Grounded 3D-Aware Spatial Vision-Language Modeling

Supplementary Material

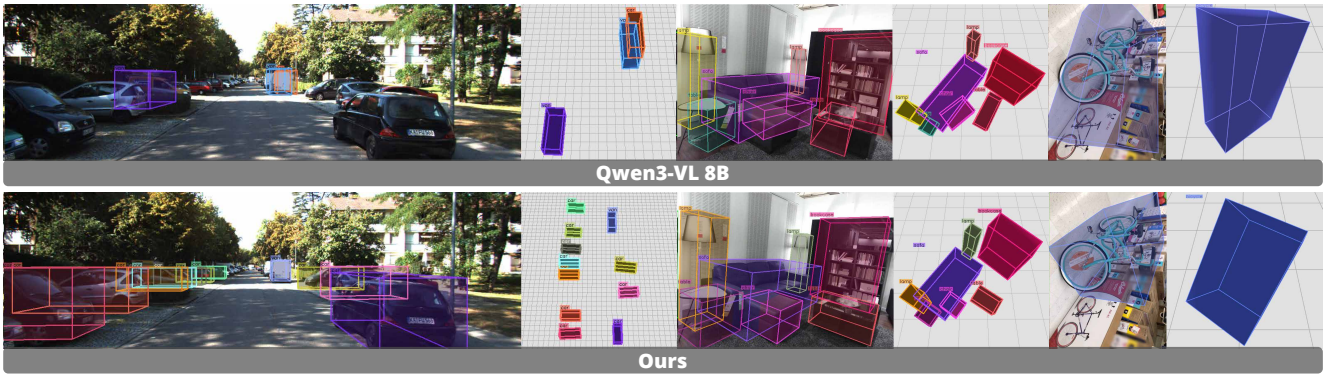

Figure 6. Qualitative comparison on 3D object detection between our model and Qwen3-VL 8B [45]. Our model produces more accurate 3D bounding boxes with fewer missed objects, demonstrating stronger spatial grounding and detection reliability.

Table of Contents

1. More Results on 3D Detection

We show a qualitative comparison on 3D object detection between GR3D and Qwen3-VL-8B [45]. As shown in Fig. 6, when multiple objects are present, GR3D produces clearly better results due to our detect-then-lift technique. For indoor scenes, GR3D also predicts 3D boxes with more accurate orientations compared to Qwen3-VL-8B.

2. More Results on 2D Grounding

Our approach decomposes 3D detection into two steps: first grounding the target in 2D, then predicting its 3D properties based on the grounded region. Because accurate 2D grounding is essential for the first step, we evaluate our model on two grounding benchmarks. We first report results on RefSpatial [23], a benchmark designed for spatial referring that includes queries about vacant regions, spatial relations (e.g., "left of", "between"), and fine-grained spatial logic. As shown in Table 6, our model achieves strong spatial referring performance and outperforms several baselines, including RoboRefer [23], demonstrating its ability to reason about complex spatial relations in 2D. We further evaluate on the widely used RefCOCO, RefCOCO+, and RefCOCOg datasets [98, 99] to measure general referring capability. These benchmarks contain diverse referring expressions involving object names, attributes, and relational descriptions. Results in Table 8 show that GR3D performs comparably to vision-specialized models and is on par with top VLMs such as InternVL-3.5 [100] or Qwen2.5-VL [64], confirming that our strong 2D grounding ability generalizes well to both spatial and standard referring benchmarks.

Method	LOCATION	PLACEMENT	UNSEEN
Gemini-2.5-Pro [101]	46.9	24.2	27.1
SpaceLLaVA-13B [65]	5.8	4.3	4.0
RoboPoint-13B [25]	22.8	9.2	8.4
Molmo-7B [102]	21.9	12.8	12.2
Molmo-72B [102]	45.7	14.7	21.2
RoboBrain-2.0-7B [24]	36.0	29.0	32.5
RoboRefer-8B [23]	52.0	53.0	37.7
GR3D-8B	**63.0**	**50.0**	**41.5**

Table 6. Performance comparison on RefSpatial [23].

3. More Results on Multi-View Understanding

3.1. Multi-View Extension of Our Framework

Our framework naturally extends from single-view to multi-view settings through a unified spatial embedding design similar to SR-3D [22]. All image tokens, regardless of the view they originate from, are mapped into the same spatial feature space using depth-based and pixel-based positional

Methods	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	Avg.
	Quantitative				Qualitative				
Random	-	-	-	-	25.0	36.1	28.3	25.0	-
Human Level†	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100	79.2
Proprietary Models (API)									
GPT-4o [58]	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0
Gemini-1.5 Flash [103]	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8	42.1
Gemini-1.5 Pro [103]	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.3
Open-source Models									
InternVL2-2B [104]	24.9	22.0	35.0	33.8	44.2				
InternVL2-8B [104]	31.3	29.0	48.9	44.2	38.0	33.4	28.9	46.4	37.5
InternVL2-40B [104]	41.3	26.2	48.2	27.5	47.6	32.7	27.8	44.7	37.0
LongVILA-8B [105]	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5	21.6
VILA-1.5-8B [2]	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8	28.9
VILA-1.5-40B [2]	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2
LongVA-7B [106]	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
LLaVA-Video-7B [107]	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6	35.6
LLaVA-Video-72B [107]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
LLaVA-OneVision-7B [108]	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4	32.4
LLaVA-OneVision-72B [108]	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
SR-3D-8B	54.9	53.8	74.5	65.1	63.5	81.8	33.5	75.9	62.9
GR3D-8B	**69.6**	**55.2**	**76.8**	**65.6**	**70.5**	**86.3**	**35.5**	**81.2**	**67.6**

Table 7. We finetune our model on multi-view datasets [56, 109] following SR-3D [22], and then evaluate multi-view global spatial scene understanding on VSI-Bench [75]. Methods marked with † are evaluated on the `Tiny` subset. GR3D outperforms all state-of-the-art baselines, demonstrating strong spatial recognition capability.

Model Name	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Vision Specialists								
Grounding-DINO-L [89]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
UNINEXT-H [110]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
ONE-PEACE [111]	92.6	94.2	89.3	88.8	92.2	83.2	89.2	89.3
Vision Language Models								
InternVL3-1B [112]	85.8	90.1	81.7	76.6	84.1	69.2	82.8	82.6
InternVL3.5-1B [100]	85.4	89.7	80.2	77.7	85.5	69.5	81.9	81.6
InternVL3-2B [112]	89.8	92.6	86.4	84.0	89.2	76.5	87.6	87.2
InternVL3.5-2B [100]	88.7	91.6	84.8	82.7	88.4	76.6	85.6	85.5
Qwen2.5-VL-3B [64]	89.1	91.7	84.0	82.4	88.0	74.1	85.2	85.7
Shikra-7B [113]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
CogVLM-G [60]	92.8	94.8	89.0	88.7	92.9	83.4	89.8	90.8
Qwen2-VL-7B [114]	91.7	93.6	87.3	85.8	90.5	79.5	87.3	87.8
Qwen2.5-VL-7B [64]	90.0	92.5	85.4	84.2	89.1	76.9	87.2	87.2
TextHawk2 [115]	91.9	93.0	87.6	86.2	90.0	80.4	88.2	88.1
InternVL3.5-8B [100]	92.4	94.7	88.7	87.9	92.4	82.4	89.6	89.4
GR3D-8B	91.8	94.5	88.8	87.5	91.4	81.0	89.5	89.7

Table 8. We evaluate GR3D 's 2D grounding on RefCOCO, RefCOCO+, and RefCOCOg [98, 99]. Baseline numbers are taken from [100, 114]. GR3D achieves grounding accuracy comparable to vision specialists [89, 110, 111] models and performs on par with top VLMs such as InternVL3.5 [100].

cues. This allows the model to maintain consistent geometric relationships across views without requiring explicit point cloud reconstruction or global world coordinates.

For multi-view inputs, the first view is processed exactly as in the single-view case and is treated as the reference frame. Unlike SR-3D, which assumes a global world coordinate system and expresses all views in that space, our approach keeps everything in the coordinate frame of the first camera. Each additional view is transformed into this reference coordinate system using its intrinsics and extrinsics, so all depth-derived 3D locations and pixel-coordinate cues are expressed in the same spatial frame. After this transformation, tokens from different views that observe the same physical point occupy nearby positions in the unified embedding space. This allows the model to reason about 3D structure, occlusion, and cross-view consistency directly from the spatial tokens.

3.2. Results on VSI-Bench

To validate this design, we finetune our stage-1 model on multi-view datasets [56, 109] following SR-3D [22], and then evaluate multi-view global spatial scene understanding on VSI-Bench [75]. As shown in Table 7, GR3D achieves strong performance with an average score of 67.6 and surpasses all state-of-the-art baselines, showing that our method can effectively handle multi-view inputs.

3.3. Results on ScanRefer, ScanQA, MMSI, SPAR

To further evaluate the 3D grounding capabilities of GR3D on multi-view datasets, we conduct studies leveraging ScanRefer [116] benchmark. However, ScanRefer assumes ac-

cess to a pre-aligned world coordinate space, which is not directly compatible with the settings of Qwen3-VL [45] and ours. We therefore adapt ScanRefer into a frame/2D box grounding followed by 3D detection in the camera coordinate space, and compare against Qwen3-VL under the same input conditions. Our method outperforms Qwen3-VL-8B and is competitive with methods that use pre-aligned 3D input. We also report results on ScanQA [109], MMSI-Bench [117] and SPAR-Bench [118], showing consistent improvements.

	ScanRefer		ScanQA		
	@0.25	@0.5	B4	C	EM
SPAR	48.8	43.1	15.3	90.7	27.7
3D-LLaVA	51.2	40.6	17.1	92.6	-
Video-3D LLM	58.1	51.7	16.2	102.1	30.1
Qwen3-VL-8B	37.7	33.2	-	-	-
GR3D-8B	52.0	46.1	18.1	105.1	29.2

	MMSI	SPAR
GPT-4o	30.3	38.1
InternVL2.5-8B	28.7	36.3
Qwen2.5-VL-7B	25.9	33.1
Qwen3-VL-8B	31.1	39.8
NVILA-8B	28.1	34.1
SR-3D-8B	25.8	32.1
GR3D-8B	29.2	43.7

Table 9. Performance comparison on ScanRefer [116], ScanQA [109], MMSI-Bench [117], and SPAR-Bench [118].

4. Implementation Details

4.1. Model Architecture

Following NVILA-Lite, we use SigLIP as the vision encoder with an input resolution of 448 and a patch size of 14, paired with a Qwen-2-7B [114] LLM backbone. For training the stage-1 model, we follow SR-3D and enable dynamic tiling with up to 12 tiles per image. We also adopt SR-3D's dynamic tiling region extractor, which provides a larger effective receptive field for regions and improves the model's ability to handle small objects. During the first stage, the vision encoder is frozen and only the remaining modules are trained. For the second CoT detection stage, the LLM is fine-tuned to learn the reasoning structure and the autoregressive 3D prediction format.

4.2. Training Hyper-parameters

Both stages use the same optimization schedule: a warmup ratio of 0.03 and a cosine learning-rate scheduler. In the stage-1 stage, we train all non-visual modules with AdamW and a base learning rate of 5×10^{-5}, while keeping the SigLIP encoder frozen. The second CoT detection stage fine-tunes only the Qwen-2-7B LLM with a smaller learning rate of 1.5×10^{-5} to stabilize chain-of-thoughts text generation. Training the stage-1 model takes approximately 4 days on 8 nodes of A100 servers, while the second stage takes about 4 hours on the same compute setup.

4.3. Data Composition

The data composition for both training stages is summarized in Table 10. Most of our training data follow NVILA's data recipe, though we use only a subset due to computational constraints. Part of the spatial data is inherited from SR-3D, while many of the 2D grounding datasets are newly introduced to the model and trained for the first time on our weights. For the 3D detection data used in stage 1, we follow DetAny3D's filtering rules on Omni3D to select high-quality training objects, and convert each scene into multi-turn conversations with up to 10 rounds. For the CoT detection data used in stage 2, we construct multi-object reasoning sequences by selecting up to 20 objects for each target.

Stage-1 Data	
Hybrid	ShareGPT4V-SFT, Molmo, The Cauldron, Cambrian, LLaVA-OneVision
Captioning	MSR-VTT, Image Paragraph Captioning, ShareGPT4V-100K
Reasoning	CLEVR, NLVR, VisualMRC
Document	DocVQA, UniChart-SFT, ChartQA
OCR	TextCaps, OCRVQA, ST-VQA, POIE, SORIE, SynthDoG-en, TextOCR-GPT4V, ArxivQA, LLaVAR
General VQA	ScienceQA, VQAv2, ViQuAE, Visual Dialog, GQA, Geo170K, LRV-Instruction, RefCOCO, GeoQA, OK-VQA, TabMVP, EstVQA
Diagram & Dialogue	DVQA, AI2D, Shikra, UniMM-Chat
Instruction	LRV-Instruction, SVIT, MMC-Instruction, MM-Instruction
Text-only	FLAN-1M, MathInstruct, Dolly, GSM8K-ScRel-SFT
Knowledge	WordART, WIT, STEM-QA
Medical	PathVQA, Slake, MedVQA
Region	RegionGPT
Spatial & 2D Grounding	RefCOCO, MGrounding, Molmo, Groma, Spatial-RGPT, RefSpatial, SAT, EmbSpatial, DepthLM
Detection	Omni3D, EmbodiedScan
Stage-2 Data	
Detection	Omni3D-CoT
Spatial	RefSpatial-CoT, MMG-CoT, EmbSpatial-CoT, Vis-CoT

Table 10. Data recipe for training GR3D.

5. More Related Work

A related line of research, recently formalized as *Thinking with Images* [90], focuses on improving complex VLM reasoning by decomposing problems into explicit, intermediate steps, treating vision as a dynamic workspace. Many such methods act as "commanders" orchestrating external visual tools [91, 92] or as "visual programmers" that generate code for custom analysis and edits [93–95]. Others generate intermediate visual representations to guide reasoning, often called Visual Chain of Thought (V-CoT) [119]. These V-CoT methods may interleave text with explicit visual groundings [120], sketch visual artifacts [96], generate subgoal images for robotics [121], or perform planning entirely

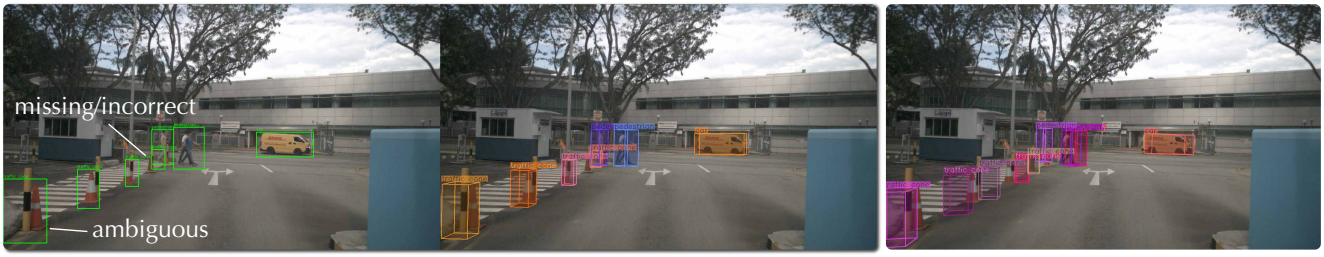

Figure 7. A failure case analysis of GR3D compared to Qwen3-VL-8B.

through visual state sequences [122]. While these methods enhance transparency and performance on complex tasks, they still focus on 2D image space, rely on coarse region-selection cues or external tools, and rarely integrate these reasoning steps with a 3D spatial framework. In contrast, our GR3D framework bypasses the need for an explicit, step-by-step visual thought process. It achieves a more seamless integration by performing implicit 2D grounding and unified 3D reasoning natively within the VLM's generative flow.

6. Discussions

6.1. Standard VLM without PE

Our method can be applied to standard VLMs, but 3D priors further improve performance. Using positional embeddings, Omni3D mAP (averaged over 6 datasets) improves from 22.9 to 25.4 compared to a standard VLM without positional embeddings, showing their benefit as a simple and effective 3D prior.

6.2. Hallucinations

We do not observe frequent hallucinated 2D boxes. The main failures are missing or ambiguous 2D grounding, which leads to incorrect predictions. We show an example above and compare them with Qwen3-VL-8B [45].

6.3. Effect of Intrinsic Estimation Errors

The effect of intrinsic normalization is modest. Since the normalization only determines the resolution size, it does not require highly accurate intrinsic estimates. In practice, off-the-shelf intrinsic estimators are sufficiently accurate: using GeoCalib for focal length prediction on Omni3D results in only a 1.2 mAP drop (averaged over 6 datasets).

6.4. CoT Data Robustness

We conduct an ablation study on the impact of data quality by training with a noisier corpus, which leads to performance drops from 74.2 to 62.8 on MM-GCoT's grounding accuracy. Human evaluation on 200 randomly sampled instances from the filtered corpus shows that 95.5% of the generated bounding boxes are accurate.

6.5. Latency Analysis

We implement multimodal prefix caching to ensure that the inference pipeline runs at a speed comparable to standard autoregressive generation. For Region Insertion, the process only extracts relevant areas from already encoded image features and passes them through a lightweight MLP projector, without re-encoding the image. We provide a latency analysis that compares our method with other baselines, tested on the same input image using a single A100 GPU. Our model is fastest among VLMs due to a more efficient dynamic tiling–based vision encoder (vs. AnyRes). The additional cost per inserted region is only 0.01 s, which is a small fraction of the total 2.7 s inference time.

	DetAny3D	VST-7B	Qwen3-VL-8B	**GR3D-8B**
Latency (s)	0.98	2.76	3.23	2.72

6.6. Limitations

Our approach has two main limitations. First, the inference speed is slower compared to vision specialists. This is mainly due to the use of a large language model backbone, our two-stage "2D grounding first" pipeline, and the fact that 3D bounding boxes are generated autoregressively as text tokens, all of which introduce additional latency. Second, current 3D detection datasets are still limited. Popular datasets such as Omni3D cover only a narrow set of environments, camera configurations, and object categories, which restricts the diversity and scale of 3D supervision our model can learn from. As a result, further progress will benefit from larger and more diverse 3D datasets with broader scene coverage and richer object annotations.