PointCloud-Text Matching: Benchmark Dataset and Baseline

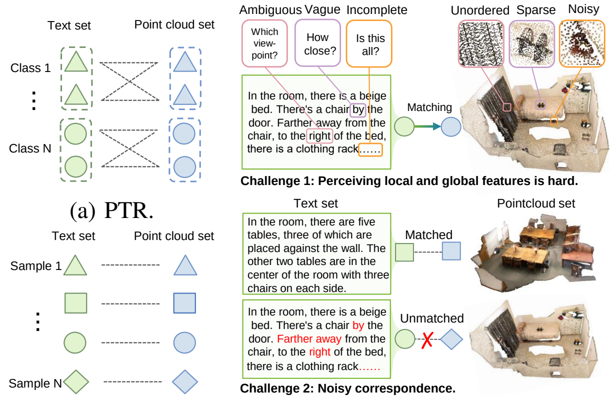
Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng and Peng Hu

***Abstract*—In this paper, we present and study a new instance-level retrieval task: PointCloud-Text Matching (PTM), which aims to identify the exact cross-modal instance that matches a given point-cloud query or text query. PTM has potential applications in various scenarios, such as indoor/urban-canyon localization and scene retrieval. However, there is a lack of suitable and targeted datasets for PTM in practice. To address this issue, we present a new PTM benchmark dataset, namely SceneDepict-3D2T. We observe that the data poses significant challenges due to its inherent characteristics, such as the sparsity, noise, or disorder of point clouds and the ambiguity, vagueness, or incompleteness of texts, which render existing cross-modal matching methods ineffective for PTM. To overcome these challenges, we propose a PTM baseline, named Robust PointCloud-Text Matching method (RoMa). RoMa consists of two key modules: a Dual Attention Perception module (DAP) and a Robust Negative Contrastive Learning module (RNCL). Specifically, DAP leverages token-level and feature-level attention mechanisms to adaptively focus on useful local and global features, and aggregate them into common representations, thereby reducing the adverse impact of noise and ambiguity. To handle noisy correspondence, RNCL enhances robustness against mismatching by dividing negative pairs into clean and noisy subsets and assigning them forward and reverse optimization directions, respectively. We conduct extensive experiments on our benchmarks and demonstrate the superiority of our RoMa.**

***Index Terms*—PointCloud-Text Matching, Noisy correspondence, Benchmark dataset.**

I. Introduction

Point clouds are a popular representation of the 3D geometry of a scene, with significant applications in computer vision, robotics, and augmented reality, such as autonomous driving [1], [2], object detection [3], and localization [4]. However, as the volume of point-cloud data continues to grow rapidly, it is urgent to have techniques that enable users to effectively and accurately find the exact matching instance/scene from large-scale point-cloud scans, especially using natural language queries, named PointCloud-Text Matching (PTM). The instance-level alignment is challenging and realistic as it reflects the need for precise and relevant information to build alignment between point clouds and texts in real-world applications, which has potential applications in indoor/urban-canyon localization, scene retrieval, and more.


Text set
Point cloud set
Class 1
Class N
Text set
Point cloud set
Sample 1
Sample N
Ambiguous
Vague
Incomplete
Which view-point?
How close?
Is this all?
Unordered
Sparse
Noisy
In the room, there is a beige bed. There's a chair by the door. Farther away from the chair, to the right of the bed, there is a clothing rack……
Matching
Challenge 1: Perceiving local and global features is hard.
Text set
Pointcloud set
In the room, there are five tables, three of which are placed against the wall. The other two tables are in the center of the room with three chairs on each side.
Matched
In the room, there is a beige bed. There's a chair by the door. Farther away from the chair, to the right of the bed, there is a clothing rack……
Unmatched
Challenge 2: Noisy correspondence.

(a) PTR.

(b) PTM.

(c) The challenges of PTM.

Fig. 1: Overview for PointCloud-Text Matching (PTM). (a) and (b) show the schematic illustrations of class-level PointCloud-Text Retrieval (PTR), and instance-level PTM, respectively. (c) illustrates the challenges faced by PTM.

Existing methods, however, struggle and lack pertinence to tackle PTM. On one hand, existing PointCloud-Text Retrieval (PTR) methods [5], [6] only focus on establishing category-level correspondence between 3D point-cloud shapes and short annotation texts as shown in Fig. 1 (a). In contrast, PTM requires exploiting the mutual information of cross-modal pairs, and achieves instance-level alignment between point-cloud scenes and detailed descriptions as shown in Fig. 1 (b). This indicates that PTM demands the ability to capture local features and instance discrimination, rendering the existing methods inapplicable. On the other hand, existing cross-modal matching works that can build instance-level cross-modal correspondence are only primarily oriented to text and 2D image modalities. According to the granularity of the established correspondence, these cross-modal matching works could be divided into two groups: coarse-grained and fine-grained matching methods. The former [7]–[9] use global features to represent the whole image and the whole text, while the latter [10]–[12] use local features to capture the fine-grained details of regions and words. Although these methods have achieved promising performance for image-text matching task, most of them ignore the specific properties and

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by NSFC under Grant 62372315, 62472295, 62176171 and U21B2040; in part by Sichuan Science and Technology Planning Project under Grant 2024YFHZ0144, 2024YFHZ0089, 2024NSFTD0047 and 2024NSFTD0038; in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant; and in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303 and CJ202403. *(Corresponding author: Peng Hu.)*

Yanglin Feng, Yang Qin, Dezhong Peng, Xi Peng, and Peng Hu are with the College of Computer Science, Sichuan University, Chengdu 610065, China, Dezhong Peng is also with Sichuan National Innovation New Vision UHD Video Technology Co., Ltd, Chengdu 610095, China (e-mail: fcyzfyl@163.com; qinyang.gm@gmail.com;pengdz@scu.edu.cn; pengx.gm@gmail.com; penghu.ml@gmail.com).

Hongyuan Zhu is with the Institute for Infocomm Research (I^{2}R), A*STAR, Singapore (e-mail: hongyuanzhu.cn@gmail.com).

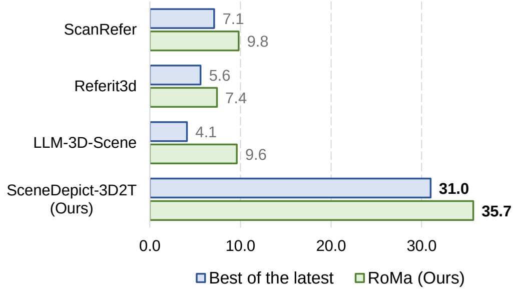


Fig. 2: PTM performance (i.e., R@1) of the latest cross-modal matching methods (i.e., DIVE [19], CHAN [20], HREM [21], and CRCL [9]) and our RoMa on existing ScanRefer, Nr3d, 3D-LLM-Scene and proposed SceneDepict-3D2T dataset.

challenges in PTM.

To the best of our knowledge, the insufficient information provided by existing datasets makes them unsuitable for PTM. To be specific, descriptions in most datasets (e.g., ScanRefer [13], Nr3d [14] primarily focus on portraying a single object for visual grounding and captioning, and a few other (e.g., LLM-3D-Scene [15]) describing several objects in isolation within the corresponding scenes. These limited descriptions match precisely with the corresponding wide-field point clouds, as demonstrated by the dismal matching performance in existing datasets depicted in Fig. 2. Therefore, we constitute a new benchmark dataset for PTM, namely SceneDepict-3D2T. The dataset contains comprehensive descriptions covering entire 3D point-cloud scenes, so they evaluate baselines more reliably and reasonably for PTM, which can be observed in Fig. 2. We also provide a comprehensive evaluation protocol and several benchmark results for PTM on the datasets as shown in Table I. From the results, we observe that point cloud-text data are more challenging than the common image-text data due to the sparsity, noise, or disorder of point clouds [16]. More specifically, these properties make it difficult to capture and integrate local and global semantic features from both point clouds and texts and may also lead to mismatched cross-modal pairs, i.e., noisy correspondence [17], [18], thus degrading the retrieval performance. The schematic illustration of the challenges is shown in Fig. 1 (c). To be specific, the existing coarse-grained matching methods fail to extract discriminative global features from the unordered point clouds and vague texts, and the fine-grained matching methods that rely on well-detected object regions cannot be generalizable to point clouds. Moreover, most existing methods are based on well-annotated data and are susceptible to overfitting noisy correspondence, resulting in performance degradation. Therefore, there is a significant gap in applying existing methods to PTM.

To tackle the aforementioned challenges, we propose a PTM baseline, named **Ro**bust PointCloud-Text **Ma**tching method (RoMa), to learn from point clouds and texts as illustrated in Fig. 7. RoMa consists of two modules: a Dual Attention Perception module (DAP) and a Robust Negative Contrastive Learning module (RNCL). DAP is proposed to adaptively capture and integrate the local and global informative features to alleviate the impact of noise and ambiguity in the data. More specifically, DAP conducts token-level and feature-level attention to adaptively weigh the patches and words to multi-grainly aggregate the local and global discriminative features into common representations, thus embracing a comprehensive perception. In addition, our RNCL is presented to adaptively divide the negative pairs into clean and noisy subsets based on the similarity within pairs, and then assign them with forward and reverse optimization directions respectively. Different from traditional contrastive learning, our RNCL only leverages negative pairs rather than positive pairs to train the model since negatives are much less error-prone than positive pairs, leading to robustness against noisy correspondence. In brief, our RNCL could utilize and focus on more reliable pairs to enhance the robustness.

In summary, our main contributions are as follows:

- We investigate a new instance-level cross-modal retrieval task, namely PointCloud-Text Matching (PTM), and propose a PTM benchmark dataset SceneDepict-3D2T and a robust baseline RoMa to learn from challenging multi-modal data for PTM.
- We present a novel Dual Attention Perception module (DAP) that adaptively extracts and integrates the local and global features into common representations by using token-level and feature-level attention, thereby achieving a comprehensive perception of semantic features.
- To handle noisy correspondence, we devise a Robust Negative Contrastive Learning module (RNCL) that adaptively identifies clean and noisy negative pairs, and assigns them correct optimization directions accordingly, thus preventing the model from overfitting noise.
- We conduct extensive comparison experiments on four pointcloud-text datasets. Our RoMa remarkably outperforms the existing methods without bells and whistles, demonstrating its superiority over existing methods.

II. Related Work

A. Cross-modal Retrieval

Cross-modal retrieval aims to search the relevant results across different modalities for a given query, e.g., image-text matching [8], [11], 2D-3D retrieval [22], [23], and visible-infrared re-identification [24], etc. Most of these works learn a joint common embedding space by applying cross-modal constraints [25], [26], which aims to pull relevant cross-modal samples close while pushing the irrelevant ones apart. These methods could roughly be classified into two groups: 1) Coarse-grained retrieval [7], [8], [21], [27], [28] typically learns shared subspaces to build connections between global-level representations, which align images and texts in a direct manner. 2) Fine-grained retrieval [11], [20], [29] aims to model cross-modal associations between local feature representations, e.g., the visual-semantic associations between word tokens and image regions. Unlike them, in this paper, we delve into a less-touched and more challenging cross-modal scenario, i.e., Pointcloud-Text Matching (PTM), which argues for building cross-modal associations between 3D space and textual space.

B. 3D Vision and Language

In contrast to image and language comprehension, 3D vision and language comprehension represent a relatively nascent frontier in research. Most existing works focus on using language to confine individual objects, e.g., distinguishing objects according to phrases [30] or detecting individual objects [31]. With the introduction of the ScanNet [32], ScanRefer [13], and Nr3d [14] datasets, more works have expanded their focus to encompass the 3D scenes. Some existing works [33], [34] have tried to locate objects within scenes based on linguistic descriptions, completing the task of 3D visual grounding. Recently, with the introduction of Scan2Cap [35], some efforts [36] focus on providing descriptions for objects about their placement. This is also known as 3D dense captioning. Recently, a few preliminary solutions [5], [6] for pointcloud-text retrieval have begun to emerge, which only establish common discrimination for coarse category-level alignment between point-cloud shapes and brief category label texts. However, these category-level methods could not be migrated to PTM. There are still scarce methods focusing on instance-level alignment and matching between wide-field point clouds and natural language texts, which requires excavating more detailed and discriminative connections within cross-modal pairs.

III. PointCloud-Text Matching

In this paper, we introduce a novel 3D vision and language task, namely PointCloud-Text Matching (PTM). The input cross-modal data of the task involves the 3D point clouds and free-form description texts. The goal of PTM is to support bi-directional retrieval between point clouds and corresponding texts, achieving instance-level cross-modal alignment.

However, the task presents notable discrepancies and task-specific challenges, which can be summarized as follows:

- **Perceiving local and global semantic features is hard.** Since sensor sampling characteristics and biases, point clouds are commonly presented as a collection of sparse, noisy, and unordered points. Compared to 2D images, point clouds encapsulate a wealth of additional objects and spatial properties, which results in more incomplete and ambiguous description texts. Such complexity makes it harder for existing models to accurately perceive local and global semantic features from both modalities.
- **Noisy correspondence.** Imperfect annotations are ubiquitous, even well-labeled datasets containing latent noisy labels, as shown by the existence of over 100,000 label issues in the ImageNet [37] and 3%-20% annotation errors in the Conceptual Captions [38]. However, due to the limitations of human perception and description of 3D space, annotation workers are unintentionally inclined to use vague expressions (such as ‘near’, ‘close to’, etc.) to describe the details of the point clouds incorrectly, introducing more correspondence annotation (i.e., noisy correspondence). Such noise would lead to insufficient learning or noise overfitting for existing models.

IV. Benchmark Datasets: SceneDepict-3D2T

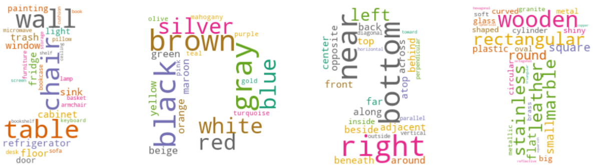


Fig. 3: Word clouds of objects, colors, spatial information, and shape material in the descriptions of SceneDepict-3D2T.

To the best of our knowledge, existing multi-modal datasets of point clouds and texts can not apply directly to PointCloud-Text Matching (PTM). On the one hand, the descriptions in most of these datasets (e.g., ScanRefer [13], Nr3d [14], and ScanQA [39]) are confined to single objects of the entire point-cloud scenes. Fig. 6 shows they only have average lengths of fewer than 15 words and each description encompasses fewer than 2 objects. However, one scene typically contains 10-30 objects [32], with many similar objects present across different scenes. This indicates that these short and inadequately informative descriptions are prone to be ambiguous, lacking the discrimination to meet the requirements of PTM. On the other hand, although several scene description datasets [15] for scene understanding with the Large Language Model (LLM) have been recently proposed, they exhibit limited data volume and lack inter-object relationships necessary for a comprehensive description for each specific scene. These limitations hinder their application in PTM, as shown in Fig. 6. We conduct PTM experiments on these existing datasets, and the matching results in Fig. 2 show that the average performance of the latest cross-modal matching methods in terms of Recall at 1 is less than 10%, confirming the above view.

To establish a reasonable evaluation of PTM with practical significance, we construct a diverse, detailed, and discriminative benchmark dataset with scene-level descriptions for comprehensive point-cloud scene understanding, namely SceneDepict-3D2T. In SceneDepict-3D2T, the point-cloud data is all based on the ScanNet [32] dataset, and the text data is derived from the ScanRefer [13], Nr3d [14], and ScanQA [39] description sets associated with ScanNet. In the following sections, we will elaborate on the collection and statistics of our proposed datasets.

A. Data Collection

We deploy a prompt-driven LLM paradigm to generate scene-level descriptions of point-cloud scene scans in ScanNet, leveraging three existing object-level description datasets (i.e., ScanRefer, Nr3d, and ScanQA). The description generation pipeline is divided into three stages, as illustrated in Fig. 4.

1) *Scene Analysis Stage*: We first divide each scene scan into multiple neighborhoods and identify discriminative objects based on their color, size, and more. We then randomly select n descriptions of n spatially related objects from different neighborhoods, creating object-level description collection. This stage arbitrarily introduces different object characteristics that guide the generated descriptions to encapsulate varied information and scene discrimination of point clouds.

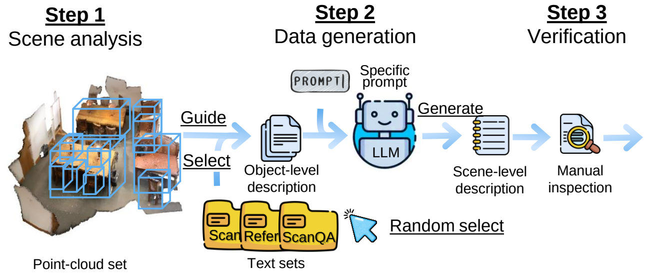
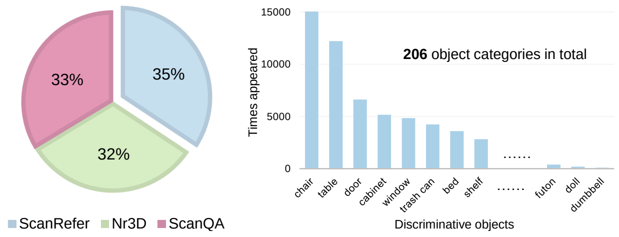


Fig. 4: Pipeline of the one scene-level description collecting process in our SceneDepict-3D2T dataset.



(a) Corpus distribution (b) Scene object distribution

Fig. 5: Data distribution statistics. (a) shows the proportion of source corpus, and (b) shows the appeared times statistics of the discriminative objects in our SceneDepict-3D2T dataset.

2) *Data Generation Stage*: In this stage, we randomly select one of the three object-level datasets and input the corresponding object-level description collection into the LLM using the tailored prompts to generate scene-level descriptions. The designed prompts align with the various linguistic characteristics of different datasets, ensuring the language style diversity and applicability of PTM across various scenarios. Specifically, for ScanRefer, the generated descriptions are objective and exhaustive, being suitable for scenarios with precise matching of the object placement throughout the scans. For Nr3d, the descriptions are concise and informative, being applicable for scenarios with matching of partial arrangement of objects within scans. For ScanQA, the descriptions detail object characteristics and relationships, which are suitable for scenarios with matching of key features of scans.

3) *Verification Stage*: In this final stage, we manually assess each generated description for discriminative accuracy, grammatical correctness, and coherence, completing the scene-level description construction.

Repeating the above process ten times, we generate ten unique scene-level descriptions for each point cloud. By following this process for all point clouds, we create the SceneDepict-3D2T dataset. Note that more construction details and data examples are provided in the Appendix.

B. Dataset Statistics

To provide a comprehensive overview of the proposed SceneDepict-3D2T dataset, we present data distribution statistics and compare it with those of existing datasets, as depicted in Figs. 5 and 6.

More specifically, the corpus adopted for description generating is evenly sourced from the datasets with distinct linguistic styles, showcasing the comprehensiveness of the constructed descriptions. Additionally, our SceneDepict-3D2T offers a wealth of grammatical scene-level descriptions suitable for PTM training and validation. On average, each description in SceneDepict-3D2T covers 10.7 objects and 8.8 object categories, which is over 5 times more than the object-level datasets (i.e., ScanRefer, Nr3d, and ScanQA). Furthermore, each description encompasses 6.6 inter-object interactions, which is 6.6 times higher than in the scene description dataset 3D-LLM-Scene. This demonstrates the sufficient scene coverage and discrimination of the descriptions in SceneDepict-3D2T. In addition, the descriptions in SceneDepict-3D2T are rich in color (85.8%), material (38.7%), shape terms (56.1%), and spatial information (99.0%), ensuring their informational depth. Consequently, benefiting from the discriminative and detailed descriptions in SceneDepict-3D2T, baselines can achieve 100% to 400% improvement in PTM performance compared to existing datasets, as shown in Fig. 2, underscoring its practical significance for PTM.

Despite meticulous verification efforts to improve the grammatical accuracy and syntactic coherence of the datasets, it is unavoidable to introduce a considerable portion of noisy correspondence because of the inherent nature of unordered point-cloud scenes and vague free-form descriptions. To assess this, we randomly selected 100 descriptions from SceneDepict-3D2T and manually checked for vague expressions that might lead to noisy correspondence. Eventually, 13 descriptions were flagged for potential issues. Thus, noisy correspondence remains an unavoidable challenge in PTM, which could result in noise overfitting, leading to performance degradation.

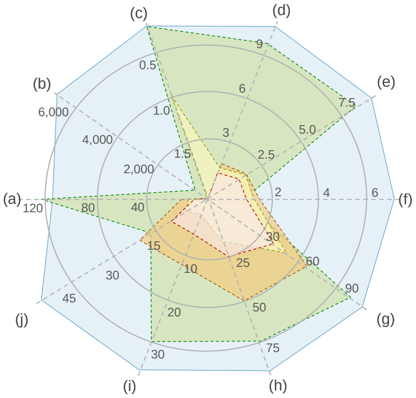
V. Robust Baseline: RoMa

A. Problem Formulation

We first define the notations for a lucid presentation. Give a PTM dataset $\mathcal{D} = \{\mathcal{P}, \mathcal{T}\}$, where $\mathcal{P} = \{X_i^p\}_{i=1}^{N_p}$ and $\mathcal{T} = \{X_j^t\}_{j=1}^{N_t}$ are the point-cloud and text sets respectively, N_p and N_t are the size of $\mathcal{P}$ and $\mathcal{T}$, X_i^p is the i^{th} point-cloud sample and X_j^t is the j^{th} text sample. There is pairwise correspondence between $\mathcal{P}$ and $\mathcal{T}$, so $\mathcal{D}$ can also be written as $\mathcal{D} = \{(X_i^p, X_j^t), y_{i,j}\}_{i,j}^{N_p,N_t}$, $y_{i,j} \in \{0, 1\}$ indicates whether X_i^p and X_j^t are matched (i.e., positive pair, $y_{i,j} = 1$) or unmatched (i.e., negative pair, $y_{i,j} = 0$). However, in practice, the unmatched pairs $(y_{i,j} = 0)$ may be mislabeled as matched ones $(y_{i,j} = 1)$, *a.k.a* noisy correspondence.

In the data encoding stage, we first employ modality-specific backbones (i.e., f_p and f_t) to extract token-wise features for the patches of point clouds and words of descriptions, i.e., $Z_i^p = f_p(X_i^p) \in \mathbb{R}^{p_n \times d_c}$ and $Z_j^t = f_t(X_j^t) \in \mathbb{R}^{t_n \times d_c}$, respectively. Z_i^p and Z_j^t are the token-wise feature sets of i^{th} point cloud and j^{th} text, p_n and t_n are the number of tokens for each sample and d_c is the dimensionality of the feature space.

In addition, to preserve spacial interactions across patches within point clouds, inspired by the sequence position representation [40], we attempt to encapsulate the 2D position



Attributes		🟦	🟩	🟧	🟨	🟥
		3D2T	3DLLM	SR	Nr3D	SQA
General	(a) Average text length	**108.9**	127.8	17.9	11.4	9.2
	(b) Scene description number	**7,050**	517	0.0	0.0	0.0
	(c) Average of grammatical errors	**0.0**	0.0	2.04	1.0	--
Scene discrimination	(d) Number of objects per text	**10.7**	10.1	1.8	1.7	1.6
	(e) Number of object categories per text	**8.8**	8.4	1.5	1.6	1.4
	(f) Number of inter-object interactions per text	**6.6**	1.0	1.2	1.1	0.9
Information richness	(g) Ratio of texts with spatial position (%)	**99.0**	96.7	69.8	47.5	46.9
	(h) Ratio of texts with color terms (%)	**85.8**	74.9	50.6	20.5	24.9
	(i) Ratio of texts with material terms (%)	**38.7**	30.6	12.3	2.1	5.5
	(j) Ratio of texts with shape terms (%)	**56.1**	17.0	20.4	6.5	11.3

Fig. 6: Statistics comparison among existing ScanRefer (SR) [13], Nr3d [14], ScanQA (SQA) [39], 3D-LLM-Scene (3DLLM) [15], and proposed SceneDepict-3D2T (3D2T) dataset benchmarks.

information of patches into a position embedding. The embedding is then used for following comprehensive token-level attention calculation, as detailed below:

$$E_i = \{f(E^x_{i,1}, E^y_{i,1}), \cdots, f(E^x_{i,p_n}, E^y_{i,p_n})\}, \tag{1}$$

where $E_i \in \mathbb{R}^{p_n \times d_c}$ is the patch position embedding of the i^{th} point-cloud sample, f denotes the fusion method (e.g., summation, concatenation, etc.) for combining the patch position embeddings, and $E^x_{i,j}$ and $E^y_{i,j}$ are the patch position embeddings calculated from the horizontal and vertical coordinates of patch centroids in j^{th} patch of i^{th} point-cloud sample. These embeddings are computed as follows:

$$E^x_{i,j,\epsilon} = \sin\left(\frac{h_{i,j} \cdot p_n}{10000^{\frac{\epsilon}{d_c}}}\right), \quad E^x_{i,j,\omega} = \cos\left(\frac{h_{i,j} \cdot p_n}{10000^{\frac{\omega-1}{d_c}}}\right), \tag{2}$$

$$E^y_{i,j,\epsilon} = \sin\left(\frac{v_{i,j} \cdot p_n}{10000^{\frac{\epsilon}{d_c}}}\right), \quad E^y_{i,j,\omega} = \cos\left(\frac{v_{i,j} \cdot p_n}{10000^{\frac{\omega-1}{d_c}}}\right), \tag{3}$$

where $h_{i,j}$ and $v_{i,j}$ are the normalized horizontal and vertical coordinates of the corresponding patch centroids, and ϵ and ω refer to the even and odd dimensionality indices of $E^x_{i,j}$ and $E^y_{i,j}$, respectively.

To tackle the task-specific challenges mentioned earlier, a robust PTM method (RoMa) is proposed to learn cross-modal associations from point clouds and texts as shown in Fig. 7. The proposed method involves two modules: 1) Dual Attention Perception (DAP) is used to comprehensively perceive semantic features with dual attention at the dataset level, and 2) Robust Negative Contrastive Learning (RNCL) is exploited to handle noisy correspondence. In the following sections, we will elaborate on each component of RoMa.

B. Dual Attention Perception

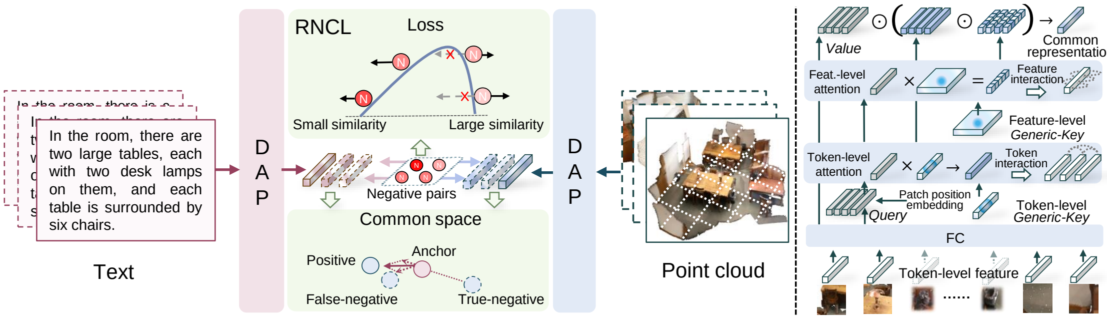
To tackle the challenge of capturing and integrating both local and global semantic features from point clouds and texts in PTM, we propose a novel dual attention mechanism. More specifically, similar to the definition in Self-Attention mechanism (SA) [40], we calculate the *Queries* $Q^p_i \in \mathbb{R}^{p_n \times d_c}$, $Q^t_i \in \mathbb{R}^{t_n \times d_c}$, along with the *Values* $V^p_i \in \mathbb{R}^{p_n \times d_c}$, $V^t_j \in \mathbb{R}^{t_n \times d_c}$ of two modalities from the i^{th} point-cloud and j^{th} text features Z^p_i and Z^t_j, using the fully connected layers g_p and g_t, respectively. However, unlike SA, our dual attention mechanism constructs learnable token-level and feature-level *Generic-Keys*, which are shared across all samples in the dataset. Benefit from this, the *Generic-Keys* could learn to model general patterns of tokens and features throughout the entire dataset. By integrating these *Generic-Keys* with the sample-specific *Queries*, our method achieves more comprehensive attention than SA, capturing interactions beyond token-wise relationships within individual samples, as depicted in Fig. 7 (b).

To be more specific, to facilitate the adaptive exploration of local semantic features, we first construct token-level *Generic-Keys* $\bar{K}^p \in \mathbb{R}^{d_c}$ and $\bar{K}^t \in \mathbb{R}^{d_c}$ upon the feature matrices of point-cloud and text modalities. We use them to model the common patterns of informative tokens (i.e., patches and words) within each modality. Similar to SA, we obtain token-level attention by measuring the token-wise similarity between *Queries* and token-level *Generic-Keys*, empowering the model to selectively focus on local key semantic units similar to the common patterns in the two modalities (e.g., foreground patches in the point clouds and keywords in the texts), which are written as:

$$\bar{a}^p_i = \varphi((Q^p_i + E_i)\bar{K}^{p\top}), \quad \bar{a}^t_i = \varphi(Q^t_i \bar{K}^{t\top}), \tag{4}$$

where $\bar{a}^p_i \in \mathbb{R}^{p_n}$ and $\bar{a}^t_i \in \mathbb{R}^{t_n}$ are the token-level attention vectors, φ is the Softmax operation along the token dimension (i.e., row-wise operation on the feature matrices). It is worth noting that we add patch position embedding in point-cloud modality to preserve the spatial interactions among patches. Based on this, we obtain the token-level attention $\bar{A}^p_i \in \mathbb{R}^{p_n \times d_c}$ and $\bar{A}^t_i \in \mathbb{R}^{t_n \times d_c}$ by stacking these attention vectors in the feature dimension.

In addition, we propose feature-level attention to capture feature semantics and enhanced cross-modal representations. Similar to token-level modeling, we introduce learnable feature-level *Generic-Keys* $\hat{K}^p \in \mathbb{R}^{d_c \times d_c}$ and $\hat{K}^t \in \mathbb{R}^{d_c \times d_c}$ for two modalities, which aims to model the interaction patterns among d_c features. We construct feature-level attention by combining *Queries* and feature-level *Generic-Keys* to grasp global discriminative features from the dimensional interrelationships in the feature space, such as distinctive object



(a) Pipeline of our RoMa

(b) Pipeline of DAP adopted by RoMa

Fig. 7: The illustration of our proposed method. (a) shows the pipeline of our RoMa, which involves two modules: Dual Attention Perception (DAP) and Robust Negative Contrastive Learning (RNCL). In DAP, comprehensive common representations could be extracted from both modalities and then matched into negative pairs. In RNCL, these negative pairs are adaptively optimized in both forward and reverse directions based on pairwise similarities, enhancing the robustness and discrimination of the common representations. (b) is the schematic illustration of DAP in point-cloud modality, which operates similarly for the text modality. *Query* and *Value* are obtained from features through a fully connected layer (FC), while *Generic-Key* is general and learnable for the whole dataset. The *Query* is combined with token-level and feature-level *Generic-Key* to obtain dual attention. Following this, the features and attentions are aggregated into common representations.

color, position, orientation, spatial relationships, etc., which is written as:

$$\hat{A}_i^p = \varphi(Q_i^p \hat{K}^{p\top}), \quad \hat{A}_i^t = \varphi(Q_i^t \hat{K}^{t\top}), \tag{5}$$

where $\hat{A}_i^p \in \mathbb{R}^{p_n \times d_c}$ and $\hat{A}_i^t \in \mathbb{R}^{t_n \times d_c}$ are the feature-level attention.

Next, we aggregate the token-level and feature-level attention into dual attention, which can be written as:

$$A_i^p = \bar{A}_i^p \odot \hat{A}_i^p, \quad A_i^t = \bar{A}_i^t \odot \hat{A}_i^t, \tag{6}$$

where A_i^p and A_i^t are dual attention in point-cloud and text modalities, and the $\odot$ is the Hadamard product operator. Subsequently, we impose dual attention upon the *Values*, aggregating them for integrated representations into common space, which are written as:

$$\boldsymbol{p}_i = Norm(\frac{1}{p_n}\sum_j^{p_n}(A_{i,j}^p \odot V_{i,j}^p)), \tag{7}$$

$$\boldsymbol{t}_i = Norm(\frac{1}{t_n}\sum_j^{t_n}(A_{i,j}^t \odot V_{i,j}^t)), \tag{8}$$

where $A_{i,j}^p$ and $A_{i,j}^t$ are the j^{th} row of dual attention A_i^p and A_i^t, $V_{i,j}^p$ and $V_{i,j}^t$ are the j^{th} row of the *Values* V_i^p and V_i^t, and $Norm(\cdot)$ is the l_2-normalization function. The common representations $\boldsymbol{p}_i \in \mathbb{R}^{d_c}$ and $\boldsymbol{t}_i \in \mathbb{R}^{d_c}$ integrate local useful semantics and global discriminative semantics, promoting comprehensive feature perception in unordered point clouds and ambiguous texts.

C. Robust Negative Contrastive Learning

Inspired by [41], we leverage the complementary contrastive learning paradigm to learn with more reliable negative pairs instead of positive pairs, thereby mitigating the negative impact of mismatched pairs and achieving robust PTM against noisy correspondence. The loss for the cross-modal complementary learning paradigm is shown below:

$$\mathcal{L}' = \mathcal{L}'_{p\to t} + \mathcal{L}'_{t\to p}, \tag{9}$$

where

$$\mathcal{L}'_{p\to t} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{i,j})\log\left(1-S_{i,j}^{p\to t}\right), \tag{10}$$

$$\mathcal{L}'_{t\to p} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{i,j})\log\left(1-S_{i,j}^{t\to p}\right), \tag{11}$$

and

$$S_{i,j}^{p\to t} = \frac{\exp\left(\boldsymbol{p}_i^\top \boldsymbol{t}_j/\tau\right)}{\sum_k^K \exp\left(\boldsymbol{p}_i^\top \boldsymbol{t}_k/\tau\right)}, S_{i,j}^{t\to p} = \frac{\exp\left(\boldsymbol{t}_i^\top \boldsymbol{p}_j/\tau\right)}{\sum_k^K \exp\left(\boldsymbol{t}_i^\top \boldsymbol{p}_k/\tau\right)}, \tag{12}$$

where $\mathcal{L}'_{p\to t}/\mathcal{L}'_{t\to p}$ is the point-cloud-to-text/text-to-point-cloud complementary learning loss term, $S_{i,j}^{p\to t}/S_{i,j}^{t\to p}$ is the similarity between the i-th point-cloud/text sample and the j-th text/point-cloud sample, K is the batch size, τ is the temperature parameter, and $1-y_{i,j}$ is the flag, making the loss only apply to negative pairs. Minimizing Eq. (9) could reduce the similarity between the samples within negative pairs, introducing common discrimination without relying on positive pairs, which are more prone to containing some erroneous information. Because of this, the model could alleviate the impact of noisy correspondence.

However, due to the similarity in object categories within the point-cloud scenes and limited differences across some parts of scenes in PTM, samples within some negative pairs unavoidably exhibit certain degrees of semantic similarity. Blindly and monotonously amplifying the gap between two samples within negative pairs would lead to error accumulation, thus impacting the formation of robust discrimination. To address this issue, we propose the Robust Negative Contrastive loss, which could prevent the model from fitting these unreliable negative pairs or even revise the wrong optimization direction. This

TABLE I: Performance comparison on the SceneDepict-3D2T dataset in terms of R@1, R@5, R@10 and the sum of them. The left side of the table shows the results of adopting Bi-GRU as the text backbone, and the right side shows the results of using BERT. The highest results are shown in **bold** and the second highest results are underlined.

Method	Bi-GRU								BERT							
	Point cloud→Text				Text→Point cloud				Point cloud→Text				Text→Point cloud			
	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum	R@1	R@5	R@10	Sum
VSE	9.9	35.5	47.5	92.9	11.1	35.5	48.6	95.2	23.4	48.2	57.2	128.8	16.2	48.7	62.0	126.9
VSE++ [7]	14.9	36.2	49.6	100.7	11.8	36.3	50.1	98.2	21.3	43.3	56.7	121.3	17.2	47.3	62.6	127.1
VSE∞ [8]	35.3	61.6	75.9	172.8	27.2	62.1	76.3	165.6	39.0	62.4	73.8	175.2	29.4	63.2	77.1	169.7
SGR [11]	2.1	6.4	13.5	22.0	2.3	9.6	19.2	31.1	17.0	47.5	61.7	126.2	18.9	50.8	66.2	135.9
SGR+NCR [17]	5.8	17.1	37.3	60.2	7.0	23.0	43.8	73.8	29.8	58.2	65.2	153.2	21.5	56.4	71.8	149.7
SAF [11]	9.9	34.0	48.9	92.8	12.8	37.5	53.2	103.5	21.3	53.2	67.4	141.9	20.1	56.5	66.1	142.7
SAF+RCL [41]	17.6	44.0	60.3	121.9	17.1	44.3	59.9	121.3	31.9	57.4	70.9	160.2	22.1	55.6	71.4	149.1
MV-VSE [42]	10.6	31.2	48.9	90.7	8.9	28.1	39.4	76.4	34.0	57.4	63.8	155.2	22.3	53.2	67.6	143.1
NAAF [12]	17.7	41.1	52.5	111.3	10.9	31.6	45.0	87.5	17.0	44.0	59.6	120.6	13.8	44.3	63.1	121.2
ESA [43]	36.2	66.0	77.3	179.5	25.8	62.6	75.5	163.9	41.8	70.2	81.4	193.4	32.0	68.6	79.9	180.5
DIVE [19]	28.4	66.7	77.1	175.2	22.7	61.3	67.8	145.8	37.6	67.4	76.6	181.6	28.4	65.3	78.0	171.7
CHAN [20]	28.8	66.6	68.3	188.7	22.1	65.6	64.4	120.5	40.1	62.7	76.3	179.1	26.6	58.2	73.4	158.2
HREM [21]	34.0	59.6	69.5	163.1	27.7	64.8	78.0	170.5	39.0	**70.9**	81.3	191.2	31.5	68.7	81.2	181.4
CRCL [9]	35.8	67.8	76.0	179.6	28.8	64.7	77.1	170.6	41.8	64.5	80.1	186.4	30.5	64.8	77.4	172.7
Ours	**42.0**	**73.0**	**84.4**	**199.4**	**29.3**	**68.9**	**82.1**	**180.3**	**44.1**	**70.9**	**82.9**	**195.9**	**32.5**	**71.0**	**82.8**	**186.3**

novel loss is non-monotonic and has a parameter-controlled inflection point. It assesses the reliability of negative pairs based on the similarity of the paired samples, dynamically and implicitly divides negative pairs into clean and noisy subsets based on their reliability by considering the inflection point as a threshold, and assigns clean subsets with forward optimization direction but provides noisy subsets with reverse optimization direction, which could be formulated as:

$$\mathcal{L} = \mathcal{L}_{p\to t} + \mathcal{L}_{t\to p}, \tag{13}$$

where

$$\mathcal{L}_{p\to t} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{i,j})(1-S_{i,j}^{p\to t})^{\frac{1}{\alpha}}\log\left(1-S_{i,j}^{p\to t}\right), \tag{14}$$

$$\mathcal{L}_{t\to p} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{i,j})(1-S_{i,j}^{t\to p})^{\frac{1}{\alpha}}\log\left(1-S_{i,j}^{t\to p}\right). \tag{15}$$

Note that $\mathcal{L}_{p\to t}$ and $\mathcal{L}_{t\to p}$ are the point-cloud-to-text and text-to-point-cloud loss terms of our Robust Negative Contrastive loss respectively, and α is the parameter that controls the inflection point. Take $\mathcal{L}_{p\to t}$ for example, its gradient calculation formula can be written as $\partial\mathcal{L}_{p\to t}/\partial S_{i,j}^{p\to t} = -\frac{1}{\alpha}(1-S_{i,j}^{p\to t})^{\frac{1-\alpha}{\alpha}}[\log\big(1-S_{i,j}^{p\to t}\big)+\alpha]$. Consequently, we can infer that when $S_{i,j}^{p\to t} = 1-e^{1-\alpha}$, the loss has a inflection point. As optimization progresses, clean negative pairs with low similarity (i.e., $S_{i,j}^{p\to t} < 1-e^{1-\alpha}$) are still separated, while pairs with high similarity (i.e., $S_{i,j}^{p\to t} > 1-e^{1-\alpha}$) are identified and brought closer, helping our RNCL filter out unreliable negative pairs adaptively. Compared to the existing loss [44], [45], our loss identifies and handles negative pairs, adaptively driving the reliable negative pairs apart in the common space, enhancing robustness against noisy correspondence in PTM.

VI. Experiments

To thoroughly evaluate our RoMa for PTM, we conduct extensive experiments on the proposed SceneDepict-3D2T dataset and three other existing datasets.

A. Experimental Settings

In this work, our RoMa is implemented in PyTorch and carried out on one GeForce RTX 3090 GPU. In the experiments, we adopt the ScanNet [32] point-cloud set along with four description sets (i.e., SceneDepict-3D2T, ScanRefer [13], Nr3d [14], and 3D-LLM-Scene [15]), obtaining corresponding four multi-modal datasets for PTM evaluation. Due to the space restriction, the details of implementation and introduction to the adopted datasets could be found in the Appendix.

In the experiments, we compared our RoMa with 14 state-of-the-art cross-modal matching methods, including VSE, VSE++ [7], VSE∞ [8], SGR [11], NCR-SGR [17], SAF [11], RCL-SAF [41], MV-VSE [42], NAAF [12], DIVE [19], CHAN [20], ESA [43], HREM [21], and CRCL [9]. In the implementations and evaluations of all the methods, we adhere to the following settings. For point-cloud data processing, we adopt the widely used DGCNN [46] to obtain patch-level features. For text data processing, without loss of generality, we follow the text processing strategy used in cross-modal matching [8], [43] and employ both Bi-GRU [47] and BERT [48] to acquire word-level features, respectively. We follow [17], [29] to compute Recall at K (R@K) as the measurement of performance. In the experiments on SceneDepict-3D2T, we report R@1, R@5, R@10, and their sum to evaluate the performance of the methods. Due to the text discrimination limitations of other existing datasets (i.e., ScanRefer, Nr3d, and LLM-3D-Scene), we report R@5, R@30, and their sum to evaluate the methods more reasonably.

B. Comparison with the State-of-the-Arts

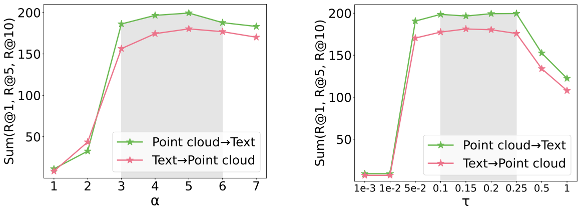
We conduct extensive PTM experiments on four datasets to evaluate the performance of our RoMa and the baselines. The experimental results are reported in Tables I and II. These results could yield the following observations: 1) General cross-modal matching methods exhibit inadequate performance. This substantiates the presence of distinct and more formidable challenges in PTM, indicating the difficulty of effectively applying these methods in PTM. 2) Some fine-grained methods

TABLE II: Performance comparison on the ScanRefer, Nr3d, and 3D-LLM-Scene datasets in terms of R@5, R@30 and the sum of them. The top of the table shows the results of adopting Bi-GRU as the text backbone, and the bottom shows the results of using BERT. The highest results are shown in **bold** and the second highest results are $\underline{\text{underlined}}$.

Method	ScanRefer						Nr3d						3D-LLM-Scene					
	Point→Text			Text→Point			Point→Text			Text→Point			Point→Text			Text→Point		
	R@5	R@30	Sum	R@5	R@30	Sum	R@5	R@30	Sum	R@5	R@30	Sum	R@5	R@30	Sum	R@5	R@30	Sum
Bi-GRU:																		
VSE∞ [8]	$\underline{38.3}$	77.3	115.6	$\underline{31.8}$	73.0	104.8	24.8	55.3	80.1	21.3	54.3	75.6	6.6	37.7	44.3	9.8	50.8	60.6
SAF+RCL [41]	29.8	68.1	97.9	28.0	71.8	99.8	19.9	52.5	72.4	19.8	56.7	76.5	$\underline{9.8}$	42.6	52.4	9.4	**57.5**	$\underline{66.9}$
ESA [43]	34.8	$\underline{80.9}$	115.7	31.6	71.5	103.1	$\underline{28.4}$	$\underline{58.9}$	$\underline{87.3}$	21.9	$\underline{58.6}$	$\underline{80.5}$	6.6	42.6	49.2	7.9	54.1	62.0
CHAN [20]	31.9	69.5	101.4	25.0	67.7	92.7	19.9	47.5	67.4	15.9	43.5	59.4	6.9	39.3	46.2	$\underline{11.5}$	51.6	63.1
HREM [21]	33.3	71.6	104.9	27.3	68.9	96.2	19.9	53.9	73.8	19.1	53.1	72.2	8.2	34.4	42.6	10.7	46.7	57.1
CRCL [9]	35.5	$\underline{80.9}$	$\underline{116.4}$	31.3	$\underline{75.4}$	$\underline{106.7}$	20.6	57.1	77.7	$\underline{22.6}$	53.6	76.2	6.6	$\underline{45.9}$	$\underline{52.5}$	9.0	$\underline{57.4}$	66.4
Ours	**41.8**	**82.3**	**124.1**	**33.1**	**76.0**	**109.1**	**31.9**	**61.7**	**93.6**	**26.0**	**61.3**	**87.3**	**15.7**	**47.5**	**63.2**	**16.2**	55.3	**71.5**
BERT:																		
VSE∞ [8]	$\underline{44.6}$	82.3	$\underline{126.9}$	33.3	76.5	109.8	$\underline{25.5}$	57.0	82.5	21.5	53.4	74.9	8.2	31.1	39.3	$\underline{9.8}$	52.5	62.3
SAF+RCL [41]	34.0	83.0	117.0	32.7	$\underline{78.9}$	111.6	18.4	56.0	74.4	18.0	$\underline{58.7}$	76.7	8.2	44.5	52.7	8.2	54.9	63.1
ESA [43]	43.4	83.3	126.7	**34.3**	76.7	111.0	24.1	48.9	73.0	20.3	53.1	73.4	6.6	$\underline{46.5}$	53.1	$\underline{9.8}$	**58.6**	$\underline{70.5}$
CHAN [20]	36.9	79.5	116.4	32.3	70.8	103.1	22.6	50.4	73.0	15.0	46.1	61.1	$\underline{13.1}$	45.8	$\underline{58.9}$	8.2	47.8	56.0
HREM [21]	40.4	$\underline{83.7}$	124.1	$\underline{34.0}$	77.9	$\underline{111.9}$	24.7	52.5	77.2	18.1	54.0	72.1	9.8	36.1	45.9	9.0	54.1	63.1
CRCL [9]	42.6	83.0	125.6	32.4	77.0	109.4	$\underline{25.5}$	$\underline{58.2}$	$\underline{83.7}$	$\underline{22.5}$	57.7	$\underline{80.2}$	7.6	39.3	46.9	9.0	55.7	64.7
Ours	**49.6**	**88.3**	**137.9**	32.8	**84.3**	**117.1**	**31.9**	**58.9**	**90.8**	**27.9**	**65.6**	**93.5**	**16.2**	**47.5**	**63.7**	**16.1**	$\underline{58.2}$	**74.3**

TABLE III: Ablation studies for our RoMa framework and DAP module adopted by our RoMa on the SceneDepict-3D2T dataset. ✓ stands for use. *w/o* stands for component removal.

Feat. Extraction			Loss			Point→Text		Text→Point		
GPO	ESA	DAP	$\mathcal{L}_c$	$\mathcal{L}'$	$\mathcal{L}$	R@1	R@5	R@1	R@5	Sum
					✓	1.4	9.9	1.1	7.6	20.0
		✓				9.7	36.5	8.2	37.4	91.8
✓			✓			35.3	61.6	27.2	62.1	186.4
✓				✓		36.8	64.5	27.4	64.1	192.8
✓					✓	37.2	66.0	27.4	65.9	196.5
	✓		✓			36.2	66.0	25.8	62.6	190.6
	✓			✓		39.7	67.4	26.7	66.0	199.8
	✓				✓	40.8	68.1	28.1	67.5	204.5
		✓	✓			37.4	59.6	27.4	63.2	187.6
		✓		✓		43.3	65.2	27.4	66.7	202.6
		w/o $\bar{A}$			✓	41.6	72.7	28.1	64.8	208.2
		w/o $\hat{A}$			✓	29.1	61.7	21.6	57.7	170.1
		w/o E			✓	37.6	72.0	28.5	68.0	206.1
		✓			✓	**42.0**	**73.0**	**29.3**	**68.9**	**213.2**



(a) Impact of different α.

(b) Impact of different τ.

Fig. 8: Performance of RoMa in terms of the sum of R@1, R@5, R@10 versus different values of α and τ on SceneDepict-3D2T. The gray box is the optimal choice range.

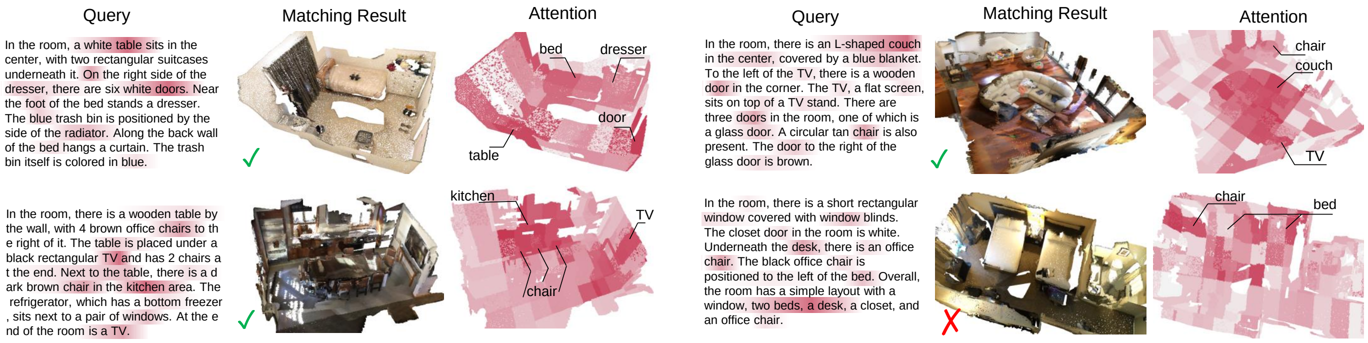
(e.g., SGR and SAF) suffered from severe performance issues in PTM. By combining these methods with robust modules, such as NCR-SGR and RCL-SAF, the performance could be remarkably improved. These results indicate that there is a large amount of noisy correspondence in PTM, which leads to the performance degradation of the non-robust methods. 3) Our RoMa achieves remarkably better results than the existing cross-modal matching methods (e.g., CRCL, HERM, etc.), demonstrating its superior effectiveness by conquering the two challenges in PTM. 4) In existing datasets, the relevant results matched by most methods rank only outside the top 5, proving the scene-specific discrimination of these datasets is limited. 5) The performance on SceneDepict-3D2T is relatively low, compared to the existing Image-Text datasets, where the state-of-the-art performance usually exceeds 80 [21], [43], in terms of R@1. This indicates that the PTM task still faces difficulties in handling unordered point clouds, vague texts, and noisy correspondence, and calls for more advanced solutions.

C. Ablation Study

In this section, we conduct an ablation study to investigate the contribution of each proposed component to PTM. Firstly, we replace the DAP module with the GPO [7] and ESA [43] feature extraction modules, and the Robust Negative Contrastive loss (i.e., $\mathcal{L}$) adopted by RNCL with the vanilla loss adopted by complementary contrastive learning paradigm (i.e., $\mathcal{L}'$) [41] and Contrastive loss (i.e., $\mathcal{L}_c$). In addition, we alternately replace patch position embedding E, token-level attention $\bar{A}$, and feature-level attention $\hat{A}$ to fairly verify their effectiveness under the premise of eliminating the influence of the number of learnable parameters. All the comparisons are conducted on SceneDepict-3D2T with the same experimental settings. The results are presented in Table III. From the table, we could draw the following observation: 1) RoMa without any component will drop matching performance, which indicates that each component contributes to our method. 2) The performances of adopting the $\mathcal{L}$ are superior to $\mathcal{L}_c$ that is widely applied in well-annotated scenarios and $\mathcal{L}'$. This proves the presence of a considerable amount of noisy correspondence in PTM and the $\mathcal{L}$ adopted by RNCL contributes to the enhanced robustness of our RoMa. 3) DAP without any one of attention and positional embedding will decrease matching performance, demonstrating that each component in DAP contributes to the comprehensive perception of features.

D. Parameter Analysis

To investigate the sensitivity of our RoMa to various parameters, we plot the sum of R@1, R@5, and R@10 of



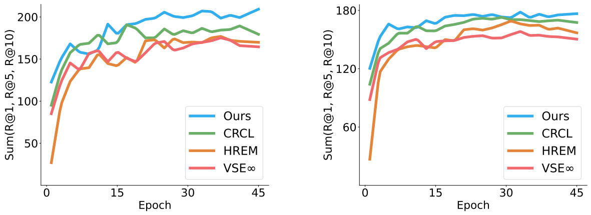
In the room, a white table sits in the
center, with two rectangular suitcases
underneath it. On the right side of the
dresser, there are six white doors. Near
the foot of the bed stands a dresser.
The blue trash bin is positioned by the
side of the radiator. Along the back wall
of the bed hangs a curtain. The trash
bin itself is colored in blue.

In the room, there is a wooden table by
the wall, with 4 brown office chairs to th
e right of it. The table is placed under a
black rectangular TV and has 2 chairs a
t the end. Next to the table, there is a d
ark brown chair in the kitchen area. The
refrigerator, which has a bottom freezer
, sits next to a pair of windows. At the e
nd of the room is a TV.

In the room, there is an L-shaped couch
in the center, covered by a blue blanket.
To the left of the TV, there is a wooden
door in the corner. The TV, a flat screen,
sits on top of a TV stand. There are
three doors in the room, one of which is
a glass door. A circular tan chair is also
present. The door to the right of the
glass door is brown.

In the room, there is a short rectangular
window covered with window blinds.
The closet door in the room is white.
Underneath the desk, there is an office
chair. The black office chair is
positioned to the left of the bed. Overall,
the room has a simple layout with a
window, two beds, a desk, a closet, and
an office chair.

Fig. 9: Some matched instances of PTM on SceneDepict-3D2T. For each text query, the top-1 ranked point cloud are displayed. The correctly matched point clouds are marked with a green tick, otherwise the red cross. In addition, we visualize the text after applying attention and present a comparison between the original point clouds and the point cloud after applying attention. **Redder** words and patches indicate higher attention weights.



(a) Point cloud→Text

(b) Text→Point cloud

Fig. 10: The performance of VSE∞, HREM, CRCL, and our RoMa on the SceneDepict-3D2T dataset.

PTM against different hyper-parameters (i.e., α, and τ) on SceneDepict-3D2T as shown in Fig. 8. The results indicate that our method achieves superior matching performance across a range of α and τ values. Notably, when α is set too low, the threshold for distinguishing between clean and noisy negative pairs becomes excessively small. This leads to a significant number of negative pairs being misclassified as noise and subjected to reverse optimization, resulting in remarkable performance degradation. Conversely, if α is too high, the RNCL struggles to differentiate potential noisy negative pairs, causing error accumulation and degraded performance.

E. Visualization Analysis

To provide a comprehensive insight into the effectiveness exhibited by our RoMa, we conduct visualization experiments in PTM. Firstly, to shed light on the reasons behind the superior performance of our RoMa, we illustrate a small handful of matching results and token-level attention visualization throughout the point clouds and texts on SceneDepict-3D2T dataset, as shown in Fig. 9. Additionally, we present a performance comparison among our RoMa and the VSE∞ [8], HREM [21], and CRCL [9] throughout the training process, as shown in Fig. 10. From the results, we could draw the following observations: 1) Our RoMa can achieve correct retrieved results in PTM. Even the mismatched pair still exhibits a strong cross-modal semantic correlation. This is attributed to our DAP, which actually focuses on useful and discriminative patches and words. 2) Throughout the whole learning process, it is evident that non-robust baselines (i.e., VSE∞ and HREM) involve performance degradation in the later training stage, impacted by the noisy correspondence. In contrast, our RoMa mitigates the negative impact comprehensively, achieving superior and robust performance.

VII. Conclusion

In this paper, we introduce a novel yet challenging task, named PointCloud-Text Matching (PTM). To facilitate the research on this promising task, we construct a benchmark dataset, namely SceneDepict-3D2T. We also propose a robust baseline, named **Ro**bust PointCloud-Text **Ma**tching method (RoMa), which consists of two novel modules: Dual Attention Perception module (DAP) and Robust Negative Contrastive Learning module (RNCL). Specifically, DAP leverages dual attention mechanisms to capture local and global features of point clouds and texts. In addition, RNCL is employed to handle noisy correspondence by distinguishing and endowing clean and noisy negative pairs with correct optimization directions. We conducted extensive experiments compared to 14 state-of-the-art methods on four datasets, demonstrating the superiority of our RoMa in the challenging PTM task.

References

[1] Y. Li, L. Ma, Z. Zhong, F. Liu, M. A. Chapman, D. Cao, and J. Li, "Deep learning for lidar point clouds in autonomous driving: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3412–3432, 2020.

[2] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.

[3] Q. Xu, Y. Zhou, W. Wang, C. R. Qi, and D. Anguelov, "Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15446–15456, 2021.

[4] G. Wang, H. Fan, and M. Kankanhalli, "Text to point cloud localization with relation-enhanced transformer," *arXiv preprint arXiv:2301.05372*, 2023.

[5] R. Huang, X. Pan, H. Zheng, H. Jiang, Z. Xie, C. Wu, S. Song, and G. Huang, "Joint representation learning for text and 3d point cloud," *Pattern Recognition*, p. 110086, 2023.

[6] C. Tang, X. Yang, B. Wu, Z. Han, and Y. Chang, "Parts2words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6884–6893, 2023.

[7] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[8] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15789–15798, 2021.

[9] Y. Qin, Y. Sun, D. Peng, J. T. Zhou, X. Peng, and P. Hu, "Cross-modal active complementary learning with self-refining correspondence," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *Proceedings of the 27th ACM international conference on multimedia*, pp. 3–11, 2019.

[11] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 1218–1226, 2021.

[12] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15661–15670, 2022.

[13] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European conference on computer vision*, pp. 202–221, Springer, 2020.

[14] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. 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*, pp. 422–440, Springer, 2020.

[15] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024.

[16] W. Liu, J. Sun, W. Li, T. Hu, and P. Wang, "Deep learning on point clouds and its application: A survey," *Sensors*, vol. 19, no. 19, p. 4188, 2019.

[17] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 29406–29419, Curran Associates, Inc., 2021.

[18] Z. Feng, Z. Zeng, C. Guo, Z. Li, and L. Hu, "Learning from noisy correspondence with tri-partition for cross-modal matching," *IEEE Transactions on Multimedia*, 2023.

[19] D. Kim, N. Kim, and S. Kwak, "Improving cross-modal retrieval with set of diverse embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23422–23431, 2023.

[20] Z. Pan, F. Wu, and B. 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*, pp. 19275–19284, 2023.

[21] Z. Fu, Z. Mao, Y. Song, and Y. Zhang, "Learning semantic relationship among instances for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15159–15168, June 2023.

[22] Y. Li, Y. Qin, Y. Sun, D. Peng, X. Peng, and P. Hu, "Romo: Robust unsupervised multimodal learning with noisy pseudo labels," *IEEE Transactions on Image Processing*, 2024.

[23] Y. Feng, H. Zhu, D. Peng, X. Peng, and P. 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*, pp. 11610–11619, 2023.

[24] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji, "Discover cross-modality nuances for visible-infrared person reidentification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4330–4339, 2021.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking.," *Journal of Machine Learning Research*, vol. 11, no. 3, 2010.

[27] C. Chen, D. Wang, B. Song, and H. Tan, "Inter-intra modal representation augmentation with dct-transformer adversarial network for image-text matching," *IEEE Transactions on Multimedia*, vol. 25, pp. 8933–8945, 2023.

[28] Y. Yang, L. Wang, E. Yang, and C. Deng, "Robust noisy correspondence learning with equivariant similarity consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17700–17709, 2024.

[29] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 201–216, 2018.

[30] P. Achlioptas, J. Fan, R. Hawkins, N. Goodman, and L. J. Guibas, "Shapeglot: Learning language for shape differentiation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8938–8947, 2019.

[31] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese, "Text2shape: Generating shapes from natural language by learning joint embeddings," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 100–116, Springer, 2019.

[32] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

[33] Z. Yuan, X. Yan, Y. Liao, R. Zhang, S. Wang, Z. Li, and S. Cui, "Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1791–1800, 2021.

[34] D. He, Y. Zhao, J. Luo, T. Hui, S. Huang, A. Zhang, and S. Liu, "Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2344–2352, 2021.

[35] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2cap: Context-aware dense captioning in rgb-d scans," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3193–3203, 2021.

[36] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-aware alignment and mutual masking for 3d-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10984–10994, 2023.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[38] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

[39] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19129–19139, 2022.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[41] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-modal retrieval with partially mismatched pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2023.

[42] Z. Li, C. Guo, Z. Feng, J.-N. Hwang, and X. Xue, "Multi-view visual semantic embedding," in *IJCAI*, vol. 2, p. 7, 2022.

[43] H. Zhu, C. Zhang, Y. Wei, S. Huang, and Y. Zhao, "Esa: External space attention aggregation for image-text retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[44] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2980–2988, 2017.

[45] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[46] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[47] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.