
Supplementary Material

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

In this supplementary material, we provide additional details, discussion, and theoretical analysis on our proposed **Interactive Text-3D Scene Retrieval Method (IDeal)**. Additionally, we then add more experiments and analysis, demonstrating the effectiveness and superiority of our proposed IDeal.

A More Details, Discussion, and Analysis on IDeal

In this section, we first provide a discussion on Density Compensated Similarity Correction in our proposed IRR (*i.e.*, corresponds to *footnote 1* in the paper). Second, we conduct a detailed analysis of the discriminability and diversity risks involved in the domain adaptation process of our IAT (*i.e.*, corresponds to *footnote 2*). Finally, we present the prompts used by the Large Language Models (LLMs) adopted by the *questioner* and *answerer* in our IDeal.

A.1 Discussion on Density Compensated Similarity Correction

To rigorously interpret the role of the density compensation mechanism, we formulate it from a Bayesian perspective, where feature density serves as a prior in similarity-based inference. Specifically, given a query feature \mathbf{u}_i and a scene feature \mathbf{v}_j , the objective is to estimate the posterior probability $\mathbb{P}(\mathbf{v}_j \mid \mathbf{u}_i)$ —the likelihood that \mathbf{v}_j corresponds to \mathbf{u}_i —based on observed feature similarity.

The introduction of the Density Compensated Factor can be theoretically motivated through a Bayesian lens. In a probabilistic retrieval framework, we are fundamentally interested in estimating the posterior probability $\mathbb{P}(\mathbf{v}_j \mid \mathbf{u}_i)$ —that is, the probability that a scene feature \mathbf{v}_j is relevant to the query \mathbf{u}_i . By Bayes’ theorem, this posterior is given by:

$$\mathbb{P}(\mathbf{v}_j \mid \mathbf{u}_i) = \frac{\mathbb{P}(\mathbf{u}_i \mid \mathbf{v}_j) \cdot \mathbb{P}(\mathbf{v}_j)}{\mathbb{P}(\mathbf{u}_i)},$$

where $\mathbb{P}(\mathbf{u}_i \mid \mathbf{v}_j)$ reflects the likelihood of observing query \mathbf{u}_i given the candidate \mathbf{v}_j , and $\mathbb{P}(\mathbf{v}_j)$ represents the prior distribution over scene features in the embedding space.

Standard similarity-based retrieval models typically estimate affinities based solely on a symmetric similarity function (e.g., cosine or dot product), which implicitly assumes a uniform prior $\mathbb{P}(\mathbf{v}_j)$. However, real-world scene feature distributions are often highly non-uniform, with significant variations in local density. In such cases, high-density regions contribute disproportionately to the affinity-based neighborhood, introducing structural bias in the entropy calculation.

To approximate a correction for this bias, we adopt:

$$\tilde{\mathcal{S}}(\mathbf{u}_i, \mathbf{v}_j) = \frac{\mathcal{S}(\mathbf{u}_i, \mathbf{v}_j)}{\sqrt{\rho(\mathbf{v}_j)}},$$

where $\rho(\mathbf{v}_j)$ is inversely proportional to the local density around \mathbf{v}_j and thus approximates the prior $\mathbb{P}(\mathbf{v}_j)$. This modification acts as a first-order approximation of dividing out the prior, thereby yielding a more posterior-like affinity measure. Substituting $\tilde{\mathcal{S}}$ into the affinity entropy computation thus gives rise to a *Density Compensated Affinity Entropy* that reflects a Bayesian-corrected view of similarity, mitigating the impact of density-induced retrieval bias.

A.2 Analysis of Discriminability and Diversity Risks in IAT

Firstly, we provide a detailed analysis of the discriminability risk term $\mathcal{R}_{\text{dis}}(\theta)$, which quantifies the misalignment between paired features in the shared representation space. Specifically, minimizing $\mathcal{R}_{\text{dis}}(\theta)$ encourages the augmented textual features $\tilde{\mathbf{u}}$ to be well aligned with their corresponding positive augmented features $\tilde{\mathbf{u}}^+$.

Under standard assumptions in representation learning [16], the distribution of positive features is modeled as:

$$\tilde{\mathbf{u}}^+ \sim \mathcal{N}(\bar{\mathbf{u}}^+, \Sigma_+^2),$$

where $\bar{\mathbf{u}}^+ = \mathbb{E}[\tilde{\mathbf{u}}^+]$ is the class mean and Σ_+^2 is the covariance matrix. Due to the semantic and visual diversity of wide-field scenes, the variance in positive features is large, i.e.,

$$\|\Sigma_+^2\|_F \gg 0.$$

The discriminability risk based on cosine distance is defined as:

$$\mathcal{R}_{\text{dis}}(\theta) = \mathbb{E}_{(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}^+)} [1 - \cos(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}^+)] = \mathbb{E} [1 - \tilde{\mathbf{u}}^\top \tilde{\mathbf{u}}^+].$$

Assuming $\tilde{\mathbf{u}}$ is approximately aligned with the mean direction $\bar{\mathbf{u}}^+$, we can decompose the risk as:

$$\mathcal{R}_{\text{dis}}(\theta) \approx 1 - \tilde{\mathbf{u}}^\top \bar{\mathbf{u}}^+ + \mathbb{E}_{\tilde{\mathbf{u}}^+} [1 - \bar{\mathbf{u}}^{+\top} \tilde{\mathbf{u}}^+],$$

where the second term reflects the *intrinsic angular spread* of the positive distribution. In fact, when $\tilde{\mathbf{u}}^+ \sim \mathcal{N}(\bar{\mathbf{u}}^+, \Sigma_+^2)$ and $\|\tilde{\mathbf{u}}^+\| \approx 1$, we have:

$$\mathbb{E} [1 - \cos(\bar{\mathbf{u}}^+, \tilde{\mathbf{u}}^+)] \propto \text{Tr}(\Sigma_+^2),$$

suggesting that the discriminability risk increases with the angular variance of the positive features. This highlights that encouraging strong alignment under high uncertainty (large Σ_+^2) incurs a high discriminability risk, potentially harming generalization in diverse visual contexts. To address the challenge, we propose substituting the ambiguous text features with corresponding point-cloud features when constructing positive pairs in *Section 3.3* of our paper’s main body.

Secondly, we analyze the upper bound of the diversity risk term $\mathcal{R}_{\text{div}}(\theta)$. Following existing analyses in domain adaptation literature [15], it is naturally related to both the expectation $\mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))$ and the variance $\mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))$ of the similarity measure. Depending on the specific distance metric employed, this upper bound may take different forms:

- (i) If distance metric is the χ^2 -divergence and

$$\delta \leq \mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-)) / [\mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))]^2, \text{ then}$$

$$\mathcal{R}_{\text{div}}(\theta) = \mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-)) + \sqrt{\delta \mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))}.$$

- (ii) If distance metric is the KL-divergence, then,

$$\mathcal{R}_{\text{div}}(\theta) = \mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-)) + \sqrt{2\delta \mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))} + \mathcal{O}(\delta),$$

with $\delta > 0$ being small.

- (iii) Suppose the discrepancy metric is chosen to be the cosine distance, defined by

$$\text{cos_dist}(\tilde{\mathbf{u}}, \tilde{\mathbf{u}}^-) = 1 - \frac{\langle \tilde{\mathbf{u}}, \tilde{\mathbf{u}}^- \rangle}{\|\tilde{\mathbf{u}}\| \cdot \|\tilde{\mathbf{u}}^-\|}.$$

If the scoring function $\mathcal{S}_\theta(\cdot)$ is L -Lipschitz continuous with respect to the Euclidean norm $\|\cdot\|$, i.e.,

$$|\mathcal{S}_\theta(\tilde{\mathbf{u}}) - \mathcal{S}_\theta(\tilde{\mathbf{u}}^-)| \leq L \|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}^-\|,$$

then the difference in the scoring outputs can be controlled by the Euclidean distance between the inputs.

By leveraging this Lipschitz continuity and the relationship between the cosine distance and the Euclidean norm, we can derive an upper bound on the robust diversity risk. Specifically, since the cosine distance can be related to the Euclidean distance via

$$\|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}^-\|^2 = 2(1 - \cos \theta) \|\tilde{\mathbf{u}}\| \|\tilde{\mathbf{u}}^-\|,$$

where θ is the angle between $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{u}}^-$, we connect the discrepancy metric to the perturbations in input space.

Consequently, the robust diversity risk admits the following upper bound:

$$\mathcal{R}_{\text{div}}(\theta) \leq \mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-)) + \sqrt{2\delta} \cdot L \cdot \sqrt{\mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\|\tilde{\mathbf{u}}^-\|)^2} + \mathcal{O}(\delta).$$

Here, the term $\sqrt{2\delta}$ arises from the concentration bound related to the discrepancy constraint, while the Lipschitz constant L scales the sensitivity of \mathcal{S}_θ , and the variance term captures the spread of the norm of the perturbations.

Up to this point, we find that the upper bound of diversity risk term $\mathcal{R}_{\text{div}}(\theta)$ is related to $\mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))$ and $\mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))$, i.e.,

$$\sup(\mathcal{R}_{\text{div}}(\theta)) \sim \{\mathbb{E}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-)); \mathbb{V}_{\tilde{\mathbf{u}}^- \sim \tilde{\mathcal{U}}^-} (\mathcal{S}(\tilde{\mathbf{u}}; \tilde{\mathbf{u}}^-))\}, \quad (1)$$

which also means that if we want to minimize the diversity risk term, we must simultaneously minimize the mean and variance of the selected negative samples.

A.3 Prompt Details

In this section, we fully disclose the prompts used for the LLMs using in *questioner* and *answerer* of IDEal to facilitate a better understanding of the interaction process.

1) Prompt in *questioner* to generates *questions for detail probe* (i.e., \mathcal{Q}_1).

Prompt
<p>Assume you are an expert in asking questions, and you need to constantly interact with users and ask questions to continuously understand the indoor room.</p> <p>User’s description content in the current round: $\{\textit{last-round description}\}$</p> <p>Requirement: Continue to ask for detailed relationships between the objects in the current description and other objects.</p> <p>Template: Continue describing and enriching the last description sentences about xxx, yyy only based information from your memory.</p> <p>Important rules:</p> <ol style="list-style-type: none"> 1. In English! 2. Don’t ask simple closed-ended questions; you want to ask more open-ended ones. 3. Don’t specify object categories, ask more broadly. 4. Just return to the question you want to ask, without any extra content or blank lines.

2) For the *questions for divergent exploration* (i.e., \mathcal{Q}_2), we do not adopt LLMs to generate specific questions, as we found that doing so often led to severely misleading hallucinations. Instead, we adopt a template-based text as a substitute: *Try to describe different objects within the passage from your memory that have not been covered in previous conversations. Only answer with about 5-7 sentences. Avoiding a repeat of before conversations. When answering, start with a focus object that is different from the previous ones. Just focus on describing spacial relationships of objects! Do not answer with any unnecessary extra descriptions beyond the spacial relationships of objects.*

3) Prompt in *answerer* in the initial round.

Prompt
<p>Assume this is the passage in your memory: $\{\textit{memory text}\}$</p> <p>Here is an original sentence describing a scene: $\{\textit{initial query}\}$</p> <p>Template for beginning of every sentence in your answer:</p>

1. this is ...
2. it is ...
3. there are ...
4. there is ...
5. xxx is ...
6. xxx are ...

Requirement: Q_1 . Supplement with $\{fit\ number\}$ additional sentences with the original sentence. Summarize all sentences to form new sentences in the original style. Follow the Important rules. Must use the original language style.

Important rules:

1. Answer $\{fit\ number\}$ English sentences at most. The returned answer is a paragraph, without blank lines.
2. Only use details that are available in your memory passage. Do not repeat the previous sentences. Do not fabricate new details!
3. If no new details about the object are present in the passage, do not fabricate new ones.
4. The sentence format should imitate the original sentence.
5. Be brief and don't answer with any unnecessary extra descriptions beyond the spacial relationships of objects.

where $\{fit\ number\}$ represents the number of items that fit the current task data. For coarse-grained memory, it is generally set to 2–4; for fine-grained memory, it is typically set to 3–6; and for summarization tasks, it is set higher, around 5–8.

4) Prompt in *answerer* in the following round.

If last-round description is uninformative (*i.e.*, $\tilde{\mathcal{E}} > \beta$), the *answerer* need to answer *questions for detail probe*:

Prompt

Q_1 . Supplement with $\{fit\ number\}$ additional sentences with the last user descriptions. Summarize all sentences to form new sentences in the original style. Follow the Important rules. Must use the original language style. Do not describe the purpose of the object, just focus on describing its placement in the room! Do not answer with any unnecessary extra descriptions beyond the spacial relationships of objects.

If last-round description is informative (*i.e.*, $\tilde{\mathcal{E}} \leq \beta$), the *answerer* need to answer *questions for divergent exploration*:

Prompt

Q_2 . Try to describe different objects within the passage from your memory that have not been covered in previous conversations. Only answer with about $\{fit\ number\}$ sentences. Avoiding a repeat of before conversations. When answering, start with a focus object that is different from the previous ones. Just focus on describing spacial relationships of objects! Do not answer with any unnecessary extra descriptions beyond the spacial relationships of objects.

B Additional Experiments

In this section, we first introduce the datasets used in our experiments, followed by a discussion of the coarse-grained and fine-grained descriptions in the memory settings to facilitate a deeper understanding of the experimental settings. Second, we present additional experiments, including: evaluations of our method under realistic conditions where source queries undergo textual domain shifts.

B.1 Introductions to the Adopted Datasets

We provide details about the point cloud and text datasets used in experiments. For the point-cloud dataset, ScanNet [5] is an instance-level 3D dataset comprising thousands of 3D point-cloud scans and nearly 2.5 million views across over 1,500 indoor room scenes. Additionally, for experiment evaluation, we follow [6] using 10 description texts per scene for training and evaluation. The text datasets used for descriptions are outlined as follows:

- *ScanRefer* [3]: It serves as a large-scale and highly discriminative dataset for 3D visual grounding and dense captioning, featuring 51,583 object descriptions from thousands of objects spanning nearly 800 ScanNet scenes. In our experiments, we use 562 scenes for training and 141 scenes for testing.
- *Nr3d/Sr3D* [1]: NR3D and SR3D are also built upon ScanNet, with SR3D consisting of 83,572 straightforward machine-generated descriptions and NR3D featuring 41,503 descriptions, closely resembling the human annotations found in ScanRefer. In the experiments, we use 511 scenes for training and 130 scenes for testing of Nr3D, 476 scenes for training and 116 scenes for testing of Sr3D.
- *SceneDepict-3D2T* [6]: Text-3D scene retrieval fine-grained text set based on ScanNet. It contains 7,030 fine-grained scene descriptions for 703 3D scenes. In the experiments, this dataset is primarily used as the source of fine-grained memory texts. Specifically, under using fine-grained memory text setting, during each query interaction, the user’s memory consists of a single SceneDepict-3D2T fine-grained description corresponding to the scene.

B.2 Detailed Discussions about Memory Texts

In this section, we provide more details and discussions on the memory settings of the *answerer*. More specifically,

1) *Using coarse-grained description as memory*: We leverage an LLM [2] to generate rich expansions of queries, serving as memory without introducing any additional information leakage. It primarily involves four types of expansions: i) sentence-level coherence refinement, ii) subject-object role reversal in object descriptions, iii) plausible scene inference based on the current description, and iv) synonym substitution for objects and attributes.

2) *Using fine-grained description as memory*: In line with existing interactive methods [12], we simulate the user’s memory in real-world scenarios using fine-grained descriptions from SceneDepict-3D2T [6], albeit with access to partial additional information. This setting simulates real-world scenarios where user queries often provide limited information [12], even though users may possess richer, latent knowledge about the scene in their minds that has yet to be expressed. This is because, in the absence of any reference, it is inherently difficult for individuals to articulate all relevant information in a single attempt. We argue that although this setting entails a certain degree of information leakage, it is comparatively weaker than the more pervasive leakage found in widely accepted interactive works [10, 9, 8], where users are granted access to ground truth (GT) information. In contrast, the fine-grained descriptions used here do not reveal the GT scene directly. As such, this setting better approximates realistic application scenarios, where evaluating a model’s ability to uncover and utilize potentially vague or implicit user intent is both acceptable and practically meaningful.

B.3 Text-3D Scene Retrieval under Textual Domain Shift

In this section, we conduct additional T3SR comparative experiments to evaluate the model’s robustness under domain shifts commonly encountered in real-world scenarios. Specifically, we introduce corruptions into the text queries to simulate realistic noise, such as character-level misspellings, omissions, and word-level deletions. The domain shift settings follow the protocol proposed in recent work [11]. More precisely, the corruptions applied to the queries can be categorized into two levels: character-level and word-level. Character-level corruptions include OCR errors, character insertion, character replacement, character swapping, and character deletion—all of which mimic typographical errors or input noise. Word-level corruptions consist of word insertion, word swapping, and word deletion, simulating variations in writing habits where users may alter sentence structure while conveying the same semantic content. We randomly apply 3–5 of the aforementioned corruptions

Table 1: Performance comparison on ScanRefer in terms of R@1, R@5, R@10, and their sum (Rsum) under domain shifted settings.

Method	ScanRefer			
	R@1	R@5	R@10	Rsum
VSE ∞	8.6	28.7	46.4	83.7
CHAN	8.5	29.4	47.4	85.3
HREM	9.7	30.5	48.2	88.4
CRCL	9.3	28.8	48.1	86.2
RoMa	10.1	29.8	49.2	89.1
IDEal	14.5	40.8	58.2	113.5
Δ	+4.4	+10.3	+9.0	+23.7

to each query in ScanRefer dataset [3] to construct corrupted (domain-shifted) textual inputs. We then evaluate a well-trained cross-modal retrieval model (*i.e.*, VSE ∞ [4], CHAN [13], HREM [7], CRCL [14], and RoMa [6]) on these corrupted queries. The experimental results are presented in Table 1.

The following observations can be drawn from the results: **1)** The introduction of domain shifts significantly impairs the ability of existing offline models to accurately interpret the true semantics of the input queries, resulting in notable performance degradation. **2)** IDEal demonstrates superior retrieval performance under such conditions. This can be attributed to IDEal’s progressive interaction with external LLMs, which enables it to adapt to and correct for domain shifts, highlighting IDEal’s robustness and effectiveness in addressing inherent query limitations.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [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] 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.
- [8] 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.

- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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*.
- [16] Hamed Zamani and Michael Bendersky. Multivariate representation learning for information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’23, page 163–173, New York, NY, USA, 2023. Association for Computing Machinery.