
Supplementary Material

Robust Cross-modal Alignment Learning for Cross-Scene Spatial Reasoning and Grounding

In this supplementary material, we provide additional information on the Cross-Scene Spatial Reasoning and Grounding (CSSRG) datasets, method, and experiments. More specifically, we first introduce existing 3D Visual Grounding datasets and our proposed CrossScene-RETR dataset, along with detailed statistics, analysis, and visual illustrations of CrossScene-RETR. We then provide a discussion and analysis on the feature aggregation and loss function of our RTSA module, and elaborate on the details and implementation process of our proposed CoRe. Additionally, we explain how baselines from related tasks are adapted for evaluation in the Cross-Scene Spatial Reasoning and Grounding (CSSRG) setting and supplement this with a parameter analysis of CoRe to further validate its effectiveness. Finally, we discuss the limitations and potential impact of our work.

A Supplementary Explanation of the Datasets

In this section, we first provide a supplementary introduction to existing 3D Visual Grounding (3DVG) datasets. Furthermore, to enhance understanding of our proposed CrossScene-RETR dataset, we add detailed descriptions of its construction process, along with more comprehensive statistical analyses and visualizations.

A.1 Introductions to the Adopted Datasets

We provide details about the point-cloud and text datasets used in dataset construction and experiments. For the point-cloud dataset, ScanNet [1] is an instance-level 3D dataset comprising thousands of 3D point-cloud scans and nearly 2.5 million views across over 1,500 indoor room scenes. Additionally, the text datasets used for descriptions are outlined as follows:

- *ScanRefer* [2]: It serves as a large-scale and highly discriminative dataset for 3D visual grounding and dense captioning, featuring 51,583 object descriptions from thousands of objects spanning nearly 800 ScanNet scenes. In our experiments, we use 562 scenes for training and 141 scenes for testing.
- *Nr3d/Sr3D* [3]: NR3D and SR3D are also built upon ScanNet, with SR3D consisting of 83,572 straightforward machine-generated descriptions and NR3D featuring 41,503 descriptions, closely resembling the human annotations found in ScanRefer. In the experiments, we use 511 scenes for training and 130 scenes for testing of Nr3D, 1,018 scenes for training and 255 scenes for testing of Sr3D.
- *CrossScene-RETR*: This is the dataset we presented for CSSRG, as detailed in *Section 4*. In our experiments, we use 571 scenes for training and 125 scenes for testing.

A.2 Details of our Dataset Construction

In Section 4.1, we have provided an overview of the generation process descriptions of our CrossScene-RETR. In this supplementary material, we further supplement the prompt details of four kinds of description instance generation.

We employ the API of GPT-4o, creating a new session for each data sample. Initially, we input a prompt template that encompasses a description of the task background, and specific requirements for cross-scene object reasoning. To generate more diverse and practical descriptions, we introduce

switch varied linguistic style requirements to produce four different stylistic descriptions, shown as follows:

- *Characteristic-focused descriptions* primarily describe the salient characteristics of objects such as color, shape, material, and the relationship with prominent objects, which could be used to train/evaluate the ability to capture discriminative information in CSSRG.
- *Spatial-information-focused descriptions* mainly detail the position of objects and inter-object relationships, which could be used to train/evaluate the ability to grasp spatial information.
- *Comprehensive descriptions* consist of exhaustive and in-depth object descriptions, which could be used to train/evaluate the overall cross-scene grounding capability.
- *Fuzzy descriptions* use vague descriptions to simulate the real-world scenario where the users have unclear memory. The descriptions could be adopted to train/evaluate the practical CSSRG ability with limited information.

Their corresponding generation prompts are as follows:

*Note: Please generate four detailed paragraphs in English, including as much information as possible, avoiding unnecessary sentences or line breaks. The content should be based on the given description of an object (**ObjectName**). The information must be factual and not fabricated. The purpose of each generated paragraph is that "I" hope others can help me find the (**ObjectName**). All start with: I want to find a (**ObjectName**). Requirements for each paragraph: (**Switch1**) First paragraph: Focus as much as possible on the main characteristics of the object, such as color, shape, material, or any related features, while also describing the spatial relationship of the object's position in relation to its surroundings. (**Switch2**) Second paragraph: Focus primarily on the spatial relationship and arrangement of the object within the space. (**Switch3**) Third paragraph: The description should be as comprehensive and detailed as possible. (**Switch4**) Fourth paragraph: Pretend as if my memory about the placement is somewhat vague, but the information is still accurate. Keep the description brief. Each paragraph should utilize as much of the given information as possible, and prioritize the most distinguishing details. Avoid any fabrication or altering the meaning of the information. Given refer corpus: (**Corpus**).*

Where **ObjectName** is the actual category name of the object being described, **Switch** represents the switchable requirements for generating descriptions in different language styles. **Corpus** refers to all the corpus in the existing datasets (*i.e.*, ScanRefer [2], Nr3D [3], Sr3D [3], and ScanQA [4]) corresponding to the object. Afterward, we based the descriptions generated by GPT-4o and divided them into four corresponding style subsets, completing the *Text Generation Phase*.

A.3 Supplementary Statistics and Analysis of CrossScene-RETR

First, we present word clouds of the descriptions in CrossScene-RETR subsets with different linguistic styles, as shown in Figure 1. Additionally, we provide statistical information for the four subsets with varying linguistic styles Figure 2 and Table 1, and visualize the sample proportions along with the top 10 category distributions across the three subsets Figure 3, which are defined by the challenge levels outlined in *Section 4.1 Scene Analysis Phase*. Together with *Table 1* in the main body, we could draw the following observations:

- Our descriptions are richer and more discriminative, making them suitable for supporting CSSRG. Specifically, although the corpus is partially drawn from the existing datasets, the richness of description and vocabulary far exceeds theirs. 81.2% of CrossScene-RETR descriptions include color, 45.0% contain shape, 38.5% mention material, and 97.5% provide spatial information. On average, each description covers 11.4 objects, 5.9 object characteristics, and 6.3 spatial relationships. Moreover, the descriptions we generate contain significantly more information points than existing texts. This ensures that the descriptions retain discrimination across all scenes.
- The descriptions of subsets with different linguistic styles emphasize distinct aspects. Specifically, for the *Characteristic-focused* subset, over 50% of the descriptive terms are related to characteristics such as color, shape, and material. In the *Spatial-information-focused* subset, 68% of the terms are related to spatial information, exceeding other subsets by more than 10%. Additionally, the average number of spatial description terms per sentence is 9.5,

with over 97.5% of the samples containing such terms, highlighting the strong emphasis on spatial understanding. Furthermore, the *Comprehensive* subset contains an average of 133.5 words, 18.7 objects, 10.0 attributive terms (e.g., color, shape), and 9.5 spatial terms, far surpassing other subsets. The *Fuzzy* subset contains the fewest information points but introduces many uncertain descriptions to enhance real-world applicability.

- The category distribution in the *regular* labeled subset closely aligns with the evaluation set, facilitating a comprehensive method evaluation. The *conspicuous* subset contains distinctive objects, like sinks and toilets, which are easier to identify. The *confusing* subset includes common objects, such as chairs, doors, and tables, prevalent in most scenes, requiring more in-depth analysis.
- The object descriptions we have constructed are not only detailed but also feature a well-structured and logical division into two kinds of subsets (i.e., different challenge levels of grounding and varied linguistic styles). This framework holds great potential for future applications in 3D Visual Grounding, 3D Dense Captioning, and Pointcloud-Text Matching tasks, offering new avenues for more precise and context-aware interactions with 3D data.

To provide a more intuitive understanding of the descriptions and to explore the diversity within our proposed dataset, we present illustrative samples of point-cloud scene objects along with four distinct styles of descriptions at the end of this appendix, as shown in Figures 7 to 9.

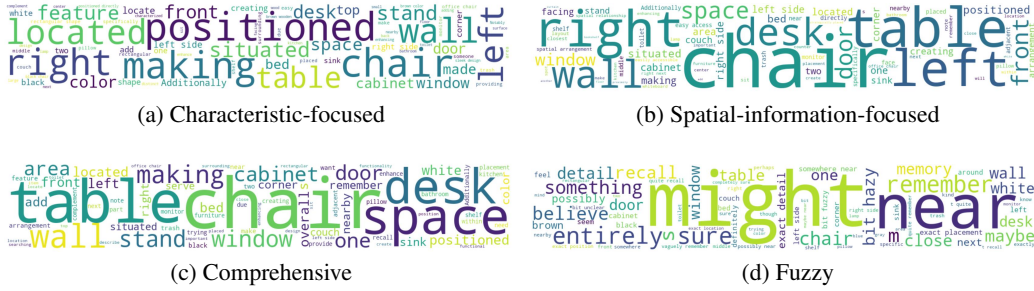


Figure 1: Word clouds of the four subsets with different linguistic styles (i.e., *Characteristic-focused*, *Spatial-information-focused*, *Comprehensive*, *Fuzzy* description subsets) in our proposed CrossScene-RETR dataset.

B Supplementary Explanation of the Method

In this section, we first include computation details of the feature aggregation adopted in our RTSA, and add a theoretical analysis and a proof showing that the loss \mathcal{L}_m in RTSA is robust to false-negative pairs. Additionally, to demonstrate reproducibility and enhance understanding of our CoRe, we provide a detailed explanation of the implementation details and the algorithmic process.

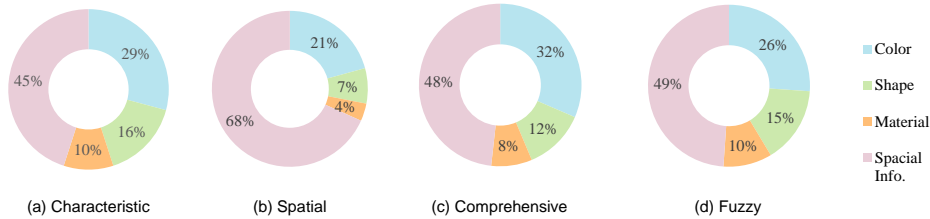


Figure 2: The distribution of descriptive terms such as color, shape, material, and spatial information across subsets with different linguistic styles (i.e., *Characteristic-focused*, *Spatial-information-focused*, *Comprehensive*, *Fuzzy* description subsets) in our proposed CrossScene-RETR dataset.

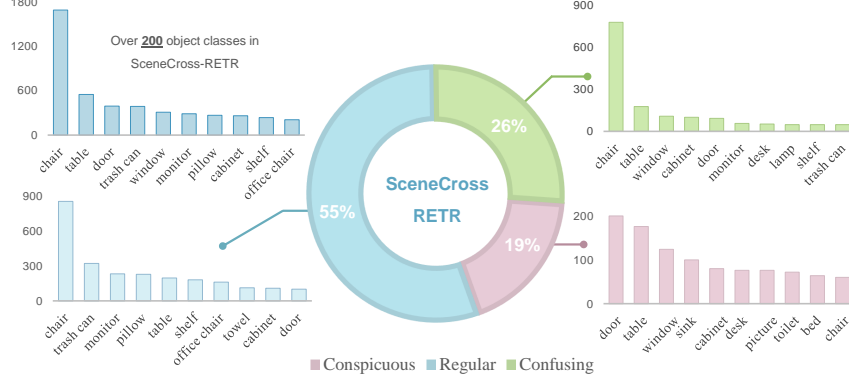


Figure 3: Top 10 category distributions of the CrossScene-RETR test set and its corresponding three subsets with different challenge levels. **Blue** represents the category distribution across the entire dataset, **red** represents the *Conspicuous* labeled subset, **light blue** represents the *Regular* labeled subset, and **green** represents the *Confusing* labeled subset.

Table 1: Statistics comparison between *Characteristic-focused* (Characteristic), *Spatial-information-focused* (Spatial), *Comprehensive*, *Fuzzy* description subsets in the proposed CrossScene-RETR dataset.

	Characteristic	Spatial	Comprehensive	Fuzzy
Average length	74.3	63.3	133.5	55.5
Number of samples	9,879	9,879	9,888	9,880
Number of objects per text	9.5	8.9	18.7	8.4
Number of attributes per text	6.8	2.6	10.0	4.4
Number of spatial info. per text	5.7	5.6	9.5	4.5
Text with spatial info. (%)	97.0	97.5	98.2	97.3
Text with color description (%)	95.3	62.2	97.5	69.7
Text with shape description (%)	61.2	17.4	65.5	35.9
Text with material description (%)	47.5	15.9	58.3	32.5
Number of info. points per text	20.7	17.1	38.12	17.2

B.1 Details of Feature Aggregation

To facilitate text-scene Cross-Modal Matching (CMM), we first aggregate the fine-grained object and word features into the scene and text representations. Its implementation details are shown in Figure 4. Specifically, inspired by [5], we utilize a learnable affinity matrix to capture the interrelationships among feature dimensions, modeling their relative contribution. Based on this, we could obtain interrelation-focused weights for each dimension. In addition, we integrate the interrelationships to derive the feature-focused weight, which is used to assess the overall importance of features. Taking the text modality as an example, the two weights are calculated as follows:

$$U_i^t = W_u^t Z_i^w, \quad V_i^t = f_v(U_i^t; \theta_v), \quad (1)$$

where Z_i^w is the fine-grained features of text with word tokens X_i^t , $U_i^t/V_i^t \in \mathbb{R}^{M_i \times d}$ is the interrelation-focused/feature-focused weights of text features, W_u^t is the learnable affinity matrix, f_v is the projection function to obtain feature-focused weight, and θ_v represent the parameters of f_v .

Subsequently, the two weights are applied to the fine-grained features, aggregating them into common representations. To leverage their complementary focus, we adaptively combine them via a discriminator, formulated as:

$$z_i^t = \theta_1^t (Z_i^w \odot U_i^t) + \theta_2^t (Z_i^w \odot V_i^t), \quad (2)$$

where z_i^t is the final common representation, $[\theta_1^t; \theta_2^t] = f_d([Z_i^w \odot U_i^t; Z_i^w \odot V_i^t]; \theta_d)$ is two adaptive weights for aggregation, $f_d(\cdot, \theta_d) : \mathbb{R}^{2(M_i \times d)} \rightarrow \mathbb{R}^2$ is the discriminator, and \odot means hadamard product. Similarly, we can obtain the scene representation z_j^s .

B.2 Robustness analysis of \mathcal{L}_m

Analysis: As q approaches 1 from 0, \mathcal{L}_m gradually converges to its lower bound, *i.e.*, MAE loss, which could prevent overfitting to noisy pairs [6]. We can define the matching risk $R(f)$ under ideal conditions:

$$R(f) = \mathbb{E}_{\mathcal{D}} (\mathcal{L}_m(f(X_i^t, X_j^s), \bar{y}_{ij}) + \mathcal{L}_m(f(X_j^s, X_i^t), \bar{y}_{ij})), \quad (3)$$

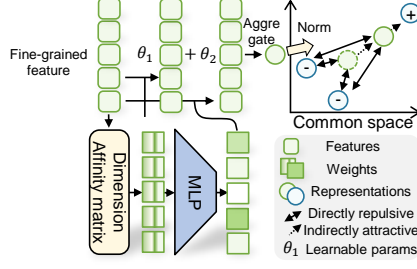


Figure 4: Details of the feature aggregation adopted in our RTSA.

where \mathbb{E} is the expectation operator, \bar{y}_{ij} is the ideal, correct correspondence label, and f is the matching function. In the case of asymmetric correspondence, where false-negative pairs occur with probability η , we can similarly define risk $R^\eta(f)$. After CMM, we could obtain global minimizer f^* and f_η^* under the two aforementioned scenarios.

When $q = 1$ and $\eta \leq (M - 1)/M$, we can prove that f^* is also a global minimizer of $R_{\mathcal{L}_m}(f)$. In other words, it is robust to false-negative pairs with a probability of less than $(M - 1)/M$.

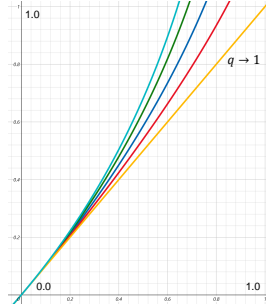


Figure 5: The change in the \mathcal{L}_m as q increases from 0 to 1.

In the limit as q approaches 1 from 0, the first partial derivative of \mathcal{L}_m is $\frac{\partial \mathcal{L}_m}{\partial s} > 0$, and the second partial derivative of $\frac{\partial^2 \mathcal{L}_m}{\partial s^2}$ is negatively correlated with q . In other words, as shown in Figure 5, the lower bound of \mathcal{L}_m occurs when $q = 1$, as shown below:

$$\mathcal{L}_m = \sum_{i,j}^K (1 - y_{ij})(\vec{s}_{ij} + \overleftarrow{s}_{ij}), \quad (4)$$

where

$$\vec{s}_{ij} = \frac{\exp(\mathbf{z}_i^t \top \mathbf{z}_j^s / \tau)}{\sum_k^K \exp(\mathbf{z}_i^t \top \mathbf{z}_k^s / \tau)}, \quad \overleftarrow{s}_{ij} = \frac{\exp(\mathbf{z}_i^s \top \mathbf{z}_j^t / \tau)}{\sum_k^K \exp(\mathbf{z}_i^s \top \mathbf{z}_k^t / \tau)}, \quad (5)$$

$\vec{s}_{ij} / \overleftarrow{s}_{ij}$ is the similarity between the i -th scene/text feature and the j -th text/scene feature, $\tau \in (0, 1]$ is the temperature parameter. Clearly, the lower bound of \mathcal{L}_m aligns with the form of the MAE loss. According to prior research on noisy label/correspondence learning [6, 7, 8], the MAE-based loss function exhibits uniform gradients across data of varying challenge levels, which facilitates its ability to handle data equally, even in the presence of noisy labels or correspondence issues.

To further illustrate the robustness of this loss, we try to prove it theoretically: When $q = 1$ and $p_\eta \leq (\eta - 1)/\eta$, f^* is also a global minimizer of $R_{\mathcal{L}_m}(f)$.

Proof. Recall that for any f ,

$$R(f) = \mathbb{E}_{\mathcal{D}} (\mathcal{L}_m(f(X_i^t, X_j^s), \bar{y}_{ij}) + \mathcal{L}_m(f(X_j^s, X_i^t), \bar{y}_{ij})),$$

where X_i^t and X_j^s are actually matched text and scene, i.e., $\bar{y}_{ij} = 1$.

Due to the wide variety of specific scenarios in which false-negative noise occurs, we follow the general theoretical derivation and prove under the condition of symmetric noise. For any f , $R^\eta(f)$ is written as:

$$\begin{aligned} R^\eta(f) &= \mathbb{E}_{\mathcal{D}_\eta} (\mathcal{L}_m(f(X_i^t, X_j^s), y_{ij}) + \mathcal{L}_m(f(X_j^s, X_i^t), y_{ij})) \\ &= \mathbb{E}_{\mathcal{D}} ((1 - \eta) (\mathcal{L}_m(f(X_i^t, X_j^s), \bar{y}_{ij}) + \mathcal{L}_m(f(X_j^s, X_i^t), \bar{y}_{ij})) + \\ &\quad \frac{\eta}{M-1} \sum_{k \neq j}^M (\mathcal{L}_m(f(X_i^t, X_k^s), \bar{y}_{ik}) + \mathcal{L}_m(f(X_k^s, X_i^t), \bar{y}_{ik}))) \end{aligned} \quad (6)$$

Note that the sum of matching predict probabilities is 1, so when $q = 1$, $\sum_k^M \mathcal{L}_m(f(X_i^t, X_k^s), \bar{y}_{ik}) = 1$. In this way, we can derive the above formula as follows:

$$\begin{aligned} R^\eta(f) &= \mathbb{E}_{\mathcal{D}} ((1 - \eta) (\mathcal{L}_m(f(X_i^t, X_j^s), \bar{y}_{ij}) + \mathcal{L}_m(f(X_j^s, X_i^t), \bar{y}_{ij})) \\ &\quad + 1/(M-1)((M-1) - (\mathcal{L}_m(f(X_i^t, X_j^s), \bar{y}_{ij}) + \mathcal{L}_m(f(X_j^s, X_i^t), \bar{y}_{ij}))) \\ &= CR(f) + C', \end{aligned} \quad (7)$$

where $C = \frac{M-1}{M}$ and C' are both computable constants. Now, for any f , $R^\eta(f^*) - R^\eta(f) = C(R_{\mathcal{L}_r}(f^*) - R_{\mathcal{L}_r}(f)) \leq 0$, where $\eta \leq \frac{N-1}{N}$ and f^* is a global minimizer of $R(f)$. This proves f^* is also the global minimizer of $R^\eta(f)$. \square

B.3 Details of Method Implementation

The method is trained following the 3DVG pipeline, with many implementation details following the 3DVG implementation in VisTA [9]. All the methods are carried out on GeForce RTX 3090 GPUs. The algorithm flow of our CoRe is shown in Algorithm 1. More specifically, training adopts 3D object ground truths, and testing uses the 3D object proposal, which is the same as the VisTA. For the optimization of our CoRe, we use the AdamW optimizer with a learning rate set to $lr = 1e^{-4}$. For the multimodal feature extractors, we use the pre-trained BERT for the text modality, PointNet++, and Spatial Transformer for the point-cloud modality. The learning rates for these models are set to $0.1 \times lr$. Additionally, the dimensions for both features and representations are set to 768. Our hyper-parameter settings are shown as follows: $\lambda_m = 1$, $\lambda_g = 1$, $\tau = 0.05$, $L = 6$. For detailed analysis, please see Appendix C.2.

C Supplementary Explanation of the Experiments

In this section, we provide additional implementation details for the different baselines used in the experiments.

C.1 Transfer details of the Existing Methods

1) Details of CMM Method Transfer: The image region features originally used in the methods are replaced with point-cloud objects, consistent with the CoRe approach. The text encoder continues to use the pre-trained BERT [10]. For the coarse-grained CMM method, we cannot solve its inference regarding the target object. However, for the fine-grained method, the top-1 matching object in the scene, as identified by the Cross-Attention mechanism, corresponds to the target object described by the query.

2) Details of 3DVG Method Transfer: We perform grounding for each scene individually, ranking the prediction scores of all objects. The object with the highest score is the one most relevant to the description across all scenes according to the 3DVG model.

3) Details of CMM+3DVG Method Implementation: Taking HREM+VisTA as an example, we train the two models using the original training methods for their original pipelines, respectively. During the inference stage, we employ a two-stage inference pipeline. Specifically, we first use HREM [11] for scene inference to obtain the scene corresponding to each query text. Then, we associate the correctly matched scene with the description text, while discarding the incorrect ones. These are then input into VisTA [9] for grounding. Finally, the grounding accuracy and the top-1 matching accuracy are jointly calculated to obtain the final CSSRG accuracy.

Algorithm 1 Main optimization process of our CoRe

Input: The training pointcloud-text data $\mathcal{D} = \{\mathcal{T}, \mathcal{S}\} = \{(X_i^t, X_j^s), y_{ij}\}_{i,j}^{N,M}$, maximal epoch number N_e and learning rate lr .

- 1: Load the pre-trained BERT encoder $f_b(\cdot, \theta_b)$, PointNet++ $f_p(\cdot, \theta_p)$, and Spatial Transformer $f_s(\cdot, \theta_s)$. Initialize the classification layer weights W_c for object discrimination maintain, affinity matrix W_u , projection layer $f_v(\cdot, \theta_v)$, discriminator $f_d(\cdot, \theta_d)$ for scene-text aligning; attention projection matrices W_q, W_k, W_v , Transformer fusion layers $f_u(\cdot, \theta_u)$, Screening Attention fusion coefficients $\{\mu_i\}_{i=1}^L$ for word-object associating.
- 2: **for** $i = 1, 2, \dots, N_e$ **do**
- 3: Obtain the fine-grained features Z_i^w and Z_j^o through the modality-specific encoders. Compute class predictions of the object features using W_c , and then obtain the object semantic aligning loss \mathcal{L}_c based on Equation (2).
- 4: Aggregate fine-grained features into text/scene global representations based on Equation (3) and Equation (4). Calculate aligning loss \mathcal{L}_m in Equation (5) with adaptive q for robust scene matching.
- 5: After obtaining top 1 retrieved scenes, facilitating word-object association based on Equation (8) and Equation (9). Then, calculate grounding loss \mathcal{L}_m based on Equation (10).
- 6: Obtaining the overall loss \mathcal{L} based on Equation (1).
- 7: Update the learnable parameters and weights $\theta_b, \theta_p, \theta_s, \theta_v, \theta_d, \theta_u, W_c, W_u, W_q, W_k, W_v$, and $\{\mu_i\}_{i=1}^L$ by minimizing the loss \mathcal{L} with descending their stochastic gradient:
$$\theta_b = \theta_b - 0.1 \cdot lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \theta_b}\right), \theta_p = \theta_p - 0.1 \cdot lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \theta_p}\right), \theta_s = \theta_s - 0.1 \cdot lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \theta_s}\right)$$
$$\theta_v = \theta_v - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \theta_v}\right), \theta_d = \theta_d - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \theta_d}\right), \theta_u = \theta_u - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \theta_u}\right)$$
$$W_c = W_c - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial W_c}\right), W_u = W_u - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial W_u}\right), W_q = W_q - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial W_q}\right)$$
$$W_k = W_k - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial W_k}\right), W_v = W_v - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial W_v}\right), \{\mu_i\}_{i=1}^L = \mu_i - lr \cdot \left(\frac{\partial \mathcal{L}}{\partial \mu_i}\right)$$
- 8: **end for**

Output: Optimized learnable parameters and weights for two modalities.

C.2 Parameter Analysis

Table 2: CSSRG performance on CrossScene-RETR obtained for different values of λ_m and λ_g , in terms of Acc@0.25 and Acc@0.5. The **Bold** represents the optimal performance.

λ_m	1	5	10	20	1	1	1
λ_g	1	1	1	1	5	10	20
Acc@0.25	22.99	21.74	21.34	19.56	22.65	21.20	20.52
Acc@0.5	21.26	20.14	19.84	18.09	21.13	19.75	19.34

To investigate the sensitivity of our CoRe to the adopted hyper-parameters, we plot the CSSRG performance in terms of Acc@0.25 and Acc@0.5 against varied hyper-parameters (*i.e.*, $\lambda_m, \lambda_g, \tau, L$) on the CrossScene-RETR dataset, as shown in Figure 6 and Table 2. The experimental results demonstrate the insensitivity of the hyper-parameters we set, with optimal settings being effective within a large and reasonable range. Notably, for the trade-off parameter lambda, larger values of λ_g are more likely to achieve superior performance. On the other hand, when λ_m is set too large, the final CSSRG performance deteriorates. We analyze that this is due to fine-grained feature interactions being more challenging to fit compared to coarse-grained representation alignment. Excessive focus on the latter may lead to underfitting of the former, causing an imbalance. Moreover, when $\tau = 0.05$, our RTSA achieves the optimal text-scene alignment, thereby enhancing scene matching and improving object reasoning performance. When $L = 6$, TWA constructs the most suitable dynamic sparse attention, leading to comprehensive filtering of redundant information.

D Limitations and Potential Impact Statement

Although our work has taken the initial step forward in CSSRG, there are some limitations that should be acknowledged. The performance of the methods on CSSRG task is relatively low. Since the 3D

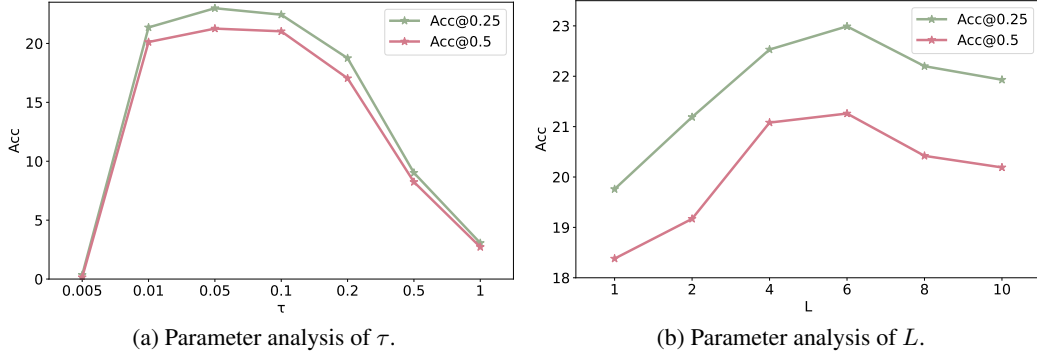


Figure 6: CSSRG Performance of CoRe in terms of the Acc@0.25 and Acc@0.5 versus different values of τ and L on CrossScene-RETR.

vision develops rapidly, it remains uncertain whether our CoRe can maintain robustness in future unknown point-cloud data. We will undertake further in-depth research.

Our work does not pose any potential impact concerns, as it utilizes the publicly available and widely adopted 3D scene dataset ScanNet [1]. Additionally, all textual data used in our research are collected and generated in accordance with established data collection and content creation standards.

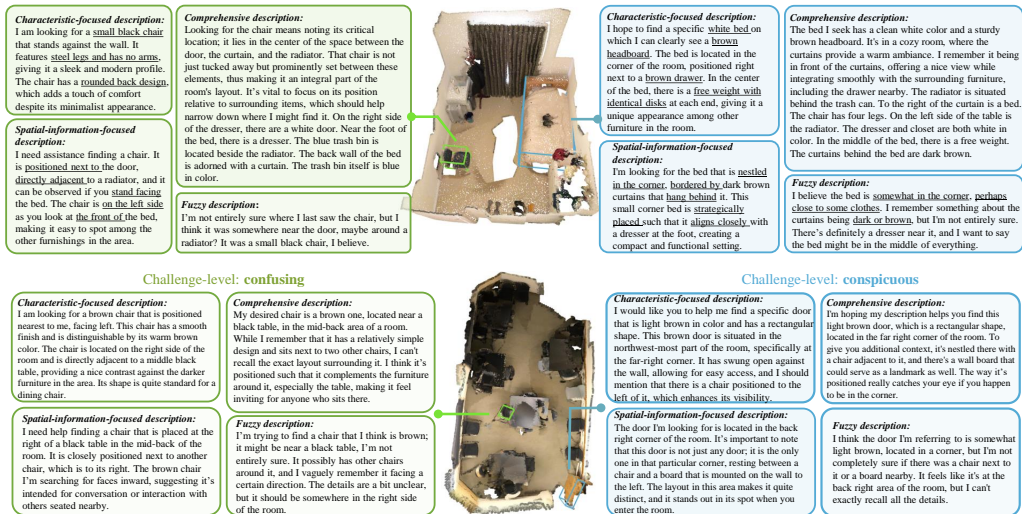


Figure 7: Some samples in our CrossScene-RETR dataset. Each sample displays a point cloud scan with two corresponding target objects, which is associated with four different linguistic style descriptions.

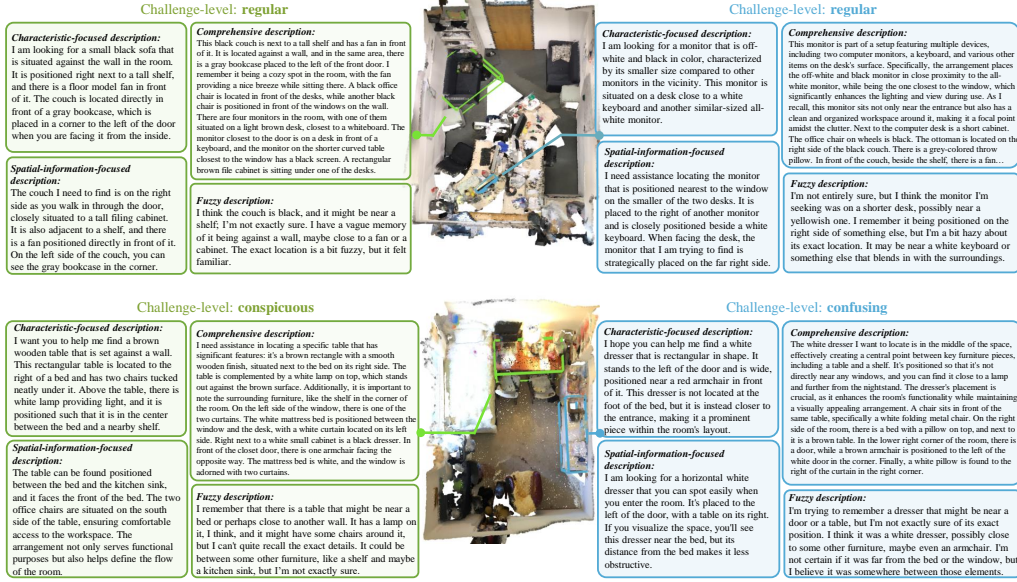


Figure 8: Other samples in our CrossScene-RETR dataset. Each sample displays a point cloud scan with two corresponding target objects, which is associated with four different linguistic style descriptions.

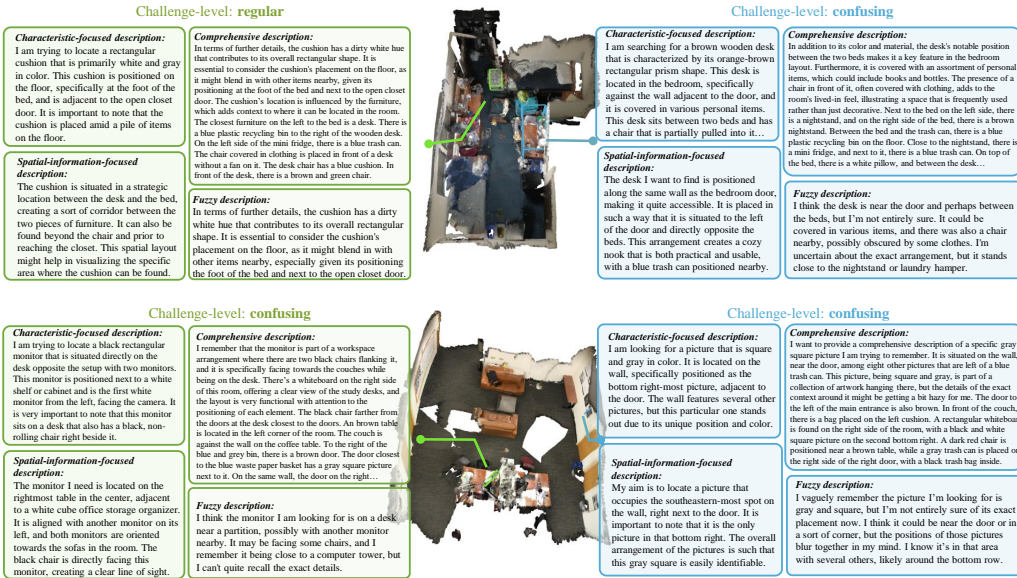


Figure 9: Other instances in our CrossScene-RETR dataset. Each sample displays a point cloud scan with two corresponding target objects, which is associated with four different linguistic style descriptions.

References

- [1] 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.
- [2] 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.

- [3] 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.
- [4] 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.
- [5] Kun Zhang, Bo Hu, Huatian Zhang, Zhe Li, and Zhendong Mao. Enhanced semantic similarity learning framework for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [6] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.
- [7] 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.
- [8] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.
- [9] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.
- [10] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [11] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15159–15168, June 2023.