

WHERE, NOT WHAT: COMPELLING VIDEO LLMs TO LEARN GEOMETRIC CAUSALITY FOR 3D-GROUNDING

Yutong Zhong

New York University
yz9760@nyu.edu

ABSTRACT

Multimodal 3D grounding has garnered considerable interest in Vision-Language Models (VLMs) [1] for advancing spatial reasoning in complex environments. However, these models suffer from a severe "2D semantic bias" that arises from over-reliance on 2D image features for coarse localization, largely disregarding 3D geometric inputs and resulting in suboptimal fusion performance. In this paper, we propose a novel training framework called What-Where Representation Re-Forming (W2R2) to tackle this issue via disentangled representation learning and targeted shortcut suppression. Our approach fundamentally reshapes the model's internal space by designating 2D features as semantic beacons for "What" identification and 3D features as spatial anchors for "Where" localization, enabling precise 3D grounding without modifying inference architecture. Key components include a dual-objective loss function with an Alignment Loss that supervises fused predictions using adapted cross-entropy for multimodal synergy, and a Pseudo-Label Loss that penalizes overly effective 2D-dominant pseudo-outputs via a margin-based mechanism. Experiments conducted on ScanRefer and ScanQA demonstrate the effectiveness of W2R2, with significant gains in localization accuracy and robustness, particularly in cluttered outdoor scenes.

Index Terms— LLM, 3D grounding, Dual-objective loss, Multimodal Fusion

1. INTRODUCTION

3D grounding matters, humans live in a 3D world and use natural language to interact with a 3D scene.[2] However, progress is constrained by data scarcity: state-of-the-art 3D models are "limited by datasets with a small number of annotated data." [3] To mitigate this, VG-LLM, recent systems pair a 2D encoder with a 3D geometry encoder, explicitly integrate[s] 3D visual geometry priors into MLLMs, processing images by "both a conventional visual encoder and the newly integrated 3D visual geometry encoder. [4] A strong geometry backbone here is VGGT,

which directly infers all key 3D attributes of scenes of its view.[5] Another line aligns modalities via a shared embedding space: ULIP unifies image–text–point-cloud representations, ULIP-2 scales training by automatically generating holistic 3D-shape descriptions, Point-Bind aligns points with image/language/audio/video for broad transfer, OpenScene co-embeds 3D points with text and pixels in CLIP space, and LERF/LangSplat build 3D language fields for open-ended queries.[3, 6, 7, 2, 8, 9]

Many claimed 3D benchmarks can be easily solved by applying a generic 2D VLM to rendered views indicating a very strong 2D bias.[10] Overreliance on 2D appearance / semantics when fusing with 3D cues has been documented as "2D-cheating"; object-centric, chain-of-analysis evaluations have been proposed, yet they remain evaluation level and do not disentangle 2D and 3D representations during training[11]

To solve this issue, we introduce W2R2, a multimodal 3D grounding training framework that changes internal representations without changing the inference architecture. First, we explicitly mine a 2D-shortcut answer through the use of attention anchors that yield a coarse 2D-only localization. Next, we introduce a Shortcut Suppression Loss that penalizes confident agreement with this shortcut and implicitly pushes the model away from the 2D route. Finally, a Representation Re-Forming Loss enforces a what–where split, where 2D features carry semantics but 3D features define grounding. This keeps the shortcut semantically similar but spatially removed from the final answer and fosters reliance on 3D cues.

2. PRELIMINARY

In contrast to 3D-LLMs that work directly with point clouds, our framework reconstructs 3D features from multi-view 2D images and combined features of the original 2D image with the reconstructed 3D geometric features is provided in the form of 3D grounding. Our method takes careful advantage of the strong pretrained semantic capacity of vision–language models (VLMs) on 2D image, but this leads to an inherent challenge that has yet to be fully resolved. The question is whether the model is actually learning to leverage the reconstructed 3D geometric information or is acting on the "more familiar" 2D semantics as a quick escape.[12, 13, 14, 15] We

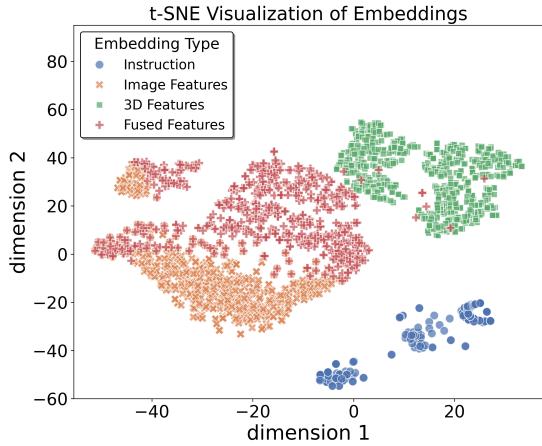


Fig. 1: t-SNE visualization of feature distributions, highlighting a 2D semantic bias. The Fused Features show a much closer proximity to the 2D Image Features than to the 3D Features.

designed the following two diagnostic experiments to test this question.

First, we carried out a behavioral diagnostic to operationalize the theoretical reliance of the model on 3D features. We ablated the 3D input branch and examined localization on the ScanRefer dataset under the condition of using 2D image features-in-other-words-and-no-3D data. To our surprise, as shown in Table 1, modern VLMs have localization accuracy that is still clearly above a random-guess baseline despite having neither 2D nor 3D data. This indicated the presence of a ubiquitous 2D semantic shortcut in which models are relying on 2D semantics and are not actually doing 3D geometry awareness.

The model therefore learns to take advantage of strong 2D semantic information without developing a deep understanding of accurate 3D geometric relations. To explore the internal reason for this shortcut behavior, we examined the model’s latent representation space via t-SNE. Results from Figure 1 establish a clear representation misalignment between modalities: 2D image features and corresponding reconstructed 3D features inhabit very different regions in latent space. Significantly, the fused representation collapses toward the 2D feature cluster, remaining far away from the clustered 3D features. This clustering results in the assumption that fusion has weight on 2D semantics over 3D geometry. Overall, these two experiments highlight the main issue of VLM-based 3D localization: there is an inherent 2D semantic bias. The evident “shortcut behavior” (Table 1) is a symptom of the misalignment of the representations (Figure 1). Since fusion occurs based on strong 2D features, the model does not have to really learn or be dependent on the more complicated 3D geometric evidence. This compromised use of 3D evidence is an impor-

Table 1: Diagnostic results on ScanRefer using only 2D inputs. All models achieve accuracy far exceeding random chance, indicating a heavy reliance on the 2D semantic features and a failure to effectively utilize 3D geometric features.

Method	ScanRefer	
	Acc@0.25	Acc@0.5
<i>Only 2D Input:</i>		
SeqVLM	47.8	39.3
VG-LLM	47.2	35.7
VLM-Grounder	50.1	38.1
<i>2D + 3D Input:</i>		
SeqVLM	55.6	49.6
VG-LLM	53.2	49.7
VLM-Grounder	62.4	53.2

tant barrier to higher accuracy 3D grounding. Therefore, the main goal of this work is to cut this bias and transform the representation space to allow for actual 3D scene understanding.

3. PROPOSED METHOD

Instead of trying to eliminate 2D semantics, we tackle 2D shortcut learning in 3D LLMs by applying What–Where Representation Re-Forming (W2R2), a training scheme that reformulates internal representations. W2R2 uses two forward passes with shared parameters: full fused and 2D-only, and applies an objective to pull-push the model towards 3D-informed predictions, while at the same time pushing against overly strong 2D-only source predictions. The push-pull structure also decouples the representation space: the 2D features have semantics/coarse cues (what), and the 3D features already were dominating precise spatial grounding (where) for encouraging the model to use 3D geometry for grounding.

3.1. Baseline & Notation

Given multi-view RGB images $I = \{I_i\}_{i=1}^V$ with camera intrinsics/extrinsics $\{K_i, R_i, T_i\}_{i=1}^V$, a 3D encoder produces geometry-aware features $F_{3D} = \mathcal{E}_{3D}(I)$ [16], while a 2D encoder provides semantic features $F_{2D} = \mathcal{E}_{2D}(I)$. We fuse the two streams via a generic operator $\Phi(\cdot, \cdot)$ (e.g., concatenation / cross-attention) and feed them with language query q to a decoder/LLM:

$$o_{\text{fused}} = \mathcal{D}(\Phi(F_{2D}, F_{3D}), q). \quad (1)$$

We supervise the fused prediction with a standard alignment loss, while CE stands for cross-entropy

$$\mathcal{L}_{\text{align}} = \text{CE}(o_{\text{fused}}, y), \quad (2)$$

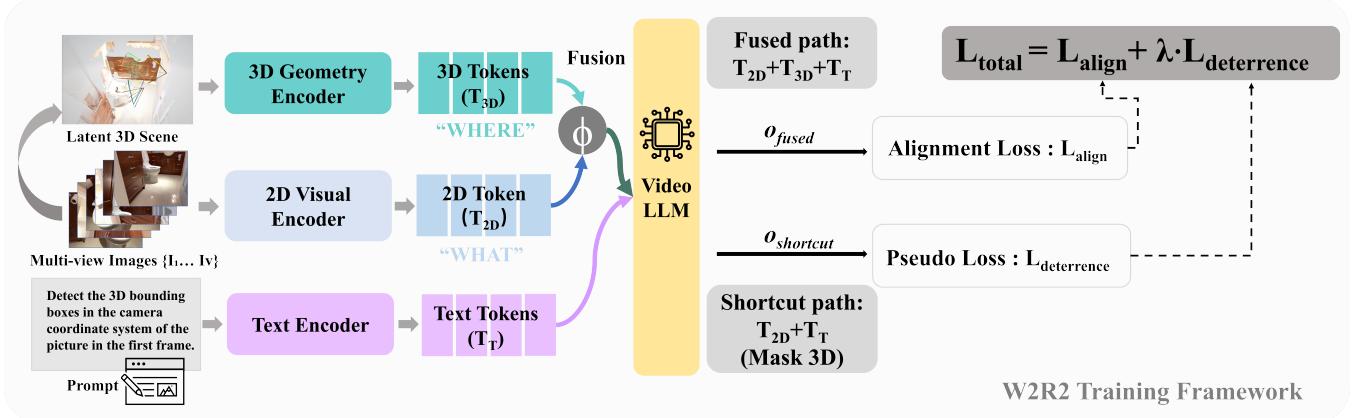


Fig. 2: Overview of our proposed W2R2 training framework.

3.2. Formalizing the 2D Shortcut

To expose the semantic shortcut, we ablate the 3D branch and define a 2D-only prediction:

$$o_{\text{short}} = \mathcal{D}(\Phi(F_{2D}, \mathbf{0}), q). \quad (3)$$

This path captures coarse localization driven by strong 2D semantics and will be used for targeted regularization; gradients through this branch will be blocked in the push term.

3.3. W2R2: Pull–Push Training

At each iteration we run two forward passes with shared parameters θ :

$$o_{\text{fused}} = f_\theta(\Phi(F_{2D}, F_{3D}), q), \quad (4)$$

$$o_{\text{short}} = f_\theta(\Phi(F_{2D}, \mathbf{0}), q). \quad (5)$$

W2R2 reshapes the representation space via a pull–push objective: (i) *Pull* aligns the fused output with the ground truth using $\mathcal{L}_{\text{align}}$. (ii) *Push* discourages over-reliance on the 2D path by penalizing a too-good 2D-only solution:

$$\mathcal{L}_{\text{deterrence}} = \max(0, s(\text{IoU3D}(o_{\text{short}}), y) - \mu) \quad (6)$$

where $s(\cdot, \cdot)$ is a task similarity (e.g., IoU_{3D} for grounding), $\mu \in (0, 1)$ is a tolerance margin, and $\text{stopgrad}(\cdot)$ blocks gradients through the shortcut branch to prevent degenerate updates that only worsen the 2D path.

3.4. Total Objective & Effect

The overall objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{deterrence}}, \quad \lambda > 0. \quad (7)$$

The pull term rewards 3D-informed fused solutions, while the push term activates only when the 2D-only output is too

close to the ground truth, thereby deterring shortcut reliance. This push–pull dynamics reshapes the representation space: 2D features retain semantics/coarse cues (*what*), whereas 3D features dominate fine-grained spatial grounding (*where*).

4. EXPERIMENT

4.1. Data and evaluation metrics

Dataset. We conduct experiments on three popular benchmarks for 3D grounding and captioning: ScanRefer[17], Scan2Cap[18], and ScanQA[19]. For 3D visual grounding, we test our model on ScanRefer, which requires localizing unique objects in single-target scenarios. For dense captioning, we use the Scan2Cap benchmark, which involves generating descriptive captions for all objects in 3D scenes. For question answering, we employ ScanQA for spatial reasoning tasks. To perform evaluations, we follow prior works [20] and use the validation of ScanRefer, Scan2Cap, and ScanQA.

Metrics. We adopt widely used evaluation metrics for each benchmark. For ScanRefer, we report threshold-based accuracy metrics, specifically Acc@0.25 and Acc@0.5, where a prediction is considered correct if its Intersection over Union (IoU) with the ground truth exceeds 0.25 and 0.5, respectively. For Scan2Cap, we apply CIDEr@0.5IoU and BLEU4@0.5IoU (denoted as C@0.5 and B-4@0.5), combining traditional image captioning metrics with the IoU between predicted and reference bounding boxes. For ScanQA, we use Exact Match accuracy (denoted as EM) to measure precise alignment with ground-truth answers.

4.2. Main Results

As shown in Table 2, our W2R2 framework achieves the best overall performance (**55.1**) across three benchmark tasks, demonstrating its superiority in complex 3D scene understanding. In the core 3D visual grounding task (ScanRefer),

Table 2: Overall comparison with state-of-the-art methods on three benchmark datasets.

Model	ScanRefer		Scan2Cap		ScanQA	Overall
	Acc@0.25	Acc@0.5	B-4@0.25	C@0.25	EM	
SPAR	48.8	43.1	-	-	-	-
SeqVLM	55.6	49.6	41.1	82.3	30.2	51.8
VLM-Grounder	62.4	53.2	45.1	70.1	29.1	52.0
SeeGround	44.1	39.4	47.1	80.6	26.9	47.2
Video-3D LLM	58.1	51.7	41.3	83.8	30.1	53.0
W2R2 LLM	63.6	54.2	44.8	82.1	30.8	55.1

W2R2 sets a new state-of-the-art with **63.6%** Acc@0.25 and **54.2%** Acc@0.5, validating the effectiveness of suppressing 2D shortcuts to enhance 3D spatial awareness. Similarly, in 3D question answering (ScanQA), W2R2 leads with a **30.8%** EM score, outperforming all competing methods. For 3D dense captioning (Scan2Cap), W2R2 delivers highly competitive results, comparable to leading approaches. These findings highlight that W2R2 not only significantly boosts localization accuracy but also maintains robust language generation, confirming the effectiveness and robustness of our Representation Re-Forming approach.

4.3. Ablation Studies

In this section, we analyze the impact of two key hyper-parameters in our framework: the suppression strength, λ , of the pseudo-path loss, and its activation threshold, μ . First, we investigate the effect of λ on 3D grounding performance using the ScanRefer dataset. As shown in Figure 3(a), increasing λ from 0.1 to 1.0 leads to a substantial improvement in the Acc@0.25 metric, which peaks at 68.1%. Interestingly, the more stringent Acc@0.5 metric continues to benefit from stronger suppression, reaching its maximum of 54.2% at $\lambda = 1.5$. This suggests that a sufficiently high suppression strength ($\lambda \geq 1.0$) is crucial for achieving high-precision localization. Next, we evaluate the activation threshold μ on the Scan2Cap dataset. The results in Figure 3(b) show that captioning quality, measured by B-4@0.25 and CIDEr (C@0.25), significantly improves as μ is increased from a low value of 0.1. The C@0.25 metric peaks at 82.1% with $\mu = 0.7$, while the B-4 score is maximized at 45.1% with $\mu = 0.9$. This validates our design rationale: an overly aggressive threshold (low μ) is detrimental, whereas $\mu = 0.7$ provides an optimal balance for activating the shortcut suppression. Consequently, for all our main experiments, we set $\lambda = 1.5$ to prioritize performance on the more challenging high-IoU metric, and $\mu = 0.7$ as our default configuration.

5. CONCLUSION

In this paper, we address the prevalent 2D semantic bias in multimodal 3D grounding tasks by proposing the What-

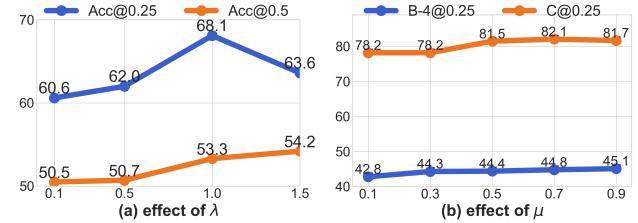


Fig. 3: Ablation studies for hyper-parameters λ and μ .

Where Representation Re-Forming (W2R2) framework. Departing from prior methods that optimize feature fusion, W2R2 fundamentally restructures the model’s representation space to enforce a clear division: 2D features are directed toward semantic identification (“what”), while 3D features are prioritized for spatial localization (“where”). This approach mitigates over-reliance on 2D cues and enhances geometric utilization. Empirical results on multiple benchmarks show substantial gains in grounding accuracy without additional inference costs, all while maintaining the model’s original captioning performance. Beyond simple suppression, future research will investigate the potential for a symbiotic relationship between the 2D and 3D pathways. We aim to explore whether the coarse localization from the 2D shortcut can actively guide the 3D geometric reasoning, transforming the shortcut from a liability into a complementary signal for mutual enhancement.

6. REFERENCES

- [1] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al., “Spatial mental modeling from limited views,” *arXiv e-prints*, pp. arXiv–2506, 2025.
- [2] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister, “Langsplat: 3d language gaussian splatting,” *arXiv preprint arXiv:2312.16084*, 2023.
- [3] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos

- Niebles, and Silvio Savarese, “Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1179–1189.
- [4] Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang, “Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors,” *arXiv e-prints*, pp. arXiv–2505, 2025.
- [5] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny, “Vggt: Visual geometry grounded transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [6] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al., “Ulip-2: Towards scalable multimodal pre-training for 3d understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27091–27101.
- [7] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al., “Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following,” *arXiv preprint arXiv:2309.00615*, 2023.
- [8] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik, “Lerf: Language embedded radiance fields,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19729–19739.
- [9] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al., “Openscene: 3d scene understanding with open vocabularies,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 815–824.
- [10] Jiahe Jin, Yanheng He, and Mingyan Yang, “Revisiting 3d llm benchmarks: Are we really testing 3d capabilities?,” *arXiv preprint arXiv:2502.08503*, 2025.
- [11] Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang, “Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24570–24581.
- [12] Haoyuan Li, Yanpeng Zhou, Yufei Gao, Tao Tang, Jianhua Han, Yujie Yuan, Dave Zhenyu Chen, Jiawang Bian, Hang Xu, and Xiaodan Liang, “Does your 3d encoder really work? when pretrain-sft from 2d vlms meets 3d vlms,” *arXiv preprint arXiv:2506.05318*, 2025.
- [13] Jiaxin Huang, Ziwen Li, Hanlve Zhang, Runnan Chen, Xiao He, Yandong Guo, Wenping Wang, Tongliang Liu, and Mingming Gong, “Surprise3d: A dataset for spatial understanding and reasoning in complex 3d scenes,” *arXiv preprint arXiv:2507.07781*, 2025.
- [14] Weipeng Deng, Jihan Yang, Runyu Ding, Jiahui Liu, Yiqiang Li, Xiaojuan Qi, and Edith Ngai, “Can 3d vision-language models truly understand natural language?,” *arXiv preprint arXiv:2403.14760*, 2024.
- [15] Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu, “Beyond semantics: Rediscovering spatial awareness in vision-language models,” *arXiv preprint arXiv:2503.17349*, 2025.
- [16] Danila Rukhovich, Anna Vorontsova, and Anton Konushin, “Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2397–2406.
- [17] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner, “Scanrefer: 3d object localization in rgbd scans using natural language,” in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 202–221, Springer International Publishing.
- [18] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang, “Scan2cap: Context-aware dense captioning in rgbd scans,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3193–3203.
- [19] 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*, 2022, pp. 19129–19139.
- [20] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner, “Scanrefer: 3d object localization in rgbd scans using natural language,” in *European conference on computer vision*. Springer, 2020, pp. 202–221.