PointCloud-Text Matching: Benchmark Datasets and a Baseline

Yanglin Feng[1] Yang Qin[1] Dezhong Peng[1] Hongyuan Zhu[2] Xi Peng[1] Peng Hu[1]*
[1]College of Computer Science, Sichuan University [2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

Abstract

*In this paper, we present and study a new instance-level retrieval task: PointCloud-Text Matching (PTM), which aims to find the exact cross-modal instance that matches a given point-cloud query or text query. PTM could be applied to various scenarios, such as indoor/urban-canyon localization and scene retrieval. However, there exists no suitable and targeted dataset for PTM in practice. Therefore, we construct three new PTM benchmark datasets, namely 3D2T-SR, 3D2T-NR, and 3D2T-QA. We observe that the data is challenging and with noisy correspondence due to the sparsity, noise, or disorder of point clouds and the ambiguity, vagueness, or incompleteness of texts, which make existing cross-modal matching methods ineffective for PTM. To tackle these challenges, we propose a PTM baseline, named **Ro**bust PointCloud-Text **Ma**tching method (RoMa). RoMa consists of two modules: a Dual Attention Perception module (DAP) and a Robust Negative Contrastive Learning module (RNCL). Specifically, DAP leverages token-level and feature-level attention 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 divides negative pairs, which are much less error-prone than positive pairs, into clean and noisy subsets, and assigns them forward and reverse optimization directions respectively, thus enhancing robustness against noisy correspondence. We conduct extensive experiments on our benchmarks and demonstrate the superiority of our RoMa.*

1. Introduction

Point clouds are a popular representation of the 3D geometry of a scene, which has important applications in computer vision, robotics, and augmented reality. For example, point clouds can be used for autonomous driving [11, 29], object detection [44], and localization [40]. However, as the volume of point-cloud data grows 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.

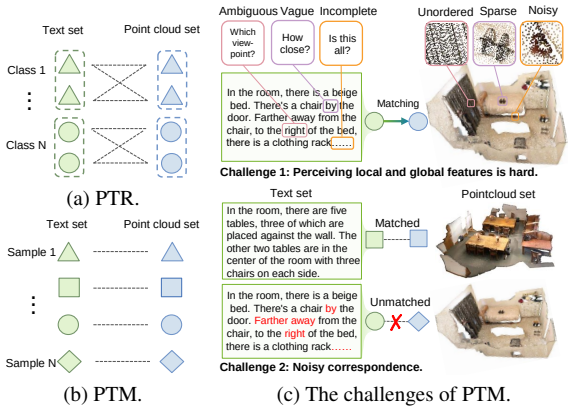


Figure 1. Overview for PointCloud-Text Matching (PTM). (a) and (b) show the schematic illustrations of class-level PointCloud-Text Retrieval (PTR), and instance-level PTR (*i.e.*, PTM), respectively. (c) illustrates the challenges faced by PTM.

However, existing methods lack pertinence and struggle to tackle PTM. On one hand, current PointCloud-Text Retrieval (PTR) methods [22, 38] can only focus on establishing a category-level correspondence between 3D point-cloud shapes and concise annotation texts as shown in Figure 1 (a). In contrast, PTM requires to exploit the mutual information of cross-modal pairs, and achieves instance-level alignment between point clouds and detailed descriptions as shown in Figure 1 (b). This indicates that PTM requires deeper capability to capture local features and instance discrimination rendering those methods entirely inapplicable. On the other hand, existing cross-modal matching works that can build instance-level cross-modal correspondences are only primarily oriented to text and 2D image modalities. According to the granularity of the established correspondences, these Image-Text Matching (ITM) works could

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

be divided into two groups: coarse-grained and fine-grained matching methods. The former [6, 17, 37] use global features to represent the whole image and the whole text, while the latter [15, 32, 48] use local features to capture the fine-grained details of regions and words. Although these methods have achieved promising performance for ITM, most of them ignore the specific properties and challenges in PTM.

To the best of our knowledge, in existing multi-modal datasets of point clouds and texts (*i.e.* ScanRefer [5], Referit3d [2], and ScanQA [3]), one description primarily focuses on describing a single point-cloud object within the corresponding scenes for visual grounding, rather than matching all objects inside the scene in PTM. These limited descriptions can hardly match the corresponding wide-field point clouds, as demonstrated by the dismal matching performance in existing datasets depicted in Figure 2. Therefore, we constitute three new benchmark datasets for PTM, *i.e.*, 3D2T-SR, 3D2T-NR, and 3D2T-QA. These datasets contain comprehensive descriptions covering entire scenes, so they evaluate baselines more reliably and reasonably for PTM, which can be observed in Figure 2. We also provide a comprehensive evaluation protocol and several benchmark results for PTM on the datasets as shown in Tables 2 and 3. From the results, we observe that point cloud-text data are more challenging than image-text data due to the sparsity, noise, or disorder of point clouds [33]. 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 [23], thus degrading the retrieval performance. The schematic illustration of the challenges is shown in Figure 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 during cross-modal learning, 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 shown in Figure 5. 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 multigrainly aggregate the local and global discriminative features into common representations, thus

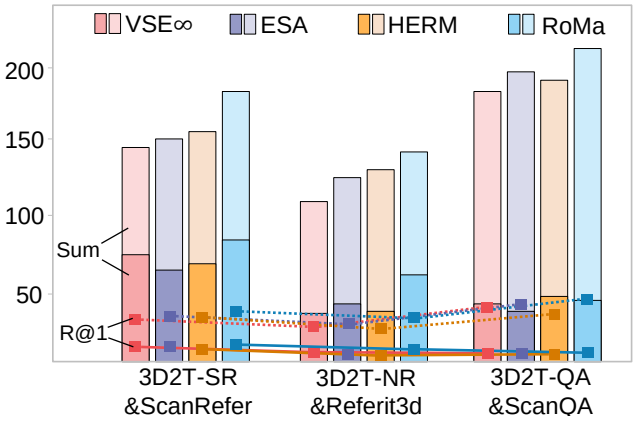


Figure 2. The PTM performance of VSE∞ [6], ESA [49], HREM [19] and RoMa on the existing ScanRefer, Referit3d, and ScanQA and proposed 3D2T-SR, 3D2T-NR, and 3D2T-QA. **Dark** and *light* bar graphs show in terms of the average of (R@1+R@5+R@10) of baselines on existing and proposed datasets respectively. The **solid** line and *dashed* line graphs show in terms of the average of R@1 on existing and proposed datasets.

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 novelty and contributions are as follows:

- We investigate a new instance-level cross-modal retrieval task, namely Point-Cloud-Text Matching (PTM), and propose three PTM benchmark datasets, *i.e.*, 3D2T-SR, 3D2T-NR, and 3D2T-QA, and a robust baseline RoMa to learn from challenging multimodal 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 the three proposed datasets. Our RoMa remarkably outperforms the existing methods without bells and whistles,

demonstrating its superiority over existing methods.

2. Related Work

2.1. Cross-modal Retrieval

Cross-modal retrieval aims to search the relevant results across different modalities for a given query, *e.g.*, image-text matching [6, 15], video-text retrieval [16], 2D-3D retrieval [26, 31], and visible-infrared re-identification [43, 45], etc. Most of these works learn a joint common embedding space by applying cross-modal constraints [4, 34, 42], 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 [6, 17, 19] typically learns shared subspaces to build connections between global-level representations, which align images and texts in a direct manner. 2) Fine-grained retrieval [15, 28, 35] 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.*, PTM, which argues for building cross-modal associations between 3D space and textual space.

2.2. 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 [1] or detecting individual objects [7]. With the introduction of the ScanNet [12] ScanRefer [5] and ReferIt3D [2] datasets, more works have expanded their focus to encompass the 3D scenes. Some existing works [18, 20, 46] have attempted to locate objects within scenes based on linguistic descriptions, completing the task of 3D visual grounding. Recently, with the introduction of Scan2Cap [9], some efforts [8, 24, 25, 47] focus on providing descriptions for objects about their placement. This is also known as 3D dense captioning. More recently, a few solutions [22, 38] for PTR have begun to emerge, which aim to establish common discrimination for coarse category-level alignment between point-cloud shapes and brief label texts. However, 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.

3. PointCloud-Text Matching

In this paper, we introduce and explore 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 [13] and 3%-20% annotation errors in the Conceptual Captions [36]. 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 correspondences). Such noise would lead to insufficient learning or noise overfitting for existing models.

4. Benchmark Datasets

To the best of our knowledge, the descriptions in existing multi-modal datasets of point clouds and texts (*i.e.*, ScanRefer [5], Referit3d [2], and ScanQA [3]) are confined to single objects of the entire point-cloud scene scans. Figure 4 shows they only have average lengths of fewer than 20 words and each description encompasses fewer than 2 objects, covering less than 10% objects in one scan. However, each scan typically contains 10-30 objects [12], with many similar objects present across different scans. This indicates that the short and inadequately informative descriptions are prone to be ambiguous, lacking the discrimination to meet the requirements of PTM. We conduct PTM experiments on existing datasets, and the matching results in Figure 2 show that all performance in terms of Recall at 1 is less than 10%, confirming the above considerations. Therefore, we constructed three PTM datasets with comprehensive descriptions, namely 3D2T-SR, 3D2T-NR, and 3D2T-QA, where the point-cloud data is all based on the ScanNet [12] point-cloud dataset and the text data derived from the description sets associated with ScanNet, ScanRefer, Referit3d, and ScanQA. In the following sections, we will elaborate on the collection and statistics of our proposed datasets.

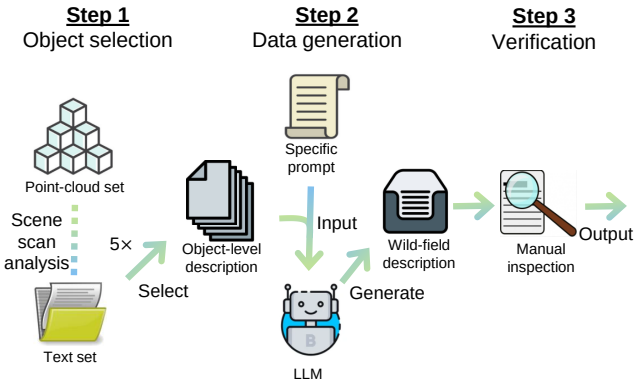


Figure 3. Pipeline of the data collection process.

4.1. Data Collection

We deploy a prompt + Large Language Model (LLM) paradigm to generate wide-field descriptions of point-cloud scans in ScanNet, based on three existing object-level description datasets. The pipeline is shown in Figure 3. More specifically, we first randomly collect n description texts of n spatially related objects in different neighborhoods within the same point-cloud scan in each of the three datasets. This ensures the generated descriptions comprehensively encapsulate the entirety of the point-cloud scans while maximizing their discrimination. Then, we input the text collections into LLM with the dataset-specific manual prompt to obtain the descriptions corresponding to the point-cloud scan, respectively. Repeating the above process five times, we obtain five descriptions corresponding to the whole point-cloud scan. By following this process for all point clouds, we obtain the initial version of the datasets. Finally, by systematically assessing the discriminative accuracy, grammatical correctness, and coherence of every description, we complete the dataset construction.

To enhance the applicability and intricacy of the PTM within various scenarios, considering the linguistic style of aforementioned text datasets, we employ different configurations of n and prompt settings, meticulously curating three unique datasets with different characteristics, which are shown as follows:

- ***3D2T-SR*** is based on ScanRefer, in which exhaustive and coherent text paragraphs delineate the placement details of objects throughout the scans.
- ***3D2T-NR*** is based on Referit3d, in which concise and informative text only depicts a partial arrangement of objects within the scan.
- ***3D2T-QA*** is based on ScanQA, in which detailed texts emphasize the object characteristics and inter-object relationships.

Note that more construction details and data examples of each dataset are put into our Complementary Materials due to space limitations.

Table 1. Statistics of the proposed datasets.

	3D2T-SR	3D2T-NR	3D2T-QA
Point-cloud scan number	703	641	633
Object number	50,061	25,666	25,871
Object number/description	14.2	8	8.17
Description number	3,515	3,205	3,165
Description number/scan	5	5	5
Vocabulary size	1,843	1,507	1,287
Avg. description length	152.5	84	84.2

4.2. Dataset Statistics

Table 1 shows the statistics of the three proposed benchmarks and Figure 4 depicts the statistics comparison of the proposed datasets with existing ones, demonstrating the descriptions we generated are comprehensive and cover a wider range of point clouds. This is reflected in the observation in Figure 2 that the proposed datasets achieve 100%-400% PTM performance gain compared to the existing datasets, indicating that our datasets build better cross-modal alignment for PTM. Additionally, The specificity of each dataset can be observed. 3D2T-SR tends to longer texts to describe placement details with an average description length of 152.5, which is 8.5 times longer than ScanRefer (17.9). 3D2T-NR prefers to encapsulate rich information in concise language. On average, each short description involves 8 objects, which is 4.7 times more than Referit3d (1.7). 3D2T-QA leans towards explaining object features and inter-object spacial relationships, utilizing massive of color (93.8%), material/shape terms (73.2%), and spatial language (99.0%) descriptions.

Although we adopt meticulous verification to improve the grammatical correctness and syntactic coherence of the datasets, it is unavoidable to introduce a considerable portion of noisy correspondence because of the unordered point-cloud scans and vague free-form descriptions. We attempt to use well-trained models and the Gaussian Mixture Module (GMM) to estimate it. The results show that there are 11.9%, 13.2%, and 13.8% data pairs with noisy correspondences in 3D2T-SR, 3D2T-NR, and 3D2T-QA, respectively. Thus, noisy correspondence is an unavoidable problem in PTM, which would cause noise overfitting, thereby leading to performance degradation.

5. Robust Baseline: RoMa

5.1. 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

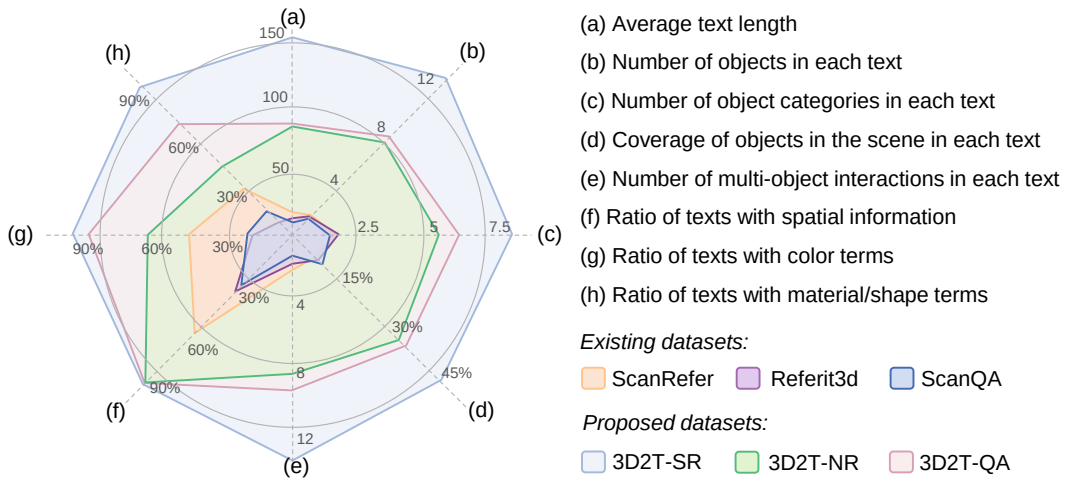


Figure 4. Statistics comparison among existing ScanRefer [5], Referit3d [2], ScanQA [3], and proposed 3D2T-SR, 3D2T-NR, 3D2T-QA dataset benchmarks.

pairwise correspondence between $\mathcal{P}$ and $\mathcal{T}$, so $\mathcal{D}$ can also be written as $\mathcal{D} = \{(X_i^p, X_j^t), y_{ij}\}_{i,j}^{N_p,N_t}$, $y_{ij} \in \{0,1\}$ indicates whether X_i^p and X_j^t are matched (*i.e.*, positive pair, $y_{ij} = 1$) or unmatched (*i.e.*, negative pair, $y_{ij} = 0$). However, in practice, the unmatched pairs ($y_{ij} = 0$) may be mislabeled as matched ones ($y_{ij} = 1$), *a.k.a* noisy correspondence.

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. The proposed method involves two modules: 1) Dual Attention Perception (DAP) is used to comprehensively perceive semantic features with dual attention, and 2) Robust Negative Contrastive Learning (RNCL) is exploited to handle noisy correspondences. In the following sections, we will elaborate on each component of RoMa.

5.2. Dual Attention Perception

We first employ modality-specific backbones (f_p and f_t) to extract token-level features for the patches of point clouds and words of textual descriptions, *i.e.*, $Z_i^p = f_p(X_i^p) \in \mathbb{R}^{p_n \times d_c}$ and $Z_i^t = f_t(X_i^t) \in \mathbb{R}^{t_n \times d_c}$, respectively. p_n and t_n are the number of token-level features for each sample and d_c is the dimensionality of the feature space.

To tackle the challenge of capturing and integrating local and global semantic features from point clouds and texts in PTM, in analogy to the ¡*Query-Key-Value*¿ definition in Self-Attention mechanism (SA) [39], we explore a novel dual attention manner. More specifically, the *Queries* $Q_i^p \in \mathbb{R}^{p_n \times d_c}$, $Q_i^t \in \mathbb{R}^{t_n \times d_c}$ and the *Values* $V_i^p \in \mathbb{R}^{p_n \times d_c}$, $V_i^t \in \mathbb{R}^{t_n \times d_c}$ of two modalities are calculated from point-cloud and word features Z_i^p and Z_i^t through the fully connected layers g_p and g_t, respectively. Different from SA, we construct general and learnable token-level and feature-level *Keys* at the dataset level, which are not restricted to combining with *Queries* to capture token-wise interaction exploring self-attention within a sample. The general *Keys* learn to model general patterns in the dataset at token and feature levels and explore comprehensive attention by integration with the sample-specific *Queries*.

To facilitate the adaptive exploration of local semantic features, we construct learnable token-level general *Keys* $\bar{K}^p$ and $\bar{K}^t$ for two 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 vectors by measuring the token-wise similarity between *Queries* and token-level *Keys*, empowering the model to selectively focus on local key semantic units (*e.g.*, foreground patches in the point clouds and keywords in the texts) similar to the common patterns in the two modalities, which are written as:

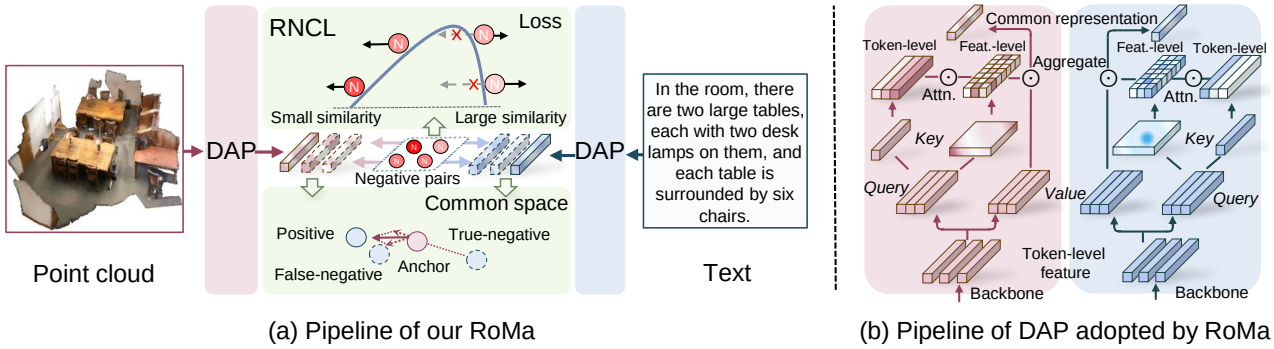
$$\bar{a}_i^p = \sigma_t(Q_i^p \bar{K}^{p\top}), \quad \bar{a}_i^t = \sigma_t(Q_i^t \bar{K}^{t\top}), \tag{1}$$

where $\bar{a}_i^p \in \mathbb{R}^{p_n}$ and $\bar{a}_i^t \in \mathbb{R}^{t_n}$ are the token-level attention vectors, σ_t is token-level softmax. Based on this, we obtain the token-level attention on feature matrices by stacking these attention vectors as follows:

$$\bar{A}_i^p = [\bar{a}_i^p, \cdots, \bar{a}_i^p], \quad \bar{A}_i^t = [\bar{a}_i^t, \cdots, \bar{a}_i^t], \tag{2}$$

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

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 general *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 *Keys* in a feature-level manner to grasp global discriminative features from the dimensional interrelationships in



(a) Pipeline of our RoMa (b) Pipeline of DAP adopted by RoMa

Figure 5. The illustration of our proposed method. (a) is the pipeline of our RoMa, which involves two modules: Dual Attention Perception (DAP) and Robust Negative Contrastive Learning (RNCL). In DAP, comprehensive Common Representations (CRs) could be extracted from both modalities. Then, CRs are matched into negative pairs. In RNCL, negative pairs are adaptively provided with forward and reverse optimization directions based on the similarity within pairs, leading to robust discrimination in the common space. (b) is the schematic illustration of DAP. *Queries* and *Values* are obtained from multimodal features, but *Keys* are general and learnable for the whole dataset. *Queries* are combined with token-level and feature-level *Keys* obtaining dual attention. Following this, features and each attention are aggregated into CRs.

the feature space, such as distinctive object color, position, orientation, spatial relationships, etc., which is written as:

$$\hat{A}_i^p = \sigma_f(Q_i^p \hat{K}^{p\top}), \quad \hat{A}_i^t = \sigma_f(Q_i^t \hat{K}^{t\top}), \tag{3}$$

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, σ_f is the feature-level softmax.

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

$$A_i^p = \bar{A}_i^p \odot \hat{A}_i^p, \quad A_i^t = \bar{A}_i^t \odot \hat{A}_i^t, \tag{4}$$

where $\odot$ 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:

$$\boldsymbol{p}_i = L2Norm(\frac{1}{p_n}\sum_j^{p_n}(A_{ij}^p \odot V_{ij}^p)), \tag{5}$$

$$\boldsymbol{t}_i = L2Norm(\frac{1}{t_n}\sum_j^{t_n}(A_{ij}^t \odot V_{ij}^t)), \tag{6}$$

where A_{ij}^p and A_{ij}^t are the j-th row of dual attention A_i^p and A_i^t, V_{ij}^p and V_{ij}^t are the j-th row of the *Values* V_i^p and V_i^t, and $L2Norm(\cdot)$ 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.

5.3. Robust Negative Contrastive Learning

Inspired by [21], 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 robustness against noisy correspondences. The loss for the cross-modal complementary learning paradigm is shown below:

$$\mathcal{L}' = \mathcal{L}'_{p\to t} + \mathcal{L}'_{t\to p}, \tag{7}$$

where

$$\mathcal{L}'_{p\to t} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{ij})\log(1-S_{ij}^{p\to t}), \tag{8}$$

$$\mathcal{L}'_{t\to p} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{ij})\log(1-S_{ij}^{t\to p}), \tag{9}$$

and

$$S_{ij}^{p\to t} = \frac{\exp(\boldsymbol{p}_i^\top \boldsymbol{t}_j/\tau)}{\sum_k^K \exp(\boldsymbol{p}_i^\top \boldsymbol{t}_k/\tau)}, S_{ij}^{t\to p} = \frac{\exp(\boldsymbol{t}_i^\top \boldsymbol{p}_j/\tau)}{\sum_k^K \exp(\boldsymbol{t}_i^\top \boldsymbol{p}_k/\tau)}, \tag{10}$$

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, $S_{ij}^{p\to t}/S_{ij}^{t\to p}$ 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 $1-y_{ij}$ is the flag, making the loss only apply to negative pairs. Minimizing Equation (7) 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, samples within some of the negative pairs unavoidably exhibit certain degrees of semantic similarity, even though negative pairs are less prone to noise. Blindly and monotonously amplifying the gap between two samples within negative pairs would lead to error accumula-

tion, thus impacting the formation of robust discrimination. To address this issue, we propose the Robust Negative Contrastive Loss (RNCL), which could prevent the model from fitting these unreliable negative pairs or even revise the wrong optimization direction. This 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:

$$\mathcal{L} = \mathcal{L}_{p\to t} + \mathcal{L}_{t\to p}, \tag{11}$$

where

$$\mathcal{L}_{p\to t} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{ij})(1-S_{ij}^{p\to t})^{\frac{1}{\alpha}}\log\left(1-S_{ij}^{p\to t}\right), \tag{12}$$

$$\mathcal{L}_{t\to p} = -\frac{1}{K}\sum_{i,j}^{K}(1-y_{ij})(1-S_{ij}^{t\to p})^{\frac{1}{\alpha}}\log\left(1-S_{ij}^{t\to p}\right). \tag{13}$$

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 RNCL respectively, and α is the parameter that controls the inflection point and helps our RNCL identify and filter out unreliable negative pairs adaptively. Our RoMa could be optimized by minimizing Equation (11), which could adaptively drive the reliable negative pairs apart in the common space, enhancing robustness against noisy correspondences.

6. Experiments

To evaluate our RoMa, we conduct extensive comparison experiments on three PTM datasets, *i.e.*, 3D2T-SR, 3D2T-NR, and 3D2T-QA.

6.1. Experimental Settings

The implementation details of our RoMa could be found in our Complementary Materials. The code and datasets will be released upon paper acceptance. All reported quantitative results represent the averages obtained from ten runs for all methods. Specific details of the datasets are presented in Section 4.

In the experiments, we compared our RoMa with 13 state-of-the-art Image-Text Matching (ITM) methods, including VSE, VSE++ [17], VSE∞ [6], SGR [15], NCR-SGR [23], SAF [15], RCL-SAF [21], MV-VSE [30], NAAF [48], DIVE [27], CHAN [35], ESA [49], and HREM [19]. In the implementations and evaluations of all the methods, we adhere to the following settings. For data processing, without loss of generality, we adopt the widely used DGCNN [41] to obtain patch-level features for point clouds and employ both Bi-GRU [10] and BERT [14] to acquire word-level features for the texts. We follow [23, 28] to compute Recall at K (R@K) as the measurement of performance. In our experiments, we report R@1, R@5, R@10, and their sum to evaluate the performance of the methods.

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

We conduct extensive PTM experiments on three datasets to evaluate the performance of our RoMa and the baselines. The experimental results are reported in Table 2 and Table 3. These results could yield the following observations:

- ITM methods exhibit inadequate performance. This substantiates the presence of distinct and more formidable challenges in PTM, indicating the difficulty of effectively applying ITM methods in PTM.
- Some fine-grained methods (*e.g.*, SGR and SAF) suffered from severe performance issues in PTM, or they could not even fit the data well. 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 correspondences in these datasets, which leads to the performance degradation of the non-robust methods.
- Our RoMa achieves remarkably better results than the embedding-based coarse-grained methods (*e.g.*, VSE∞, ESA, *etc.*) and fine-grained methods (*e.g.*, NAAF, CHAN, *etc.*), which are used in ITM, demonstrating its superior effectiveness by conquering the two challenges in PTM.
- The performance on PTM datasets is relatively low, compared to the existing ITM datasets, where the state-of-the-art performance usually exceeds 80 [19, 49], in terms of R@1. This indicates that the PTM task still faces difficulties in handling unordered point clouds, vague texts, and noisy correspondences, and calls for more advanced solutions.

6.3. 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 [17] and ESA [49] 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.*, $\mathcal{L}'$) and Contrastive loss (*i.e.*, $\mathcal{L}_c$). In addition, we alternately replace token-level attention $\bar{A}$ and feature-level attention $\hat{A}$ to fairly verify the effectiveness of $\bar{A}$ and $\hat{A}$ under the premise of eliminating the influence of the number of learnable parameters. All the comparisons are conducted on 3D2T-QA with the same experimental settings. The results are presented in Table 4. From the table, we could draw the following observation: 1) RoMa with-

Table 2. Performance comparison on the three datasets in terms of R@1, R@5, R@10 and the sum of them. All the methods use Bi-GRU as the text backbone. The highest results are shown in **bold** and the second highest results are underlined.

Method	3D2T-SR							3D2T-NR							3D2T-QA						
	Point→Text			Text→Point				Point→Text			Text→Point				Point→Text			Text→Point			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
VSE	8.2	22.0	34.8	7.7	23.7	37.6	134.0	8.5	24.6	37.7	7.4	24.3	38.1	148.6	15.5	42.3	54.9	16.9	25.6	63.1	218.3
VSE++ [17]	12.1	24.8	34.0	11.1	24.5	38.3	144.8	8.5	29.2	43.1	9.8	34.0	48.9	173.5	20.3	47.3	60.6	18.3	51.2	61.2	258.9
VSE∞ [6]	26.7	55.1	63.3	24.6	55.7	64.9	290.3	16.9	38.5	50.8	17.4	42.2	55.6	221.4	38.5	66.4	76.7	37.2	68.5	75.5	362.8
SGR [15]	1.2	6.8	13.8	0.9	7.0	13.2	42.9	0.8	3.1	6.9	1.1	3.5	6.5	21.9	1.4	9.9	14.1	2.3	8.5	16.3	52.5
NCR-SGR [23]	13.5	34.0	49.6	10.9	36.6	52.5	197.1	8.2	20.3	43.2	8.4	21.7	40.1	141.9	32.4	63.4	78.9	26.2	62.3	79.4	342.6
SAF [15]	0.7	4.3	12.2	2.1	5.5	11.7	36.5	1.1	6.2	10.0	1.4	4.0	10.3	33.0	2.8	8.5	18.3	2.1	8.8	17.2	57.7
RCL-SAF [21]	16.3	43.3	64.5	12.5	46.4	66.2	249.2	14.6	39.2	54.6	15.5	42.8	59.8	226.5	35.2	67.6	80.3	32.4	68.7	87.0	365.6
MV-VSE [30]	9.7	23.5	36.0	10.2	27.7	42.1	149.2	6.6	20.4	36.4	6.0	24.6	35.5	129.5	23.3	33.8	54.9	17.4	32.7	55.1	217.2
NAAF [48]	13.8	28.4	39.0	13.5	32.2	46.5	173.4	8.3	25.5	37.6	9.5	31.5	46.8	159.2	26.8	53.5	60.6	25.5	46.2	62.0	274.6
DIVE [27]	23.4	50.5	66.5	20.0	49.2	65.9	275.5	20.2	43.1	55.4	13.1	42.2	55.9	229.9	39.4	74.7	**90.1**	32.7	71.1	83.7	391.7
CHAN [35]	12.1	31.1	44.0	11.8	29.1	43.0	171.1	9.9	24.8	38.4	10.2	25.1	39.4	147.8	23.7	47.9	66.2	18.5	53.2	67.9	277.4
ESA [49]	28.4	55.9	67.9	25.3	55.3	67.7	300.5	21.2	42.3	58.5	20.2	47.5	58.8	248.5	44.7	73.1	84.2	36.2	72.7	84.8	395.7
HREM [19]	24.1	61.7	72.3	24.0	58.4	72.1	312.6	20.0	44.9	59.5	20.5	48.6	60.5	254.0	31.0	70.4	78.9	32.4	73.2	88.5	374.4
Ours	**36.4**	**70.5**	**80.3**	**27.1**	**67.9**	**80.7**	**362.9**	**24.1**	**49.0**	**62.0**	**21.2**	**54.7**	**65.6**	**276.6**	**51.5**	**78.3**	**90.1**	**42.3**	**78.4**	**90.5**	**431.1**

Table 3. Performance comparison on the three datasets in terms of R@1, R@5, R@10 and the sum of them. All the methods use BERT as the text backbone. The highest results are shown in **bold** and the second highest results are underlined.

Method	3D2T-SR							3D2T-NR							3D2T-QA						
	Point→Text			Text→Point				Point→Text			Text→Point				Point→Text			Text→Point			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
VSE	17.7	34.8	50.4	15.6	44.1	54.6	217.2	14.6	36.9	46.9	12.5	38.8	54.0	203.7	31.0	66.2	73.2	24.8	66.2	81.7	343.1
VSE++ [17]	18.4	41.4	51.8	15.6	42.4	59.9	229.5	20.0	34.6	42.3	14.2	36.9	49.5	197.5	35.2	66.2	77.5	31.0	65.1	80.3	355.3
VSE∞ [6]	26.2	58.9	73.0	21.7	58.8	67.8	306.4	22.3	45.4	58.5	15.8	46.6	62.2	250.8	45.1	71.1	81.7	44.5	69.4	82.7	394.5
SGR [15]	1.2	6.8	13.8	0.9	7.0	13.2	41.9	1.2	3.5	7.5	1.5	4.0	6.8	24.5	2.8	7.0	12.7	1.7	8.7	16.6	49.5
NCR-SGR [23]	17.0	44.0	61.0	18.2	49.6	70.5	260.3	20.8	42.3	57.7	16.9	46.6	66.6	250.9	35.2	62.0	84.5	26.5	69.6	86.3	364.1
SAF [15]	2.1	5.0	10.6	2.3	5.2	10.5	35.7	0.8	7.4	12.1	1.5	6.2	10.2	38.2	3.4	9.5	21.1	2.3	9.9	19.4	65.6
RCL-SAF [21]	17.7	53.9	69.5	20.4	53.9	71.3	286.7	20.0	43.1	56.9	16.0	44.9	62.8	243.7	33.8	71.8	87.3	32.1	72.4	88.2	385.6
MV-VSE [30]	13.5	43.3	56.7	16.6	41.8	59.0	230.9	13.5	32.3	47.7	12.3	36.3	52.8	194.9	38.0	68.7	75.6	32.4	65.1	77.7	357.5
NAAF [48]	12.6	36.2	46.8	10.5	36.0	52.1	194.2	10.0	27.7	39.2	13.2	34.2	49.8	174.1	28.6	63.4	75.7	26.6	69.3	83.7	347.3
DIVE [27]	26.0	55.3	70.9	24.0	56.5	69.1	301.8	19.2	40.0	56.9	15.7	44.9	62.8	239.5	46.3	74.2	83.7	39.1	72.9	83.2	399.4
CHAN [35]	16.2	39.0	55.3	15.9	38.9	54.3	219.6	18.5	36.9	55.4	12.5	36.3	53.2	212.8	25.4	57.7	78.9	20.3	56.6	77.9	316.8
ESA [49]	34.9	66.7	75.9	27.1	61.1	74.5	340.2	23.8	43.8	55.4	19.7	48.6	62.8	251.1	45.1	73.2	83.1	36.3	74.4	86.5	398.6
HREM [19]	36.5	65.5	76.7	27.5	63.7	77.6	347.5	26.9	51.5	**65.5**	22.0	54.3	66.6	286.8	52.9	69.0	81.7	37.2	79.2	87.9	407.9
Ours	**42.3**	**71.6**	**83.2**	**30.8**	**69.9**	**83.0**	**380.8**	**30.5**	**52.2**	63.3	**23.2**	**58.2**	**74.2**	**301.6**	**56.2**	**84.1**	**90.7**	**46.5**	**85.5**	**94.6**	**457.6**

out 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 each attention will decrease matching performance, demonstrating that each attention contributes to the comprehensive perception of features.

6.4. 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 visualize the token-level attention throughout the point clouds and texts, as shown in Figure 6. Due to space limitations, more PTM results are displayed in our Complementary Materials. Additionally, we present a performance comparison among our RoMa and the VSE∞ [6], ESA [49], and HREM [19] through-

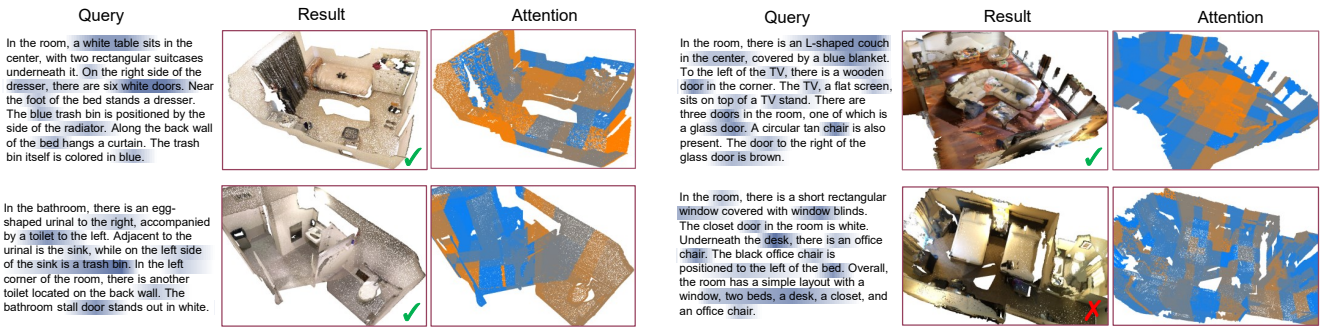
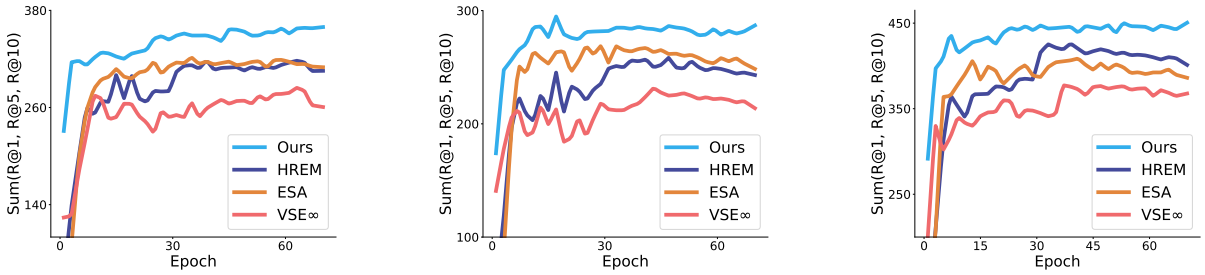


Figure 6. Some retrieved examples of PTM on 3D2T-QA. For each text query, the top-1 ranked point cloud. The correctly matched point clouds are marked with a green tick, otherwise the red cross. In addition, we visualize the text after applying attention, where **darker** colors signify increased attention weights, and we present a comparison between the original point clouds and the point cloud after applying attention. **Brighter** patches indicate higher attention weights.



(a) 3D2T-SR (b) 3D2T-NR (c) 3D2T-QA

Figure 7. The performance of VSE∞, ESA, HREM, and our RoMa on the three datasets.

Table 4. Ablation studies for our RoMa framework and DAP module adopted by our RoMa on the 3D2T-QA datasets. ✓ stands for use. $\bar{A}$ and $\hat{A}$ stand for using only one of them in DAP.

Feat. Extraction			Loss			Point→Text		Text→Point		
GPO	ESA	DAP	$\mathcal{L}_c$	$\mathcal{L}'$	$\mathcal{L}$	R@1	R@10	R@1	R@10	Sum
					✓	1.4	9.9	1.1	7.6	20.0
		✓				9.7	36.5	8.2	37.4	91.8
✓			✓			38.5	76.7	37.2	75.5	227.9
✓				✓		39.8	77.2	38.1	79.4	234.5
✓					✓	42.3	80.3	36.3	82.6	241.5
	✓		✓			42.7	84.2	37.2	84.8	246.9
	✓			✓		43.4	86.2	38.1	86.7	254.4
	✓				✓	45.2	88.7	37.8	87.1	258.8
		✓	✓			45.4	87.3	40.6	86.9	260.2
		✓		✓		49.2	87.3	41.2	88.5	266.2
		$\bar{A}$			✓	47.0	82.6	37.2	85.4	252.2
		$\hat{A}$			✓	50.8	89.5	37.2	90.3	267.8
		✓			✓	**51.5**	**90.1**	**42.3**	**90.5**	**274.4**

out the training process, as shown in Figure 7. 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 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 other baselines involve performance degradation in the later training stage, impacted by the noisy correspondence. In contrast, our RoMa mitigates the negative impact, achieving superior and robust performance.

7. 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 three benchmark datasets, namely 3D2T-SR, 3D2T-NR, and 3D2T-QA. 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 13 state-of-the-art methods on the three datasets, demonstrating the superiority of our RoMa in the challenging PTM task.

References

[1] Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8938–8947, 2019. 3

[2] 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, 3, 5

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 2, 3, 5

[4] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010. 3

[5] 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. 2, 3, 5

[6] 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. 2, 3, 7, 8

[7] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio 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*, pages 100–116. Springer, 2019. 3

[8] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023. 3

[9] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 3

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 7

[11] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 722–739, 2021. 1

[12] 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. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[14] 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. 7

[15] 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*, pages 1218–1226, 2021. 2, 3, 7, 8

[16] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9346–9355, 2019. 3

[17] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2, 3, 7, 8

[18] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3722–3731, 2021. 3

[19] 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, 2023. 2, 3, 7, 8

[20] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 3

[21] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. 6, 7, 8

[22] Rui Huang, Xuran Pan, Henry Zheng, Haojun Jiang, Zhifeng Xie, Cheng Wu, Shiji Song, and Gao Huang. Joint representation learning for text and 3d point cloud. *Pattern Recognition*, page 110086, 2023. 1, 3

[23] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems*, pages 29406–29419. Curran Associates, Inc., 2021. 2, 7, 8

[24] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining

for dense captioning in 3d scenes. In *European Conference on Computer Vision*, pages 528–545. Springer, 2022. 3

[25] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie 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*, pages 10984–10994, 2023. 3

[26] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021. 3

[27] 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. 7, 8

[28] 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. 3, 7

[29] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (8):3412–3432, 2020. 1

[30] Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Xijun Xue. Multi-view visual semantic embedding. In *IJCAI*, page 7, 2022. 7, 8

[31] Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11405–11415, 2021. 3

[32] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, pages 3–11, 2019. 2

[33] Weiping Liu, Jia Sun, Wanyi Li, Ting Hu, and Peng Wang. Deep learning on point clouds and its application: A survey. *Sensors*, 19(19):4188, 2019. 2

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[35] 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. 3, 7, 8

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu 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)*, pages 2556–2565, 2018. 3

[37] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 2

[38] Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi 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*, pages 6884–6893, 2023. 1, 3

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[40] Guangzhi Wang, Hehe Fan, and Mohan Kankanhalli. Text to point cloud localization with relation-enhanced transformer. *arXiv preprint arXiv:2301.05372*, 2023. 1

[41] 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. 7

[42] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016. 3

[43] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 3

[44] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R Qi, and Dragomir Anguelov. Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15446–15456, 2021. 1

[45] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022. 3

[46] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang 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*, pages 1791–1800, 2021. 3

[47] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022. 3

[48] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15661–15670, 2022. 2, 7, 8

[49] 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. 2, 7, 8