}_{n_c})]^\top, \tag{3}$$

where $\hat{p}(\boldsymbol{u}_i, \boldsymbol{v}_j) = \exp(\mathcal{S}(\boldsymbol{u}_i, \boldsymbol{v}_j)) / \sum_{l=1}^{n_c} \exp(\mathcal{S}(\boldsymbol{u}_i, \boldsymbol{v}_l))$ denotes the probability that the i-th text retrieves the j-th scene. Accordingly, our *retriever* first utilizes the initial query to compute an *initial retrieval prediction* $\hat{\boldsymbol{p}}_1(\boldsymbol{u}_i)$. Subsequently, the feature-level and semantic-level fusion of interactive responses is conducted to achieve more precise scene retrieval.

For the fusion of interactive response feature, on one hand, considering responses to $\mathcal{Q}_1$ are refinements of the previous-round descriptions, we apply a weighted linear fusion to incorporate supplementary information. This strategy enables the preservation of core semantic cues from previous rounds while emphasizing newly introduced details: $\boldsymbol{u}_i^j = \alpha \boldsymbol{u}_i^j + (1-\alpha)\boldsymbol{u}_i^{j-1}$, where $q_i^j \in \mathcal{Q}_1$, α is a trade-off weight. On the other hand, benefiting from $\mathcal{Q}_2$, the other response features and the aforementioned fused features capture variations across different regions of the scene. To fuse them into a comprehensive feature, inspired by [46], we model their distribution around the target scene by encapsulating them within a minimum enclosing hypersphere:

$$(\boldsymbol{o}_i^*, R_i^*) = \arg\min_{\boldsymbol{o}_i, R_i} \left\{ R_i : \boldsymbol{u}_i^j \boldsymbol{o}_i^\top \leq R_i, \forall j \right\}, \tag{4}$$

where $\boldsymbol{o}_i^*$ and R_i^* are the center and radius of the hypersphere, respectively. Based on this, features near the hypersphere boundary are grouped into a boundary set $\mathcal{U}_i^1$, while the remainder constitute the central set $\mathcal{U}_i^2$. To balance fusion robustness and feature discrimination, potentially noisy boundary features in $\mathcal{U}_i^1$ are aggregated at the hypersphere center and averaged with cleaner features in $\mathcal{U}_i^2$ to yield the final fused response feature: $\bar{\boldsymbol{u}}_i = \frac{1}{2}\left(\boldsymbol{o}_i^* + \frac{1}{|\mathcal{U}_i^2|}\sum_{\boldsymbol{u}_i^j \in \mathcal{U}_i^2} \boldsymbol{u}_i^j\right)$. This fused feature is then input into Equation (3) to obtain an *interactive feature prediction* $\hat{\boldsymbol{p}}_2(\bar{\boldsymbol{u}}_i)$.

However, the aforementioned feature-level fusion can not fully capture the holistic semantics of the responses. To address this limitation, we leverage an LLM to reconstruct the 3D scene from all responses in textual space. More specifically, inspired by Chain-of-Thought (CoT) [47], we decompose this process into object extraction and scene reconstruction for a more stable and comprehensive scene summary. The LLM first identifies the scene objects across multi-round responses and then summarizes an object-centric scene reconstruction text. Finally, the texts are encoded into feature $\boldsymbol{s}_i$, from which a *interactive semantic prediction* $\hat{\boldsymbol{p}}_3(\boldsymbol{s}_i)$ is computed using Equation (3).

Finally, the initial and interactive predictions are combined through weighted fusion to obtain the final scene retrieval prediction as follows:

$$\hat{\boldsymbol{p}}_c(\boldsymbol{u}_i) = \lambda_1 \hat{\boldsymbol{p}}_1(\boldsymbol{u}_i) + \lambda_2 \hat{\boldsymbol{p}}_2(\bar{\boldsymbol{u}}_i) + \lambda_3 \hat{\boldsymbol{p}}_3(\boldsymbol{s}_i), \tag{5}$$

where $\hat{\boldsymbol{p}}_c(\boldsymbol{u}_i)$ is the final retrieval prediction, λ_1, λ_2, and λ_3 are trade-off parameters. Benefiting from the adaptive questioning and the comprehensive retrieval information fusion, our IDeal can alleviate the limitations of initial queries through progressive interaction.

3.3 Interaction Adaptation Tuning

Although IRR can exploit interactions to promote retrieval quality, the limited text domain of the *retriever* remains a bottleneck that restricts further performance improvements. To overcome this limitation, we propose an Interaction Adaptation Tuning strategy (IAT), which enhances texts to approximate the domain of interaction texts.

We begin by integrating information and descriptions of the same scenes from the training data to construct simulated memory for text augmentation. Following IRR, we first provide a training-data-based answerer (*i.e.*, an LLM) with the constructed memory and initial queries. We then simulate the

IRR interaction process by iteratively posing a fixed number of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ questions. The response descriptions yield augmented texts that closely approximate the interaction scenario.

After obtaining the enriched augmented texts, inspired by the contrastive domain adaptation paradigm [48, 49], we try to minimize the theoretical risk $\mathcal{R}(\boldsymbol{\theta})$ associated with our *retriever* among the augmented text features, as formulated below. It facilitates model adaptation of the *retriever* to the augmented text domain without requiring access to its implementation details.

$$\mathcal{R}(\boldsymbol{\theta}) = \mathcal{R}_{dis}(\boldsymbol{\theta}) + \mathcal{R}_{div}(\boldsymbol{\theta}) = \mathbb{E}_{\tilde{\mathcal{U}}}\left[\left(-\mathbb{E}_{\tilde{\mathcal{U}}^+}\left\{\mathcal{S}(\tilde{\boldsymbol{u}}_i^+, \tilde{\boldsymbol{u}}_i)\right\}\right) + \left(\mathbb{E}_{\tilde{\mathcal{U}}^-}\left\{\mathcal{S}(\tilde{\boldsymbol{u}}_i^-, \tilde{\boldsymbol{u}}_i)\right\}\right)\right], \tag{6}$$

where the two components $\mathcal{R}_{dis}(\boldsymbol{\theta})$ and $\mathcal{R}_{div}(\boldsymbol{\theta})$ respectively reflect discriminability and diversity risks, $\mathbb{E}$ denotes expectation, $\mathbb{E}_{\tilde{\mathcal{U}}}$ is taken with respect to the distribution for the target domain features (*i.e.*, augmented text features $\tilde{\mathcal{U}} = \{\tilde{\boldsymbol{u}}_i\}_{i=1}^{n_t}$), and $\tilde{\mathcal{U}}^+$ and $\tilde{\mathcal{U}}^-$ is the distribution for the corresponding positive features $\tilde{\boldsymbol{u}}_i^+$ and negative features $\tilde{\boldsymbol{u}}_i^-$, respectively.

On the one hand, minimizing the discriminability risk $\mathcal{R}_{dis}(\boldsymbol{\theta})$ requires encouraging the augmented text features to align closely with those belonging to the same scenes. However, the scale and complexity of scenes often lead to substantial variability even among features corresponding to the same scenes. This introduces significant uncertainty in the selection of positive samples, complicating the risk optimization process.[3] To handle this, we adopt the corresponding 3D scene features as substitutes for the text features to construct positive pairs. This is based on the assumption that the scene features encoded by the well-trained model are more stably located near the center of the corresponding description distribution. Finally, we mitigate the aforementioned discriminability risk $\mathcal{R}_{dis}(\boldsymbol{\theta})$ by minimizing a negative log-based proxy loss term $\mathcal{L}_{dis}$, which could be written as follows:

$$\mathcal{L}_{dis} = -\sum_{i=1}^{b}\sum_{j=1}^{n_c} y_{ij} \log \mathcal{S}(\tilde{\boldsymbol{u}}_i, \boldsymbol{v}_j), \tag{7}$$

where b is the size of the mini-batch. *On the other hand*, motivated by [50], we attempt to approximate the minimization of the divergence risk $\mathcal{R}_{div}(\boldsymbol{\theta})$ by minimizing its upper bound, *i.e.*,

$$\sup\left(\mathcal{R}_{div}(\boldsymbol{\theta})\right) \sim \left\{\mathbb{E}_{\tilde{\boldsymbol{u}}^- \sim \tilde{\mathcal{U}}^-}\left(\mathcal{S}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^-)\right); \mathbb{V}_{\tilde{\boldsymbol{u}}^- \sim \tilde{\mathcal{U}}^-}\left(\mathcal{S}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^-)\right)\right\}, \tag{8}$$

where $\sup(\cdot)$ means the upper bound and $\mathbb{V}$ is the variance. We can obviously see that the divergence risk is affected by the mean and variance of the selected negative samples. Yet existing methods [48] usually treat others within the same mini-batch as negative samples for contrastive learning, thereby minimizing the expectation term. Due to the stochasticity of mini-batch sampling, similar samples may be mistakenly chosen as negatives, which increases the variance of negative samples, enlarging the upper bound of the divergence risk.

To tackle it, we propose a weighted complementary contrastive loss as a surrogate objective to achieve divergence risk optimization more robustly, which can be formulated as:

$$\mathcal{L}_{div} = \sum_{i=1}^{b}\sum_{j\neq i}^{b} \underbrace{\exp\left(-\max\left(0, \mathcal{S}(\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{u}}_j) - \gamma\right)\right)}_{\text{Weighting term}} \underbrace{\log\left(1 - \mathcal{S}(\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{u}}_j)\right)}_{\text{Complementary contrastive term}}, \tag{9}$$

where γ is a threshold, above which samples are assigned lower weights. Minimizing the complementary contrastive term optimizes the expectation over negative pairs, while the weighting component can mitigate the impact of high-variance false-negative samples.

Finally, we combine both terms to obtain our loss for domain adaptation tuning, as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{dis} + (1 - \lambda)\mathcal{L}_{div}, \tag{10}$$

where λ is a hyperparameter to control the contribution of each component. Minimizing this proxy loss facilitates the reduction of domain adaptation risk, thereby enabling the *retriever* to better adapt to the domain of interaction text.

[3]The analysis can be found in our Supplemental Material.

Table 1: Performance comparison on ScanRefer, Nr3D, and Sr3D in terms of R@1, R@5, R@10, and their sum (Rsum). † denotes the use of coarse-grained descriptions as memory.

Methods	ScanRefer				Nr3D				Sr3D			
	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	Rsum
w/ coarse-grained descriptions:												
VSE∞ (CVPR'21)	9.7	33.1	50.2	93.0	5.8	21.5	32.5	59.8	5.5	18.9	27.4	51.8
CHAN (CVPR'23)	9.4	32.3	52.1	93.8	7.5	15.0	32.1	54.6	5.4	21.3	34.8	61.5
HREM (CVPR'23)	10.2	34.0	51.4	95.6	7.3	18.1	31.9	57.3	5.4	21.9	33.6	60.9
CRCL (NeurIPS'23)	10.3	32.4	49.8	92.5	8.1	22.5	33.2	63.8	4.9	19.5	31.9	56.3
RoMa (TMM'25)	11.4	34.8	54.4	100.6	6.5	24.8	37.6	68.9	7.8	27.3	39.3	74.4
IDeal	**16.0**	**42.7**	**59.8**	**118.5**	**11.7**	**34.8**	**50.4**	**96.9**	**10.3**	**30.2**	**48.5**	**89.0**
w/ fine-grained descriptions:												
ChatIR† (NeurIPS'23)	21.8	55.4	73.1	150.3	15.6	40.3	58.1	114.0	12.1	35.9	50.3	98.3
Rewrite† (ICMR'24)	17.4	47.0	63.7	128.1	17.1	31.4	45.8	94.3	12.4	28.1	40.4	80.9
MERLIN† (EMNLP'24)	31.1	68.8	83.8	183.7	21.8	55.0	71.8	148.6	14.2	42.0	60.3	116.5
BASELINE: IR†	29.9	68.0	83.3	181.2	18.0	51.6	60.2	129.8	14.6	39.2	55.3	109.1
BASELINE SUM†	34.4	69.5	85.1	189.0	22.5	55.2	67.5	145.2	16.4	41.5	62.1	120.0
IDeal†	**37.8**	**71.8**	**86.4**	**196.0**	**26.4**	**62.7**	**78.7**	**167.8**	**20.2**	**43.1**	**63.1**	**126.4**

4 Experiments

4.1 Experimental Setting

Datasets, baselines, and evaluation metrics: We adopt the ScanNet 3D scene set along with several description sets (*i.e.*, ScanRefer [51], Nr3D [52], Sr3D [52], and SceneDepict-3D2T [6]) to conduct experiments, where ScanRefer, Nr3D, and Sr3D are employed as query sets, and SceneDepict-3D2T is employed to simulate fine-grained memory. To verify the superiority of our IDeal, we introduce eleven comparative baseline methods: five *conventional offline cross-modal matching methods* (*i.e.*, VSE∞ [45], CHAN [53], HREM [54], CRCL [37], and RoMa [6]), three *interactive cross-modal retrieval methods* (*i.e.*, ChatIR [8], Rewrite [41], and MERLIN [16]), and two *additional strong interactive baselines* (IR and SUM). More specifically, the Iterative Reranking (IR) involves multi-round interaction, where the results are iteratively re-ranked based on the response of each round. The Summary reranking (SUM) also involves interaction, but ultimately aggregates all answers into a comprehensive description for matching.

In addition, we follow [55, 56] to report R@1, R@5, R@10, and their summation (Rsum) as the evaluation metrics. Due to the space limitation, more details of datasets, prompts, and additional experiments are provided in the Supplemental Material.

Table 2: Performance comparison on ScanRefer and Nr3D in terms of R@1, R@5, R@10, and their sum. *+IDeal* indicates plugging the model into our IDeal. † denotes the use of fine-grained descriptions as memory.

Methods	ScanRefer				Nr3D			
	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	Rsum
VSE∞	9.7	33.1	50.2	93.0	5.8	21.5	32.5	59.8
+IDeal	13.3	38.9	57.6	109.8	8.7	27.5	42.1	78.3
VSE∞†	14.9	42.3	61.5	118.7	16.4	47.5	55.2	119.1
+IDeal†	35.8	70.6	85.0	191.4	21.2	52.1	68.4	141.7
CRCL	10.3	32.4	49.8	92.5	8.1	22.5	33.2	63.8
+IDeal	13.4	35.5	56.1	105.0	7.4	25.4	38.3	71.1
CRCL†	17.5	45.1	58.3	120.9	13.4	44.5	51.5	109.4
+IDeal†	31.7	66.9	83.5	182.1	15.8	50.4	64.4	130.6
RoMa	9.7	33.1	50.2	93.0	8.3	27.9	37.2	73.4
+IDeal	16.0	42.7	59.8	118.5	11.7	34.8	50.4	96.9
RoMa†	16.7	44.8	61.6	123.1	17.4	48.5	57.5	123.4
+IDeal†	37.8	71.8	86.4	196.0	25.4	60.7	75.7	161.8

Implementation details: All methods are implemented in PyTorch and carried out on GeForce RTX 3090 GPUs. We adhere to the experimental settings of [6] for all method implementations. We adopt widely-used DGCNN [57] and BERT [58] to obtain fine-grained features for 3D point clouds and texts, respectively. To implement interaction, we explore two approaches to constructing memory: **1)** *Coarse-grained description*: We leverage an LLM to generate rich expansions of queries, serving as memory without introducing any additional information leakage. **2)** *Fine-grained description*: In line with existing interactive methods [43, 13], we simulate the user's memory in real-world scenarios using fine-grained scene descriptions, albeit with access to partial additional information. In our experiments, we utilize Qwen-7B-Instruct [59] as our investigated LLM for the interaction experiments.

4.2 Comparison on Text-3D Scene Retrieval

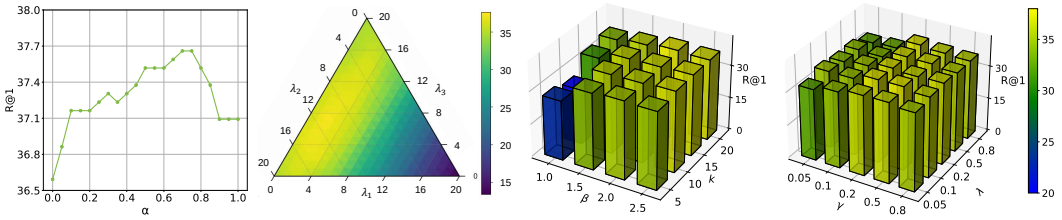
Table 1 presents a comparison between our IDeal and conventional and interactive cross-modal matching methods under two memory settings. Table 2 further demonstrates the performance gains brought by integrating our interactive framework into conventional single-round methods. These results could yield the following observations: **1)** Even without additional fine-grained information, our IDeal achieves competitive performance, highlighting its ability to uncover complementary information implicitly embedded in the queries, gradually alleviating inherent query limitations. **2)** Compared to existing interactive methods, our IDeal also achieves superior performance with access to fine-grained descriptions. This demonstrates that our interactive questioning and retrieval strategies enable an ongoing understanding of user requirements and support a comprehensive interpretation of complex scenes. **3)** Our IDeal can be seamlessly integrated into conventional cross-modal retrieval methods and yields substantial performance gains under both memory settings. This suggests that, beyond equipping offline models with interactive capabilities, IDeal empowers them to more effectively comprehend complex descriptions through interaction.

4.3 Ablation Study

In this section, we conduct an ablation study to evaluate the contribution of each proposed component to our IDeal. Specifically, we first ablate the router in the *questioner*, restricting it to ask either $\mathcal{Q}_1$ or $\mathcal{Q}_2$ continuously. In addition, we remove each of the three retrieval prediction strategies, and we examine removing CoT prompting in the reconstruction in the proposed *retriever*. Finally, we investigate the effect of not using the IAT strategy for domain adaptation and sequentially ablate its two loss terms. The results in Table 3 lead to the following observation: **1)** Removing or replacing any component from IDeal results in performance degradation, highlighting the contribution of each component. Specifically, the adaptive questions generated by our *questioner* facilitate a meticulous and comprehensive understanding of scenes. The various feature aggregation strategies in the *retriever* contribute to precise scene matching. **2)** Removing or substituting IAT components consistently leads to performance degradation, underscoring the necessity of text domain alignment and adaptation risk minimization in our IAT.

Table 3: Ablation studies for components of our IDeal on ScanRefer. RSum is the sum of R@1, R@5, R@10. *w/o* stands for without use.

Configurations		ScanRefer R@1	R@5	R@10	Rsum
Questioner	*w/o* $\mathcal{Q}_1$	36.0	71.2	86.7	193.9
	w/o $\mathcal{Q}_2$	26.3	61.5	77.3	165.1
Retriever	*w/o* $\hat{\boldsymbol{p}}_1(\boldsymbol{u}_i)$	35.2	70.1	86.5	191.8
	w/o $\hat{\boldsymbol{p}}_2(\bar{\boldsymbol{u}}_i)$	28.1	63.3	80.6	172.0
	w/o $\hat{\boldsymbol{p}}_3(\boldsymbol{s}_i)$	31.8	67.8	84.2	183.8
	w/o CoT	35.7	71.5	85.9	193.1
Adaptation	*w/o* IAT	16.6	48.4	64.4	129.4
	w/o $\mathcal{L}_{dis}$	34.9	69.5	84.4	188.8
	w/o $\mathcal{L}_{div}$	35.1	69.4	84.1	191.7
Full	IDeal	37.8	71.8	86.4	196.0

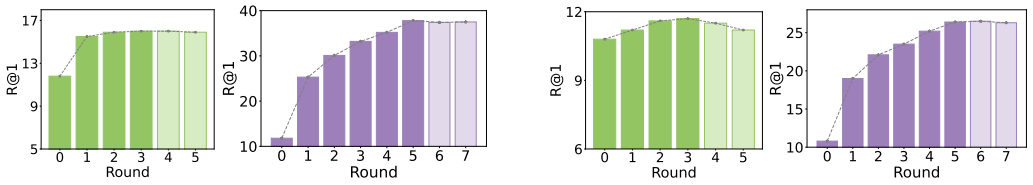


(a) α in *retriever*. (b) λ_1, λ_2, λ_3 in *retriever*. (c) k, β in *questioner*. (d) λ, γ in IAT.

Figure 3: PTM performance in terms of R@1 versus different values of the parameters of our IDeal on ScanRefer. (a) and (b) display α, λ_1, λ_2, and λ_3 in our *retriever*. (c) shows k and β in our *questioner*. (d) shows λ and γ in IAT.

4.4 Parameter Analysis

To evaluate the sensitivity of our IDeal to different hyperparameter settings, we plot the retrieval performance versus different values on ScanRefer, as shown in Figure 3. The experimental results lead to the following observation: **1)** For our *retriever*, tuning greater weights to interactive and reconstruction predictions helps achieve a well-balanced trade-off that fully leverages the interactive responses. Additionally, a higher feature fusion weight (*e.g.*, $\alpha = 0.75$) represents a emphasis on



(a) ScanRefer (b) Nr3D

Figure 4: Performance (R@1) versus rounds on two datasets. Round 0 indicates the setting without interaction. The green and purple bars represent the cases with coarse-grained and fine-grained memory descriptions, respectively. The *lighter* bars indicate no performance gain.

the integration of discriminative features from the refined descriptions, leading to more effective interaction feature fusion. **2)** For our proposed *questioner*, using reasonable and moderate settings of k and β (e.g., $k = 20$, $\beta = 2.0$) enables accurate identification of informative descriptions, thereby supporting reasonable decisions on question types in the next round. **3)** During domain adaptation tuning, a relatively wide range of λ and γ values in IAT (*i.e.*, $\lambda \in [0.2, 0.5]$ and $\gamma \in [0.1, 0.8]$) ensures effective contrastive adaptation and mitigates the impact of false negatives.

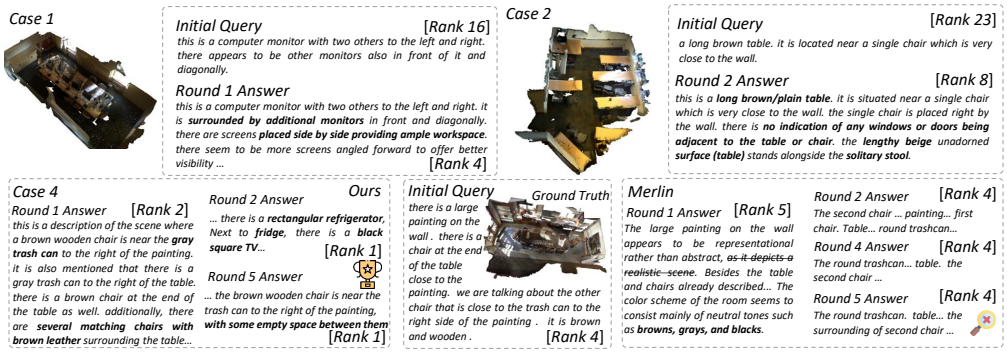


Figure 5: Case illustrations of interactive cross-modal scene matching process of IDeal on ScanRefer. Cases 1–2 and 3–4 are with coarse-grained and fine-grained memory descriptions, respectively.

4.5 Visualization Analysis

To provide a comprehensive analysis of IDeal, we conduct a series of visualization experiments. Specifically, we first present the changes in retrieval performance of IDeal across multiple rounds of interaction, as shown in Figure 4, to analyze the incremental gains brought by each interaction. In addition, we visualize several representative cases, as illustrated in Figure 5. The observations can be drawn from the results: 1) Interaction consistently improves performance over the first several rounds (five rounds with fine-grained and three rounds with coarse-grained descriptions). Although redundant interactions may inevitably cause performance to saturate or even slightly degrade due to LLM hallucinations or excessively long texts, these results indicate that a reasonable number of interaction steps can effectively enhance the query and improve retrieval performance. 2) Under the setting with coarse-grained memory texts, IDeal can infer and decompose object attributes and relationships within the queries through interaction, leading to improved retrieval performance. Moreover, with the integration of fine-grained memory, IDeal leverages targeted and switchable questioning to elicit informative responses, continually improving retrieval precision. In contrast, MERLIN [16] frequently generates redundant descriptions confined to local details of scenes.

5 Conclusion

In this paper, we propose a novel **I**nteractive Text-3**D** Scene Retrieval Method, namely IDeal, to address the Text-3D Scene Retrieval (T3SR). Our IDeal integrates two components: the Interactive Retrieval Refinement Framework (IRR) and the Interaction Adaptation Tuning strategy (IAT). Specifically, IRR continuously conducts adaptive questioning and comprehensive response fusion,

enabling holistic exploration of 3D scenes for more precise retrieval. IAT performs contrastive domain adaptation for the retriever toward realistic texts, overcoming the performance bottleneck during interaction. Extensive experiments demonstrate the superiority of our IDeal in T3SR task.

Limitations and Potential Impact Statement: Although our work has taken the initial step forward in interactive T3SR, there are some limitations and potential impacts that should be acknowledged. First, the performance of the methods is relatively low. Second, we employ LLMs, and more stable and unbiased LLMs and interaction approaches merit further exploration in the future.

Acknowledgments

This work was supported in part by NSFC under Grant 62472295 and 62372315; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202403; in part by Sichuan Science and Technology Planning Project under Grant 24NSFTD0130, 2024ZDZX0004, 2024NSFTD0049, and in part by the Chengdu Science and Technology Project under Grant 2023-XT00-00004-GX.

References

[1] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[2] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.

[3] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024.

[4] Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22314–22324, 2024.

[5] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[6] Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng, and Peng Hu. Pointcloud-text matching: Benchmark dataset and baseline. *IEEE Transactions on Multimedia*, 2025.

[7] Jiaqi Chen, Daniel Barath, Iro Armeni, Marc Pollefeys, and Hermann Blum. "where am i?" scene retrieval with language. In *European Conference on Computer Vision*, pages 201–220. Springer, 2024.

[8] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36:61437–61449, 2023.

[9] Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Xiting Liu, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. *arXiv preprint arXiv:2410.15624*, 2024.

[10] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *International Conference on Machine Learning*, pages 27890–27902. PMLR, 2024.

[11] Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11091–11101, 2023.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[13] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–809, 2024.

[14] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26679–26688, 2024.

[15] Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 356–365, 2022.

[16] Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, and Nojun Kwak. Merlin: Multimodal embedding refinement via llm-based iterative navigation for text-video retrieval-rerank pipeline. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 547–562, 2024.

[17] Junsheng Wang, Tiantian Gong, Zhixiong Zeng, Changchang Sun, and Yan Yan. C3cmr: Cross-modality cross-instance contrastive learning for cross-media retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4300–4308, 2022.

[18] Zhenqiu Shu, Yibing Bai, Kailing Yong, and Zhengtao Yu. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*, 2024.

[19] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 852–861, 2024.

[20] Chao Su, Huiming Zheng, Dezhong Peng, and Xu Wang. Dica: Disambiguated contrastive alignment for cross-modal retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20610–20618, 2025.

[21] Ruitao Pu, Yang Qin, Xiaomin Song, Dezhong Peng, Zhenwen Ren, and Yuan Sun. She: Streaming-media hashing retrieval. In *Forty-second International Conference on Machine Learning*.

[22] Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15184–15192, 2024.

[23] Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14354–14363, 2024.

[24] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022.

[25] Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. Robust self-paced hashing for cross-modal retrieval with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19969–19977, 2025.

[26] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. *Advances in neural information processing systems*, 35:38655–38666, 2022.

[27] Weike Jin, Zhou Zhao, Pengcheng Zhang, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. Hierarchical cross-modal graph consistency learning for video-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*, pages 1114–1124, 2021.

[28] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023.

[29] Yongxiang Li, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Romo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*, 2024.

[30] Hongguang Zhu, Chunjie Zhang, Yunchao Wei, Shujuan Huang, and Yao Zhao. Esa: External space attention aggregation for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[31] Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Xijun Xue. Multi-view visual semantic embedding. In *IJCAI*, volume 2, page 7, 2022.

[32] Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:824–836, 2023.

[33] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226, 2021.

[34] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23422–23431, 2023.

[35] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022.

[36] Changkun Liu, Shuai Chen, Yash Sanjay Bhalgat, Siyan Hu, Ming Cheng, Zirui Wang, Victor Adrian Prisacariu, and Tristan Braud. Gs-cpr: Efficient camera pose refinement via 3d gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2024.

[37] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Ryoya Nara, Yu-Chieh Lin, Yuji Nozawa, Youyang Ng, Goh Itoh, Osamu Torii, and Yusuke Matsui. Revisiting relevance feedback for clip-based interactive image retrieval. *arXiv preprint arXiv:2404.16398*, 2024.

[39] Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems*, 37:42369–42387, 2025.

[40] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer, 2024.

[41] Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 978–987, 2024.

[42] Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14390–14399, 2025.

[43] Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025.

[44] Christopher R Palmer and Christos Faloutsos. Density biased sampling: An improved method for data mining and clustering. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 82–92, 2000.

[45] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021.

[46] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[48] Shiqi Yang, Shangling Jui, Joost Van De Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35:5802–5815, 2022.

[49] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in neural information processing systems*, 34:3635–3649, 2021.

[50] Gezheng Xu, Hui Guo, Li Yi, Charles Ling, Boyu Wang, and Grace Yi. Revisiting source-free domain adaptation: a new perspective via uncertainty control. In *The Thirteenth International Conference on Learning Representations*.

[51] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[52] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[53] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19275–19284, 2023.

[54] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15159–15168, June 2023.

[55] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[56] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29406–29419. Curran Associates, Inc., 2021.

[57] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[59] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations are discussed in the Conclusion Section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The Analysis and proof are provided in the supplementary material.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The implementation and training details are clearly described for reproduction in our main paper and supplementary material.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The code will be released publicly after in-peer review.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experiment settings are clearly presented in the paper and supplementary material.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive for experiments involving LLMs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The compute resources are reported in the experiment settings.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The broader impacts are discussed with the limitations.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: All datasets and models used in this paper are publicly available.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Proper citations are provided throughout the document and the licenses will be included with the code when it is released.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The document will accompany the code upon its release.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: We have fully disclosed the details of the use of the adopted LLMs in our supplementary material.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.