Towards Edge Device Perception with Vision-Language Models for 3D Visual Grounding

Manan Arora Vu Nguyen Sarang Rajeev
aromanan@umich.edu nvtnghia@umich.edu sarangr@umich.edu

1 Introduction

3D Visual Grounding (3DVG) is the task of localizing a specific object within a 3D scene given a natural language query. This capability is essential for machines to meaningfully perceive and interact with the physical world, with direct applications in autonomous vehicles, mobile robotics, drones, and augmented reality.

Historically, 3DVG methods required large-scale annotated 3D datasets and specialized architectures trained end-to-end on sparse point cloud data. Recent work has dramatically changed this landscape. LangSplat [6] (CVPR 2024) demonstrated open-vocabulary object localization in 3D Gaussian Splatting representations without task-specific training. SeeGround [1] (CVPR 2025) extended this paradigm to standard point cloud scenes, achieving state-of-the-art zero-shot grounding by bridging 3D data to a pre-trained 2D VLM through query-aligned rendering and spatial text descriptions.

Despite SeeGround's impressive results, its original evaluation relied exclusively on Qwen2-VL-72B [3] as the reasoning backbone, a model with 72 billion parameters that demands approximately 144 GB of GPU memory at `bfloat16` precision, necessitating multi-GPU server infrastructure well beyond any edge platform. The SeeGround paper's own ablations [1] confirm that backbone capacity matters: swapping Qwen2-VL-72B for Qwen2-VL-7B on Nr3D drops overall accuracy from 46.1% to 33.3%, a gap of nearly 13 points. This sensitivity raises a natural question: is the performance cliff a consequence of *scale*, or of *architecture*?

We address this directly by substituting the 72B backbone with Qwen3-VL [2], a newer model generation that replaces the standard M-RoPE positional encoding of Qwen2-VL with an Interleaved Multimodal Rotary Position Embedding (Interleaved-MRoPE) [2], distributing temporal, height, and width positional signals across attention heads rather than concatenating them. This architectural change is directly relevant to SeeGround's spatially structured visual prompts, which embed numeric object IDs at precise pixel locations within rendered point cloud images. We evaluate Qwen3-VL at the 4B and 2B parameter scales, models that fit within 8–16 GB of unified memory, as drop-in replacements for the 72B backbone, with all other pipeline components held fixed.

2 Background

2.1 The SeeGround Framework

SeeGround [1] is a training-free, zero-shot pipeline that bridges 3D point cloud data to a 2D VLM through four stages (Figure 1).

1. **Query Parsing.** A VLM decomposes the query into a *target* object and an *anchor* reference object, constraining subsequent viewpoint selection to a semantically relevant scene region.
2. **Perspective Adaptation.** A virtual camera is positioned relative to the anchor's 3D centroid via a look-at-view transform [1], producing a query-aligned rendered view. This strategy improves Nr3D accuracy by +2.8 points over bird's-eye view [1].
3. **3D-to-2D Rendering.** The point cloud is projected via PyTorch3D [8] and each object is labeled with a unique numeric ID using depth-aware occlusion filtering, forming a visual prompt.
4. **Fusion Alignment and Grounding.** The labeled image is combined with structured spatial metadata from an Object Lookup Table (OLT) and passed to the VLM, which predicts the target object ID. This module contributes a +3.6 point accuracy gain on Nr3D over text-only input [1].

2.2 The Edge Deployment Problem

SeeGround's original evaluation used Qwen2-VL-72B [3], requiring approximately 144 GB of GPU memory at `bfloat16` precision and at least four NVIDIA A100 80 GB GPUs consuming 1–2 kW continuously. This is incompatible with edge platforms on three axes:

- **Memory.** Even the highest-end Jetson AGX Orin provides only 64 GB of unified memory; the 72B model exceeds this by 2× at minimum.
- **Power.** Jetson TDP envelopes of 10–60 W are one to two orders of magnitude below the 300–400 W per GPU of server hardware.
- **Latency.** Cloud offloading introduces network round-trip delays unacceptable in safety-critical navigation and manipulation.

The SeeGround paper's own ablations show that replacing the 72B backbone with Qwen2-VL-7B drops Nr3D accuracy by nearly 13 points [1]. However, this compares models from the same architectural family, conflating scale with generation. Whether a newer sub-4B architecture can close this gap is precisely the question this work addresses.

3 Method

This work evaluates Qwen3-VL at the 2B and 4B parameter scales as drop-in replacements for the Qwen2-VL-72B backbone within the SeeGround framework, extending the backbone scaling curve from four data points to six and probing the sub-7B regime that is actually accessible on edge hardware.

3.1 Backbone Substitution

The original SeeGround paper [1] evaluated Qwen2-VL-7B, InternVL2-8B, InternVL2-26B, and Qwen2-VL-72B, with the 72B model achieving 46.1% overall accuracy on Nr3D. We substitute this backbone with **Qwen3-VL** [2] at the **2B** and **4B** scales using identical prompts and pipeline components, making the substitution fully drop-in.

Qwen3-VL is motivated by two considerations. First, as the direct successor to Qwen2-VL, it retains full input format and API compatibility with SeeGround's pipeline. Second,

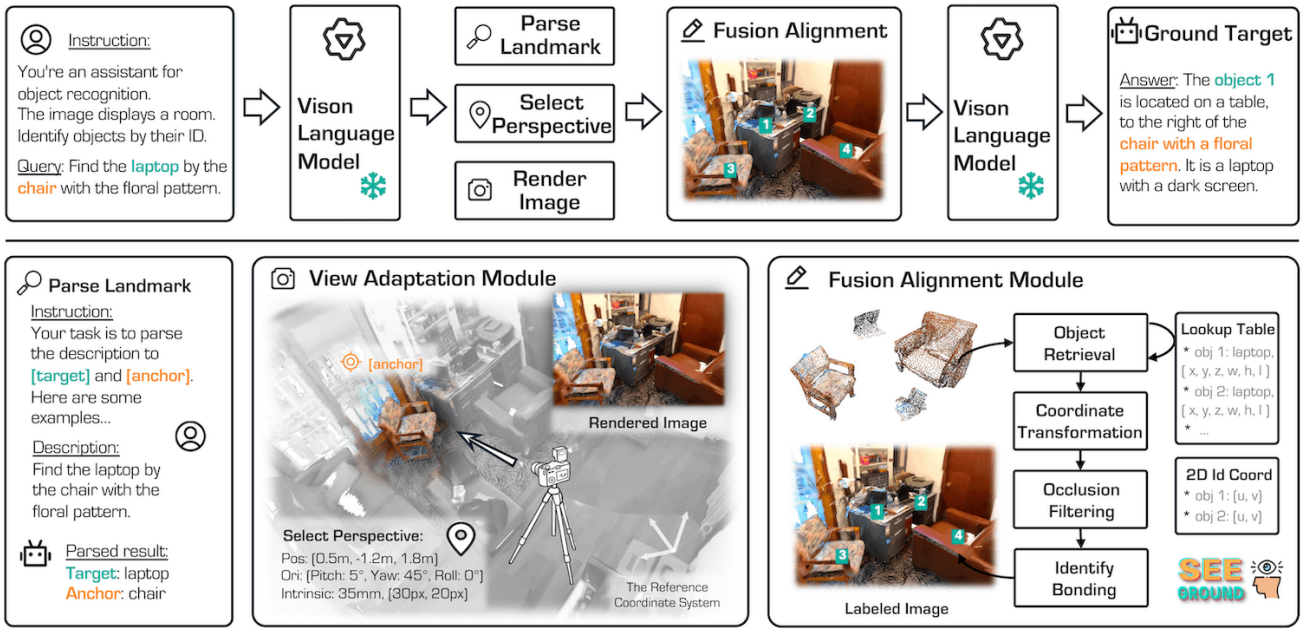


Figure 1: Overview of the SeeGround framework [1]. *Top:* full pipeline from query parsing to target ID prediction. *Bottom:* detail of the Parse Landmark, View Adaptation, and Fusion Alignment modules.

it introduces Interleaved-MRoPE [2], which distributes temporal, height, and width positional signals across attention heads rather than partitioning them. This produces finer spatial localization within rendered images, directly relevant to reading the numeric object IDs that SeeGround overlays at precise pixel locations. We hypothesize that this improvement, combined with stronger instruction following in the Qwen3 generation, allows smaller models to compensate for reduced parameter count on SeeGround's structured reasoning task.

We acknowledge that comparing Qwen3-VL-4B against Qwen2-VL-72B conflates model scale with model generation. A clean separation would require a Qwen3-VL-72B or Qwen2-VL-4B evaluation, which we identify as the most important direction for future work.

3.2 Controlled Experimental Design

All pipeline components outside the VLM backbone are held fixed across every evaluated model, making this a clean ablation over backbone scale and architecture:

- **Object Lookup Table.** Built from Mask3D [7] instance predictions on ScanNet scenes, identical to the original SeeGround setup [1].
- **Perspective Adaptation.** Camera placement heuristics, look-at-view transform parameters, and rendered image resolution of 680×680 pixels are unchanged.
- **Fusion Alignment.** The depth-aware visual prompting scheme, occlusion filtering, and structured spatial text description format are identical across all configurations.
- **Inference.** All models are served via `vLLM` [5] on a single GPU. The 4B and 2B models occupy approximately 8 GB and 4 GB at `bfloat16` precision, requiring no quantization and no tensor parallelism.

3.3 Evaluation Protocol

We evaluate on **Nr3D** [4], which provides 41,503 natural language referral queries over ScanNet scenes with ground-truth 3D bounding boxes supplied directly for all candidate objects. This removes 3D detection accuracy as a confounding variable, isolating the grounding capability of the backbone. Performance is reported at IoU thresholds of 0.25 and 0.50 across the Unique, Multiple, and Overall subsets, consistent with [1].

4 Results

4.1 Quantitative Results

Table 1 reports grounding accuracy on Nr3D across the Unique, Multiple, and Overall subsets. The headline finding is immediate: Qwen3-VL-4B, a model 18× smaller than the original backbone, *surpasses* the Qwen2-VL-72B baseline on every reported metric except Unique@25, where it trails by a marginal 1.54 points. At Acc@25 and Acc@50, the 4B model achieves 47.23% and 42.84% respectively, outperforming the 72B baseline by +3.13 and +3.44 points. The 2B model, at 36× smaller, achieves near-parity with the 72B baseline at Acc@25 (42.41% vs. 44.1%) and matches it almost exactly at Acc@50 (39.15% vs. 39.4%), making it a viable option for the most memory-constrained edge deployments.

Table 1: Nr3D results. **Bold** = best per metric.

Metric	72B	4B *(Ours)*	2B *(Ours)*
Unique@25	75.7	74.16	73.03
Unique@50	68.9	67.98	**68.54**
Multiple@25	34.0	**38.14**	32.07
Multiple@50	30.0	**34.35**	29.22
Acc@25	44.1	**47.23**	42.41
Acc@50	39.4	**42.84**	39.15

Accuracy (%) at stated IoU. Baseline uses Qwen2-VL; ours use Qwen3-VL.

Unique Subset. The Unique split isolates pure object recognition: only one instance of the target class exists in the scene, so no relational reasoning is required. Performance here is remarkably stable across all model sizes. The 72B baseline achieves 75.7% at Unique@25; our 4B and 2B models trail by only 1.54 and 2.67 points respectively. Strikingly, the 2B model *exceeds* the 72B baseline at Unique@50 (68.54% vs. 68.9%), a gap well within noise. This stability confirms that SeeGround's rendered visual prompts convey sufficient object identity information for even a 2B model to reliably localize unambiguous targets.

Multiple Subset. The Multiple split is where the hypothesis predicted the sharpest degradation, and where our results are most surprising. This split requires the model to disambiguate a target from multiple distractors of the same class, demanding genuine spatial and relational reasoning. Rather than degrading, Qwen3-VL-4B *outperforms* the 72B baseline by +4.14 points at Multiple@25 and +4.35 points at Multiple@50, the largest absolute gains in the entire evaluation. The 2B model also remains competitive, trailing the 72B baseline by only 1.93 points at Multiple@25. This is the opposite of what a pure capacity argument would predict, and points directly to the role of Interleaved-MRoPE's finer spatial token resolution in distinguishing between spatially proximate objects with overlapping visual prompts.

Overall Accuracy. Aggregating across subsets, the gains on Multiple more than compensate for the marginal losses on Unique. Qwen3-VL-4B achieves 47.23% at Acc@25 and 42.84% at Acc@50, surpassing the 72B baseline by +3.13 and +3.44 points respectively. The 2B model achieves 42.41% at Acc@25, just 1.69 points below the 72B baseline, and 39.15% at Acc@50, essentially matching the baseline at a 36× reduction in parameter count. Both models occupy a strongly Pareto-optimal position: higher or comparable accuracy at a fraction of the memory and power cost.

4.2 Backbone Scaling Analysis

Table 2: Backbone scaling on Nr3D. VRAM at `bf16`. †from [1]. **Bold** = best.

Backbone	Params	VRAM	Acc.
Prior work [1]			
Qwen2-VL-7B†	7B	14 GB	33.3
InternVL2-8B†	8B	16 GB	34.3
InternVL2-26B†	26B	52 GB	38.0
Qwen2-VL-72B†	72B	144 GB	46.1
Ours (Qwen3-VL)			
Qwen3-VL-2B	2B	4 GB	42.41
Qwen3-VL-4B	4B	8 GB	**47.23**

Overall accuracy (%) on Nr3D at IoU@0.25.

Table 2 places our results within the full backbone scaling curve, combining the four data points from the SeeGround paper [1] with our two new evaluations. Within the Qwen2-VL and InternVL2 families, accuracy increases monotonically with scale — from 33.3% at 7B to 46.1% at 72B — consistent with standard scaling behaviour. Our Qwen3-VL-4B breaks this trend entirely, surpassing the 72B ceiling at less than a tenth of the parameter count.

This cross-generation jump, from 46.1% with Qwen2-VL-72B to 47.23% with Qwen3-VL-4B, indicates that architectural improvements in the Qwen3 generation contribute meaningfully to grounding accuracy, independent of scale. In particular, the jump from Qwen2-VL-7B (33.3%) to Qwen3-VL-2B (42.41%) is a gain of +9.11 points at a *smaller* parameter count, a result that cannot be explained by scale alone. Disentangling generation effects from scale effects would require a Qwen3-VL-72B or Qwen2-VL-4B comparison, which we identify as the most important direction for future work.

4.3 Qualitative Analysis

Figure 2 presents eight representative examples in which Qwen3-VL-4B correctly localizes the target object while Qwen3-VL-2B fails. A consistent pattern emerges across these cases: the 4B model succeeds on queries that require integrating spatial context from multiple objects simultaneously — identifying the correct desk among several by its relationship to a specific monitor, localizing a cabinet by its position relative to a couch against a wall, or selecting the correct ottoman by reasoning about the sofa behind it. The 2B model tends to select a plausible but incorrect object that satisfies part of the description without fully resolving the relational constraint, confirming the Multiple subset gap observed quantitatively. The IoU scores in these examples are stark: the 4B model achieves IoU values of 0.940–1.000 across all eight cases, while the 2B model scores 0.000 on seven of the eight, indicating complete localization failure rather than marginal misalignment.

5 Discussion

The hypothesis was partially contradicted. We predicted degradation on Hard and View-Dependent queries for smaller models. The Multiple subset — the closest available proxy — directly contradicts this: Qwen3-VL-4B *improves* upon the 72B baseline by +4.14 points at Multiple@25, the largest absolute gain in the evaluation, precisely where we expected the sharpest decline. The 2B model also remains competitive, trailing by under 2 points despite being 36× smaller.

Architecture versus scale. Two factors plausibly explain this. First, Qwen3-VL's Interleaved-MRoPE distributes spatial positional signals more finely across attention heads, improving pixel-level ID reading in SeeGround's rendered prompts — an advantage that is disproportionately beneficial on the Multiple subset where nearby objects carry overlapping labels. Second, SeeGround's pipeline pre-structures the task substantially: the query is already parsed, the scene rendered from an anchor-aligned viewpoint, and objects labeled in structured text. This scaffolding may reduce reasoning demand to a level a well-architected 4B model can satisfy without 72B-scale capacity. Separating these effects would require a Qwen3-VL-72B or Qwen2-VL-4B evaluation, which we identify as the most important direction for future work.

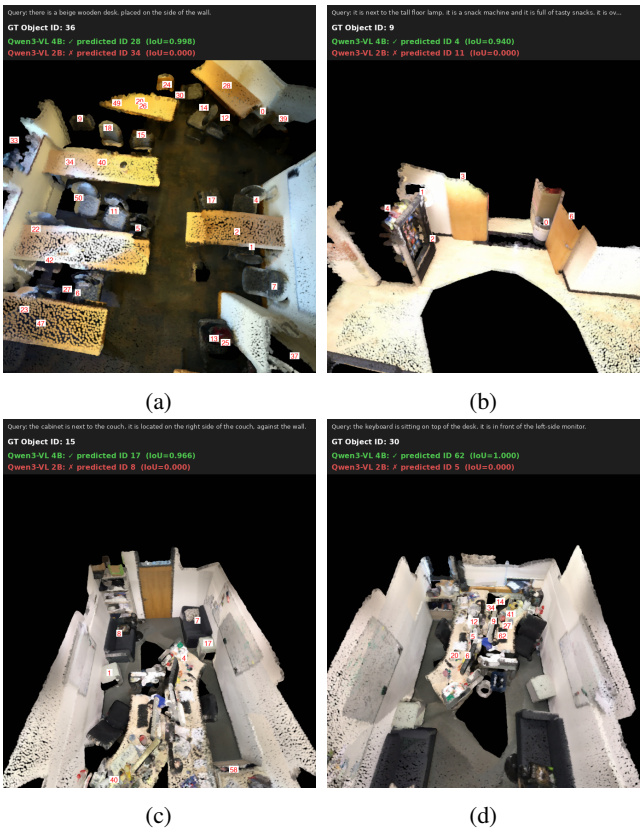


Figure 2: Qualitative grounding examples where Qwen3-VL-4B succeeds and Qwen3-VL-2B fails on Nr3D. Object IDs are overlaid as SeeGround visual prompts.

Implications for edge deployment. Qwen3-VL-4B occupies ∼8 GB at `bfloat16` and achieves state-of-the-art zero-shot 3DVG accuracy on a single consumer GPU. The 2B model at ∼4 GB fits within a Jetson AGX Orin without quantization. Both eliminate the multi-GPU infrastructure that made the original SeeGround deployment impractical. The accuracy-efficiency frontier has shifted: there is no longer a meaningful tradeoff between grounding quality and edge deployability at the 4B scale.

6 Conclusion and Future Work

This project demonstrates that the Qwen2-VL-72B reasoning backbone in the SeeGround 3D visual grounding framework can be replaced by Qwen3-VL models at the 2B and 4B parameter scales without sacrificing — and in the case of the 4B model, while surpassing — state-of-the-art zero-shot grounding accuracy. Qwen3-VL-4B achieves 47.23% overall accuracy on Nr3D at Acc@25, outperforming the 72B baseline by +3.13 points, while fitting within the memory envelope of a single edge GPU. The 2B model matches the 72B baseline at Acc@50 at 36× fewer parameters. These results establish that high-accuracy zero-shot 3D visual grounding is achievable today on hardware-constrained platforms, opening the door to deployment on autonomous vehicles, mobile robots, and drones without cloud connectivity or multi-GPU infrastructure.

Future Work. Several directions follow naturally from this study:

- **Disentangling scale from architecture.** Evaluating Qwen3-VL-72B or Qwen2-VL-4B would isolate whether performance gains stem from Interleaved-MRoPE or scale reduction.
- **Quantization.** INT4/INT8 quantization would shrink the 4B model to 2–4 GB, enabling deployment on highly constrained devices like the Jetson Orin NX with headroom for the full perception stack.
- **Latency and ScanRefer.** End-to-end latency profiling on target hardware is needed to confirm real-time viability, and a ScanRefer evaluation would test whether Nr3D gains transfer to a dataset where detection accuracy is also a factor.

References

[1] R. Li, S. Li, L. Kong, X. Yang, and J. Liang, "SeeGround: See and Ground for Zero-Shot Open-Vocabulary 3D Visual Grounding," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2025. https://arxiv.org/abs/2412.04383

[2] S. Bai et al., "Qwen3-VL Technical Report," *arXiv preprint arXiv:2511.21631*, 2025. https://arxiv.org/abs/2511.21631

[3] P. Wang et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv preprint arXiv:2409.12191*, 2024. https://arxiv.org/abs/2409.12191

[4] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes," in *Proc. European Conf. Computer Vision (ECCV)*, 2020.

[5] W. Kwon et al., "Efficient Memory Management for Large Language Model Serving with PagedAttention," in *Proc. ACM Symp. Operating Systems Principles (SOSP)*, 2023. https://arxiv.org/abs/2309.06180

[6] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "LangSplat: 3D Language Gaussian Splatting," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3D: Mask Transformer for 3D Semantic Instance Segmentation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2023. https://arxiv.org/abs/2210.03105

[8] N. Ravi et al., "Accelerating 3D Deep Learning with PyTorch3D," *arXiv preprint arXiv:2007.08501*, 2020. https://arxiv.org/abs/2007.08501