G^3-LQ: Marrying Hyperbolic Alignment with Explicit Semantic-Geometric Modeling for 3D Visual Grounding

Supplementary Material

1. Overview

In this supplementary material, we supplement significant ingredients for our proposed G^3-LQ (explicitly models **G**eometric-aware visual representations and **G**enerates fine-**G**rained **L**anguage-guided object **Q**ueries) framework and provide more quantitative results and visualizations to further verify the effectiveness of the G^3-LQ approach in 3D visual grounding tasks. We introduce more implementation details (*i.e.*, training details, framework details and loss details) of our method in Section 2. In Section 3, we present additional experiments, *e.g.*, hyper-parameters selection and ablation studies on the widely-used benchmarks *ScanRefer* and *Sr3D/Nr3D* to better demonstrate the superiority of the proposed G^3-LQ method. In Section 4, we provide more visualization results in regard to the One-Stage paradigm, Sr3D/Nr3D and failure cases of our method. Finally, we describes the limitation analysis and future improvement of our G^3-LQ framework in Section 5.

2. Implementation Details

2.1. Datasets

ScanRefer. The ScanRefer benchmark [3], stands as a pivotal milestone in the realm of 3D visual grounding, which is composed of 51,583 human-written utterances of 11,046 3D objects annotated across 800 ScanNet [5] indoor scenes. In each scene, an average of 13.81 objects along with 64.48 descriptions are provided. Depending on whether the target object is a unique object class in the given scene, the dataset can be categorically divided into the "Unique" and the "Multiple" subset. Following prior works, we evaluate the experimental results with the Acc@mIoU metric (Acc) with the threshold $m \in \{0.25, 0.5\}$, which quantifies the proportion of the predicted boxes whose Intersection over Union (IoU) with the ground-truth (GT) boxes exceeds m.

Nr3D and Sr3D. The Nr3D and Sr3D [2] sub-datasets within ReferIt3D contribute to the comprehensive understanding of 3D visual grounding. Both sub-datasets are built upon ScanNet scenes, enriching the dataset with diverse and representative real-world environments. Nr3D (Natural Reference in 3D) provides 41,503 texts collected by ReferItGame. SR3D (Natural Reference in 3D) includes 83,572 synthetic utterances based on a target-spatial relationship-anchor template and utilizes spatial relation to localize a referred object. Nr3D and Sr3D have different test subsets: "Easy" and "Hard" splits align with the "Unique" and the "Multiple", while the "View-dependent" (VD) and "View-independent" (VID) splits are discerned whether the referring expression hinges upon the speaker's viewpoint. The evaluation metric of both datasets is the accuracy, i.e., whether the model correctly identifies the target object.

2.2. Additional Details

Training Details. The code is implemented on the PyTorch [14] platform equipped with 4 NVIDIA 12GB TITAN Xp GPUs. We freeze the text backbone RoBERTa, while keeping the rest of the network trainable. For the ScanRefer dataset, we train our G^3-LQ model with the AdamW optimizer, which configs an initial learning rate of 2×10^{-3} for the pre-trained PointNet++ [15] tokenizer, 2×10^{-4} for other layers, and the weight decay as 5×10^{-4}. Our model is optimized for 80 epochs and we utilize a learning rate decay strategy at epoch $\{50, 75\}$ with a rate of 0.1. For the Sr3D dataset, we train the network for around 60 epochs and the initial learning rate is set to 1×10^{-3} and 1×10^{-4}, which is reduced at epoch $\{30, 40\}$ with a rate of 0.1. For the Nr3D dataset, which comprises complex free-form descriptions, the proposed network is trained for about 180 epochs and the initial learning rate is set to 1×10^{-3} and 1×10^{-4}, declining at epoch 150 with a rate of 0.1.

Framework Details. For all the datasets, the xyz coordinates and RGB values are the input into the overarching network. We utilize PointNet++ [15] as point cloud tokenizer to perform set-abstraction and subsample points to $n = 1024$. Similar to EDA [20], the maximum length of text tokens $\mathcal{T}$ is set to 256, and the absence bit of the position label is padded with 0. In the two-stage paradigm of our proposed G^3-LQ framework, we utilize the information obtained from the location and category of the detected boxes, which are embedded separately and concatenated as the box token $\mathcal{B} \in \mathbb{R}^{b \times d}$. The number of channel dimensions d, and candidate object embeddings, box tokens are empirically set to 288, 256 and 132, respectively. The Geometric and Language Encoder with 3 layers, facilitate cross-modal feature extraction and interaction, while the Geometric Decoder equipped with 6 layers generates candidate object features. Please refer to BUTD-DETR [10] for more details.

Loss Details. In addition to the Poincaré Semantic Alignment loss $\mathcal{L}_{\text{psa}}$ introduced in the manuscript, we also adopt the commonly-used object location loss $\mathcal{L}_{\text{loc}}$, the KPS point sampling loss $\mathcal{L}_{\text{pts}}$ [11] and the position-aligned loss $\mathcal{L}_{\text{pa}}$ proposed in EDA [20]. Specifically, $\mathcal{L}_{\text{loc}}$ is composed of $\mathcal{L}_{\text{box}}$ and $\mathcal{L}_{\text{iou}}$. The former corresponds the L1 loss to regress the position and size of the target object. The latter rep-

resents the 3D IoU loss between the predicted and ground truth loss. $\mathcal{L}_{\text{pa}}$ aims to align the language-modulated visual features consistent with the text descriptions. For more details of the parameters searching about $\mathcal{L}_{\text{pa}}$, please refer to EDA [20]. Finally, we optimize the proposed G^3-LQ model with the following total loss:

$$\mathcal{L} = \left(\alpha\left(\mathcal{L}_{\text{pa}} + \mathcal{L}_{\text{psa}}\right) + \mathcal{L}_{\text{loc}}\right) / \left(N_D + 1\right) + 8\mathcal{L}_{\text{pts}} \quad (1)$$

where N_D is the layer number of the multi-modal decoder. As described in EDA, α exhibits distinct values depending on the significance in bounding box detection. Specifically, within the Sr3D/Nr3D dataset, α assumes a value of 1, while in the case of the ScanRefer dataset, it is set to 0.5.

3. Additional Experiments

Impact of Hyper-parameters. In this section, we undertake a meticulous exploration of the 3D visual grounding performance across a battery of hyper-parameter configurations on ScanRefer and Nr3D/Sr3D datasets. Specifically, we examine the influence of various hyper-parameters, including the number of neighboring points denoted as k within the PAGE module, the curvature parameter c, and the temperature coefficient τ within the PSA loss function:

1) During the geometric feature extraction, the number of neighbor points k is of vital importance that determines the local structure and embedding scope of 3D point cloud [18, 19]. As shown in Table 1, too many neighbors lead to elevated computation burden and over-smoothing of geometric features, thus declining the grounding performance. Moreover, surplus neighboring nodes can introduce redundant information, making it more challenging for our G^3-LQ model to discern critical geometric features, ultimately diminishing the accuracy of 3D visual grounding. On the other hand, too few neighbors cannot guarantee the representative ability of the PAGE module to comprehend *the scene-level 3D semantics* and *object-level spatial relationships*. A shortage of neighboring nodes can likewise result in an oversimplified representation of the point proxy, impeding its capacity to encapsulate the intricate structures within the 3D scene and consequently limiting the performance of the visual grounding task. When $k = 5$, our G^3-LQ model achieves the best performance.

2) Table 2 shows the model performance depending on the curvature c in the PSA loss. The curvature of the manifold is a pivotal factor that exposes and provides insight into the radius of the Poincaré ball [6, 13]. Intuitively, the smaller c value making the Poincaré ball degrade to the Euclidean space, aiding in understanding semantic relationships in language descriptions. The bigger c value makes the Poincaré space curved, shedding light on the geometric structure and aiding in accurately capture the point cloud shapes and spatial positions. Moreover, bigger c value exhibits heightened sensitivity to geometric features, which encourages our G^3-LQ model to adept at capturing subtle geometric variations, such as minor object displacements, discrepancy in angles, or intricate shape details. To strike a balance, we opt for a curvature value of $c = 0.1$, yielding the optimal performance.

Neighbor k	ScanRefer 0.25	ScanRefer 0.5	Sr3D(@0.25)	Nr3D(@0.25)
3	56.48	44.39	72.59	56.37
5	**56.90**	**45.58**	**73.10**	**56.97**
8	56.15	44.23	72.08	55.76
10	55.87	44.16	71.30	55.29

Table 1. Experiments on the ScanRefer and Sr3D/Nr3D benchmarks for different neighboring points number k of the point proxy construction in PAGE module.

Curvatures c	ScanRefer 0.25	ScanRefer 0.5	Sr3D(@0.25)	Nr3D(@0.25)
0.05	55.94	44.08	70.38	55.22
0.10	**56.90**	**45.58**	**73.10**	**56.97**
0.15	56.24	44.39	72.75	55.61
0.20	54.93	42.66	69.94	54.56

Table 2. Experiments on the ScanRefer and Sr3D/Nr3D datasets for different curvatures c of the hyperbolic space in the PSA loss.

Temperature τ	ScanRefer 0.25	ScanRefer 0.5	Sr3D(@0.25)	Nr3D(@0.25)
0.05	56.14	45.06	71.60	56.34
0.10	**56.90**	**45.58**	**73.10**	**56.97**
0.15	55.94	44.76	70.38	55.99
0.20	55.57	44.19	67.57	55.07

Table 3. Experiments on the ScanRefer and Sr3D/Nr3D benchmarks for the temperature coefficients of multi-modal contrast learning τ in the PSA loss.

3) In Table 3, we conduct an in-depth investigation into the effectiveness of the temperature coefficient τ within the PSA loss. Given the intricate nature of 3D scenes and language features, the choice of temperature coefficient τ will exert a more significant impact on the performance of the 3D visual grounding task. Considering the complexity of 3D scenes which encompass diverse geometric information and viewpoint variations, a higher τ help balance the similarity between different geometric features, thereby enhancing the robustness of the contrastive loss. However, it may also result in ambiguous similarity measurements, affecting precise localization of the target. Conversely, a lower τ may better capture fine-grained geometric information but also increase the model's sensitivity to 3D noise and disturbed objects. Hence, we have opted for $\tau = 0.1$, achieving higher performance while maintaining a balance between these competing factors.

ID	Language Components			Unique(%)		Multiple(%)	
	Obj.	Attri.	Rel.	0.25	0.5	0.25	0.5
(a)	—	—	—	86.75	69.76	50.11	38.58
(b)	✓	—	—	86.89	70.47	50.25	38.87
(c)	✓	✓	—	87.03	71.03	50.36	39.02
(d)	✓	—	GCN	87.31	71.81	50.52	39.68
(e)	✓	✓	GCN	**87.66**	**72.16**	**50.87**	**40.05**

Table 4. Ablation study of the language components effectiveness in the Flan-QS module on the ScanRefer dataset.

Method	Subsets			Overall	
	Attri	Rel	Attri+Rel	@0.25	@0.5
ScanRefer [3]	11.17	10.53	10.29	10.51	55.22
TGNN [8]	10.52	13.32	11.35	11.64	56.97
InstanceRefer [22]	14.74	13.71	13.81	13.92	55.61
BUTD-DETR [10]	12.30	12.11	11.86	11.99	8.95
EDA [20]	25.40	25.82	26.96	26.50	21.20
G^3-LQ	**26.61**	**26.92**	**27.88**	**27.55**	**21.89**

Table 5. Performance of grounding without object name proposed by EDA [20]. The accuracy of subsets (attribution and relationship) is measured by Acc@0.25IoU.

Method	Unique		Multiple		Overall	
	0.25	0.5	0.25	0.5	0.25	0.5
BUTD-DETR	85.62	68.64	46.07	35.51	52.0	40.5
EDA	90.91	75.33	51.71	40.66	57.6	45.8
G^3-LQ	**91.58**	**78.57**	**53.95**	**44.10**	**59.3**	**49.2**

Table 6. Performance on the *ScanRefer* benchmark using ground truth 3D boxes. The proposed method G^3-LQ demonstrates notable advantages over other remarkable methods.

Method	mAP@0.25	mAP@0.50
DETR+KPS+iter	59.9	—
3DETR with PointNet++	61.7	—
BUTD-DETR	63.0	43.8
EDA	64.1	45.3
G^3-LQ	**66.8**	**47.4**

Table 7. Performance of our proposed G^3-LQ method on the 3D object detection task (trained on the *ScanRefer* dataset).

Components of Language Scene Graph. As shown in Table 4, we perform ablation analysis of the text components in language scene graph construction on the ScanRefer benchmark. (a) represents the proposed G^3-LQ method without the Flan-QS module and the PSA loss. Firstly, the densely-aligned sub-methods (b) and (c) outperform the vanilla setting in (a). It provides compelling evidence of the effectiveness achieved by explicitly guiding query generation the query generation via decoupled language priors. Moreover, comparison between (b) and (c) indicates that attributions can identify the target object with additional color, texture, and shape features, resulting in improved performance. Secondly, when incorporating the relation components (shown in (c) and (e)), we observe a remarkable improvement of 1.13% (@0.5) and 1.03% (@0.5) on the unique and multiple subsets. The relation-aware phrases updated with the message passing mechanism, facilitate to understand the spatial layouts of 3D objects and encode long-range dependencies, which further encourages to generate precise object queries related to the descriptions.

Performance of grounding without object name. To evaluate the performance of our model, we conduct further experiments on the "*Grounding without Object Name*" setting proposed by EDA [20]. In this setting, object names are replaced with the generic term "*object*" to test whether our model can still accurately identify the referred object without relying on the specific object names. The language set is partitioned into four distinct subsets: only mentioning object attributes (∼15%), only mentioning spatial relationships (∼20%), mentioning both attributes and relationships (∼63%), and others (∼2%). It is an challenging yet practical task since it more closely resembles real-world scenarios where object labels may be missing or unreliable. The experimental results presented in Table 5 demonstrate that even without retraining, our method continues to outperform other state-of-the-art EDA in this setting, with 1.05% and 1.69% improvements of the overall performance. It further verifies the potential of our G^3-LQ method in capturing the intrinsic geometric properties of objects and complex spatial relationships.

Evaluation on the ScanRefer with GT box. The ScanRefer dataset serves as an essential benchmark for evaluating models' proficiency in comprehending natural language instructions and accurately localizing objects within 3D scene. However, it is important to note that the ScanRefer dataset does not provide additional GT boxes for candidate objects. Similar to EDA [20], we provide evaluation on the ScanRefer dataset by GT boxes for fair comparison. As shown in Table 6, the performance of our G^3-LQ model demonstrates a remarkable improvement particularly in the overall setting where the accuracy reaches **59.3**% and **49.2**%, respectively. To explain, GT bounding boxes provide precise information about the object's location, perfectly matching the real-world position. This accuracy helps avoid uncertainties introduced by errors in the object detection algorithm (*e.g.*, GroupFree [11]). By directly using GT bounding boxes for visual localization, the model can establish a more precise correspondence between textual descriptions and objects. The impressive results achieved without retraining underline the potential of our G^3-LQ model to significantly improve the efficiency and scalability of 3D

VG systems given more accurate object detection.

Detailed results on the Sr3D/Nr3D dataset. Regrettably limited by the length constraints, we have exclusively showcased the experimental findings pertaining to the "*overall*" and "*Hard*" subsets of the Nr3D and Sr3D datasets in the manuscripts. While this provides valuable insight into the performance of our approach, for a more comprehensive analysis, a detailed comparison of our method with existing outstanding approaches spanning all subsets (i.e., *easy*, *hard*, *view-dependent*, and *view-independent*) can be found in Table. 8. Specifically, in the Nr3D dataset, descriptions exhibit noteworthy intricacy (reference and spatial relationships) and details (geometric attributes), inducing additional challenges to 3D VG task. The experimental findings of Nr3D dataset showcased in the Table. 8 demonstrate the superiority of our G^3-LQ method in understanding complex scenes and utterances understanding. To explain, our proposed PAGE module contributes to distinguishing ambiguous objects by effectively modeling the underlying geometric shape and unraveling the complex relationships between vision-text features.

3D object detection performance on ScanNet. We adhere to the consistent experimental framework established by BUTD-DETR [10] and EDA [20] for assessing the efficacy of 3D object detection within the ScanNet dataset. Noteworthy, our proposed G^3-LQ approach is not explicitly crafted for 3D object detection. The remarkable performance shown in Table. 7 employs the identical model as utilized in the 3D visual grounding task, which further proves that our method enjoys high superiority. We analyze from the following aspects: (1) *Enriched visual representation*. Our approach enriches the visual representation of 3D objects, thereby elevating the semantic understanding of 3D scenes. Through the explicit integration of geometric features, our model achieves enhanced precision in determining the spatial position and attributes of 3D objects, resulting in a refined semantic comprehension of the target entities. (2) *Improved spatial understanding*. The explicit integration of geometric features in our method facilitates a more precise localization and attribute estimation of objects. This, in turn, enhances the model's ability to understand the spatial relationships between objects, leading to improved 3D object detection performance.

4. Qualitative Analysis

4.1. Visualization of Two-Stage Paradigm

The qualitative results of the 3D visual grounding task conducted on the ScanRefer dataset are visually showcased in Fig. 1, providing an intuitional depiction of the model's effectiveness in accurately localizing and distinguishing objects within the intricate 3D indoor scenes.

1) Our method demonstrates exceptional visual perception, allowing for the accurate identification of objects based on their geometric attributes (such as **appearance, size, and shape**) among multiple candidates belonging to the same class. This advancement is achieved through the explicit exploration of the abundant topological structures and geometric details present in 3D point clouds.

(2) The proposed G^3-LQ method demonstrates outstanding proficiency in complex spatial awareness, including relative distances, proximity, containment, and contact relationships. This is attributed to our proposed Poincare Semantic Alignment (PSA) loss, which captures the nonlinear mapping of visual-textual correlations in hyperbolic space and models the correspondence between visual and textual elements at different hierarchical levels.

4.2. Visualization of One-Stage Paradigm

In Fig. 2, we present a detailed visualization of the experimental results on the ScanRefer [3] dataset for our method and EDA [20] in the context of one-stage 3D visual grounding setting. These visualizations provide compelling evidence that our proposed G^3-LQ exhibits superior performance in localizing 3D objects, even without the assistance of detected bounding boxes. The key factor contributing to our method's performance is the proposed PSA loss, which effectively captures the complex nonlinear relationships between 3D scene and text descriptions. The PSA loss enables a comprehensive alignment of these two modalities, facilitating geometric-semantic consistency based solely on text descriptions. Moreover, the Fine-grained Language-guided Query Selection (Flan-QS) generates object proposals guided by fine-grained text priors, allowing our model to align rich semantic information with the visual features of the 3D scene in a more precise and granular manner.

4.3. Visualization on the Sr3D and Nr3D

The grounding results on the Sr3D and Nr3D datasets are visualized in Fig. 3 and Fig. 4, respectively. We can observe the effectiveness of our G^3-LQ approach in accurately localizing objects based on simple and natural descriptions. In the Sr3D dataset, the language descriptions are not only concise but also contain many expressions related to relative distances, such as "*near*", "*farthest*", "*close to*" and "*next to*" (see Fig. 3). Our G^3-LQ approach addresses the challenge by leveraging PAGE module to capture fine-grained spatial details and relationships between objects. Additionally, the Nr3D dataset provides more natural language descriptions for the localization of 3D objects, including descriptions of their relative sizes (*e.g.*, largest), heights(*e.g.*, higher), and ambiguous orientations (*e.g.*, above, beside, closest). Fig. 4 visually demonstrates the effectiveness of our approach in handling these challenging text descriptions. However, the existing EDA falls short in explicitly incorporating the geometric properties and spatial relation-

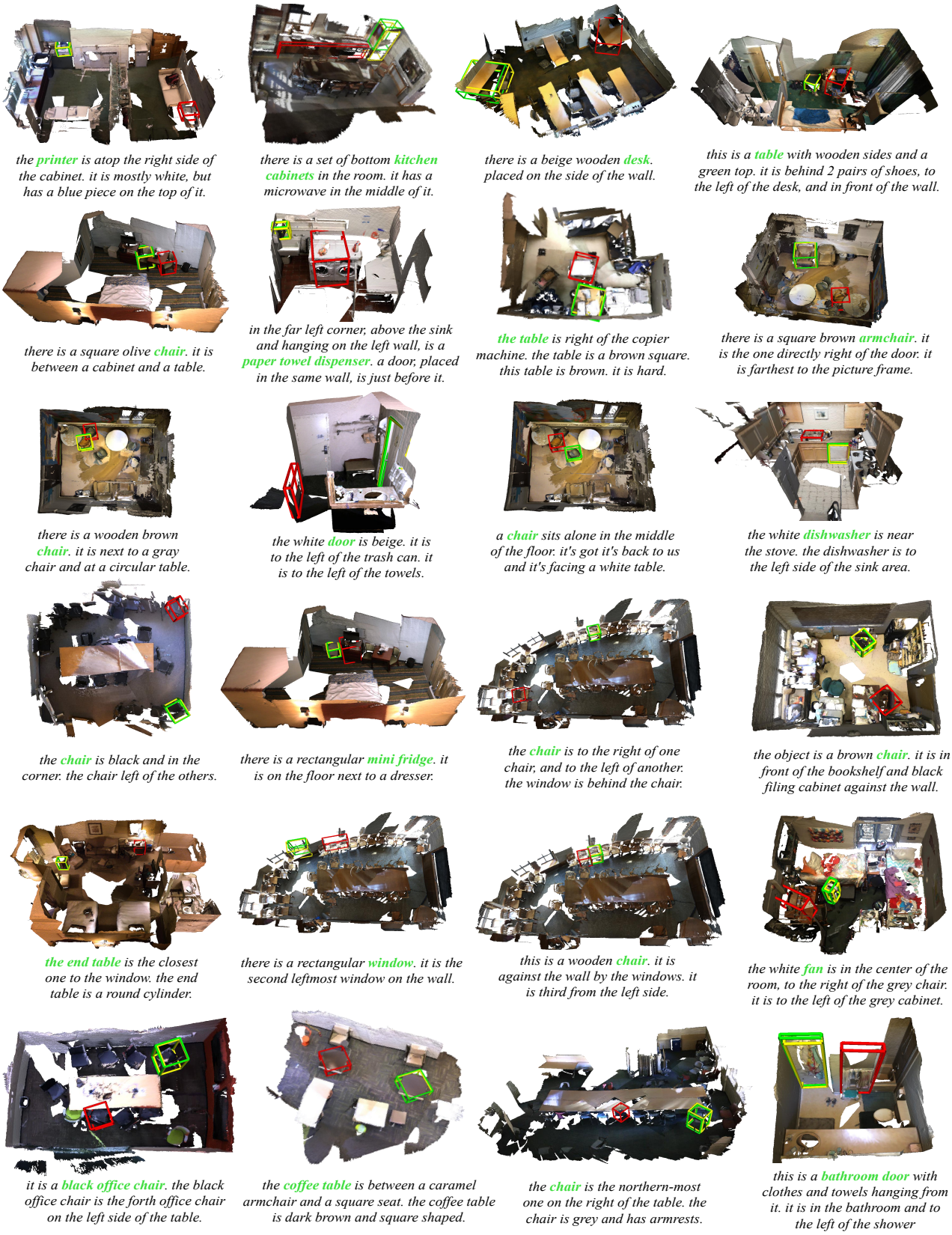

Figure 1. Additional visual experimental results on the ScanRefer dataset. The green bounding boxes denote the ground truth annotations, the red ones represent the grounding results obtained by EDA [20], and the yellow boxes indicate the localization results of our G^3-LQ method. Our results showcase outstanding performance.

Method	Venue	Nr3D					Sr3D				
		Overall	Easy	Hard	VD	VID	Overall	Easy	Hard	VD	VID
TGNN [8]	AAAI21	37.3	44.2	30.6	35.8	38.0	45.0	48.5	36.9	45.8	45.0
InstanceRefer [22]	ICCV21	38.8	46.0	31.8	34.5	41.9	48.0	51.1	40.5	45.4	48.1
3DVG [23]	ICCV21	40.8	48.5	34.8	34.8	43.7	51.4	54.2	44.9	44.6	51.7
LanguageRefer [17]	CoRL21	43.9	51.0	36.6	41.7	45.0	56.0	58.9	49.3	49.2	56.3
TransRefer3D [7]	MM21	48.0	56.7	39.6	42.5	50.7	57.4	60.5	50.2	49.9	57.7
SAT [21]	ICCV21	49.2	56.3	42.4	46.9	50.4	57.9	61.2	50.0	49.2	58.3
LAR [4]	NeurIPS22	48.9	58.4	42.3	47.4	52.1	59.4	63.0	51.2	50.0	59.1
3DRef [1]	WACV22	47.0	50.7	38.3	44.3	47.1	39.0	46.4	32.0	34.7	41.2
3D-SPS [12]	CVPR22	51.5	58.2	45.1	48.0	53.2	62.6	56.2	65.4	49.2	63.2
MVT [9]	CVPR22	55.1	61.3	49.1	54.3	55.4	64.5	66.9	58.8	58.4	64.7
BUTD-DETR [10]	ECCV22	54.6	60.7	48.4	46.0	58.0	67.0	68.6	63.2	53.0	67.6
EDA [20]	CVPR23	52.1	58.2	46.1	50.2	53.1	68.1	70.3	62.9	54.1	68.7
3D-VisTA(Scratch) [24]	ICCV23	57.5	65.9	49.4	53.7	59.4	69.6	72.1	63.6	57.9	70.1
G^3-LQ (Ours)	—	57.8	62.0	50.7	53.8	57.1	73.1	74.7	66.3	57.2	74.0

Table 8. Performance on SR3D/NR3D datasets by Acc@0.25IoU as the metric. We have highlighted the top-performing three methods in red (best viewed in colors). Our G^3-LQ method showcases an admirable performance compared with most prevailing methods.

Figure 2. Illustration of the **One-Stage** grounding results of our proposed G^3-LQ framework on the ScanRefer [3] dataset. The green marks are the Ground Truth boxes and the red marks denote the grounding results of the SOTA EDA [20] method, while the yellow marks represent the detected boxes of our method.

ships inherent in 3D objects. Consequently, it fails to capture the intricate details essential for achieving precise 3D visual grounding.

4.4. Visualization of Failure Cases

While our method has achieved state-of-the-art performance, it is important to acknowledge that there are still

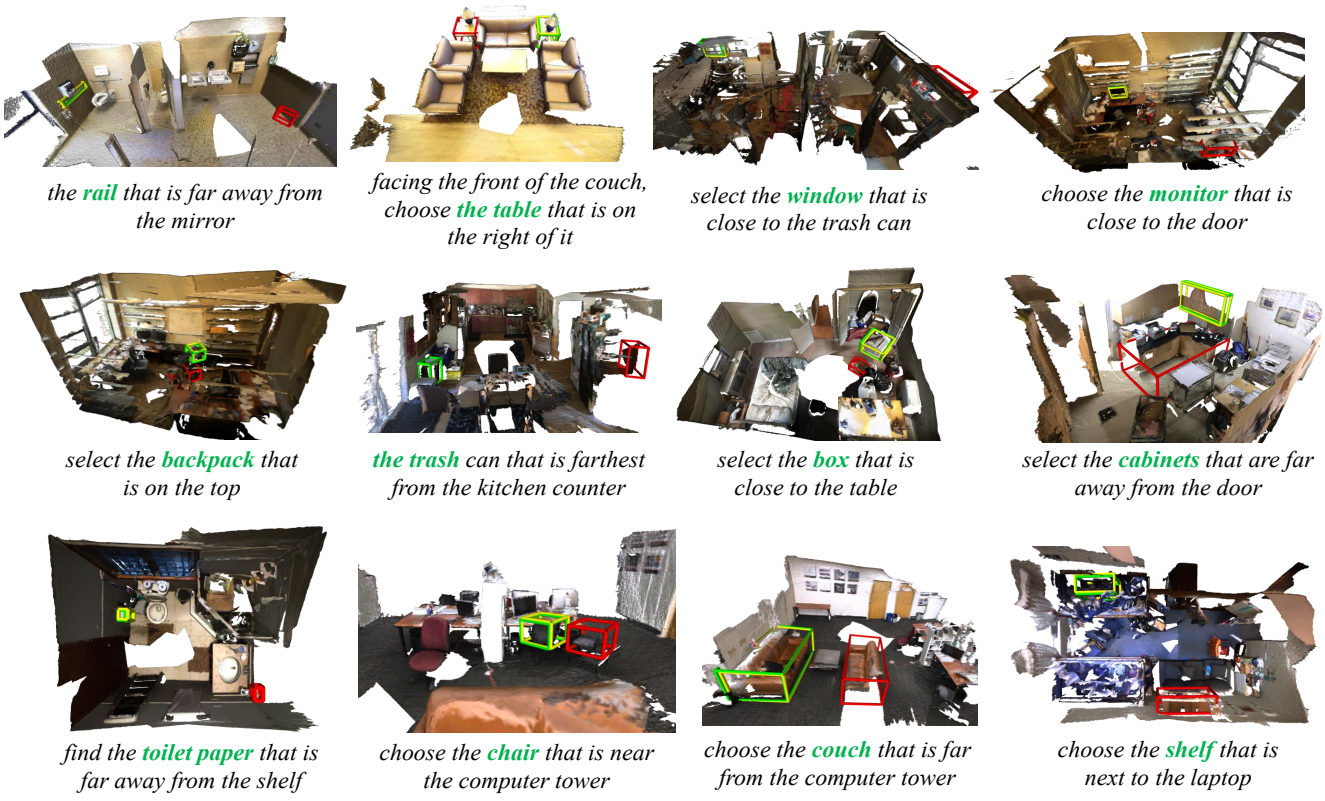

Figure 3. Qualitative results of the EDA [20] and our proposed G^3-LQ on the Sr3D dataset. The green marks are the ground truth boxes. The yellow and the red marks denote the grounding results of our G^3-LQ methods and EDA method.

a significant number of instances where failures occur. One aspect that we have identified as contributing to the failure occurrences is the low precision in **detecting both large and small objects**. This issue stems from the performance limitations of the 3D object detector used in our method. To address this limitation, we are exploring alternative detection frameworks [11, 16] and evaluating their effectiveness in enhancing the accuracy and robustness of our method. Another factor that may contribute to our model's failure in certain scenarios is the lack of consideration of multi-scale point cloud visual representations, which enables to capture the full range of spatial context and details. The problem of ambiguous language descriptions is a significant challenge for 3D visual grounding, as it can lead to poor understanding of the semantic content of text and imprecise localization of referring objects. This challenge is particularly pronounced in situations where there are multiple similar objects, object occlusions, or cluttered scenes.

5. Limitation and Future Study

Limitation. In this paper, we propose a G^3-LQ framework that explicitly models **G**eometric-aware visual representations and **G**enerates fine-**G**rained **L**anguage-guided object **Q**ueries. While this approach has demonstrated promising experimental performance in **indoor** 3D scenes, its effectiveness have underexplored in complex **outdoor** environments (*e.g.*, robot navigation and autonomous driving) as well as multi-view scenarios. Furthermore, as our G^3-LQ method predominantly focuses on individual object grounding, its performance is balanced on a knife-edge when confronted with scenarios involving multiple objects referenced within the text descriptions. Finally, the proposed G^3-LQ approach is dedicated to the 3D visual grounding task centered around *point cloud* modeling, and it does not venture into the exploration of 3D visual grounding tasks that rely on *volumetric* or *multi-view image* representations.

Future Study. In the future, there are still significant problems worthy to be explored. Among this, future research on building larger, more diverse, and complex benchmarks to comprehensively explore the performance and generalization capabilities of 3D VG models is desperately in need. Such datasets should encompass a variety of data types, including point clouds, voxels, multi-view images, fine-grained textual information. Based on datasets, our aspiration is to develop a all-encompassing 3D multi-modal large model that can accommodate a wide range of 3D vision-language tasks. This endeavor is envisioned to make a substantial contribution to the broader research community.

Figure 4. Qualitative results of the EDA [20] and our proposed G^3-LQ on the Nr3D dataset. The green marks are the ground truth boxes. The yellow and the red marks denote the grounding results of our G^3-LQ methods and EDA method.

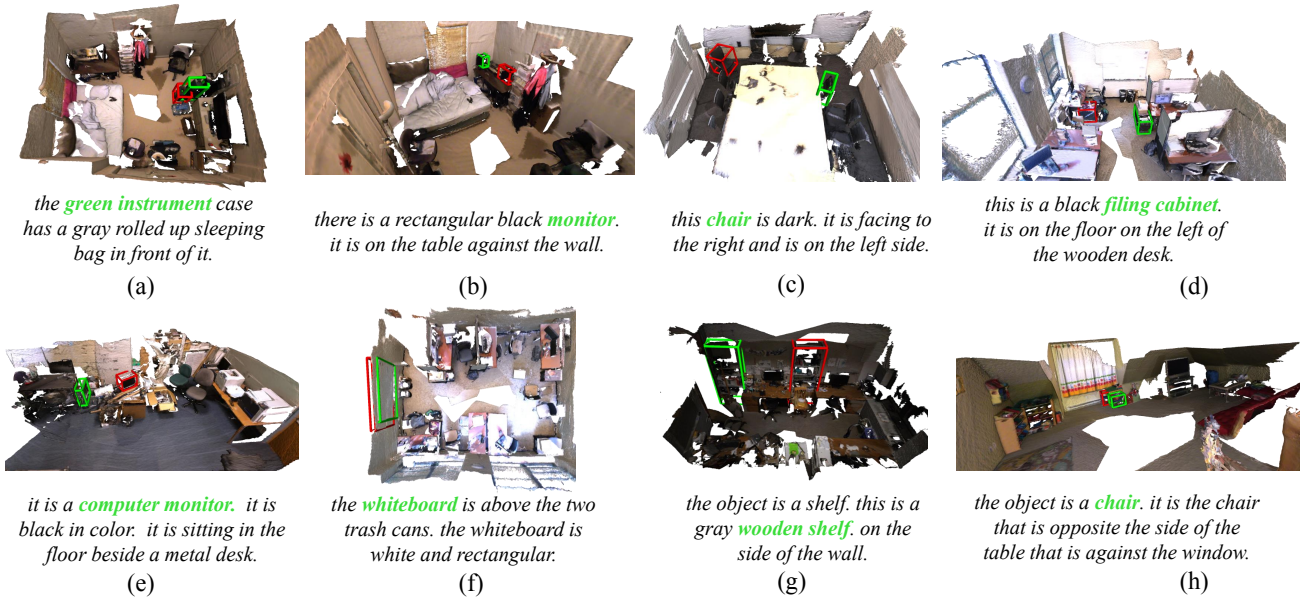

Figure 5. Illustration of the failure cases of our proposed G^3-LQ framework on the ScanRefer [3] dataset. The green marks are the Ground Truth boxes and the red marks denote the grounding results of our G^3-LQ methods

References

[1] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3dref-transformer: Fine-grained object identification in real-world scenes using natural language. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 3941–3950, 2022. 6

[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 *Proceedings of the European Conference on Computer Vision*, pages 422–440. Springer, 2020. 1

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision*, pages 202–221. Springer, 2020. 1, 3, 4, 6, 8

[4] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems*, 35:20522–20535, 2022. 6

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1

[6] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7409–7419, 2022. 2

[7] 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. 6

[8] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1610–1618, 2021. 3, 6

[9] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 6

[10] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *Proceedings of the European Conference on Computer Vision*, pages 417–433. Springer, 2022. 1, 3, 4, 6

[11] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2949–2958, 2021. 1, 3, 7

[12] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022. 6

[13] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[15] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[16] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019. 7

[17] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 6

[18] 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. 2

[19] Yuan Wang, Min Cao, Zhenfeng Fan, and Silong Peng. Learning to detect 3d facial landmarks via heatmap regression with graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2595–2603, 2022. 2

[20] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning. *arXiv preprint arXiv:2209.14941*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[21] Zhengyuan Yang, Songyang Zhang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1856–1866, 2021. 6

[22] 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 International Conference on Computer Vision*, pages 1791–1800, 2021. 3, 6

[23] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer:relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2928–2937, 2021. 6

[24] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. *arXiv preprint arXiv:2308.04352*, 2023. 6