Grounded 3D-Aware Spatial Vision-Language Modeling

An-Chieh Cheng[1,3*] Yang Fu[1] Yatai Ji[3] Ligeng Zhu[3] Guanqi Zhan[3] Zhuoyang Zhang[2,3]
Zhaojing Yang[1] Song Han[2,3] Yao Lu[3] Pavlo Molchanov[3] Vidya Nariyambut Murali[3] Jan Kautz[3]
Xiaolong Wang[1] Hongxu Yin[3] Sifei Liu[3]
[1]UCSD [2]MIT [3]NVIDIA

https://www.anjiecheng.me/gr3d

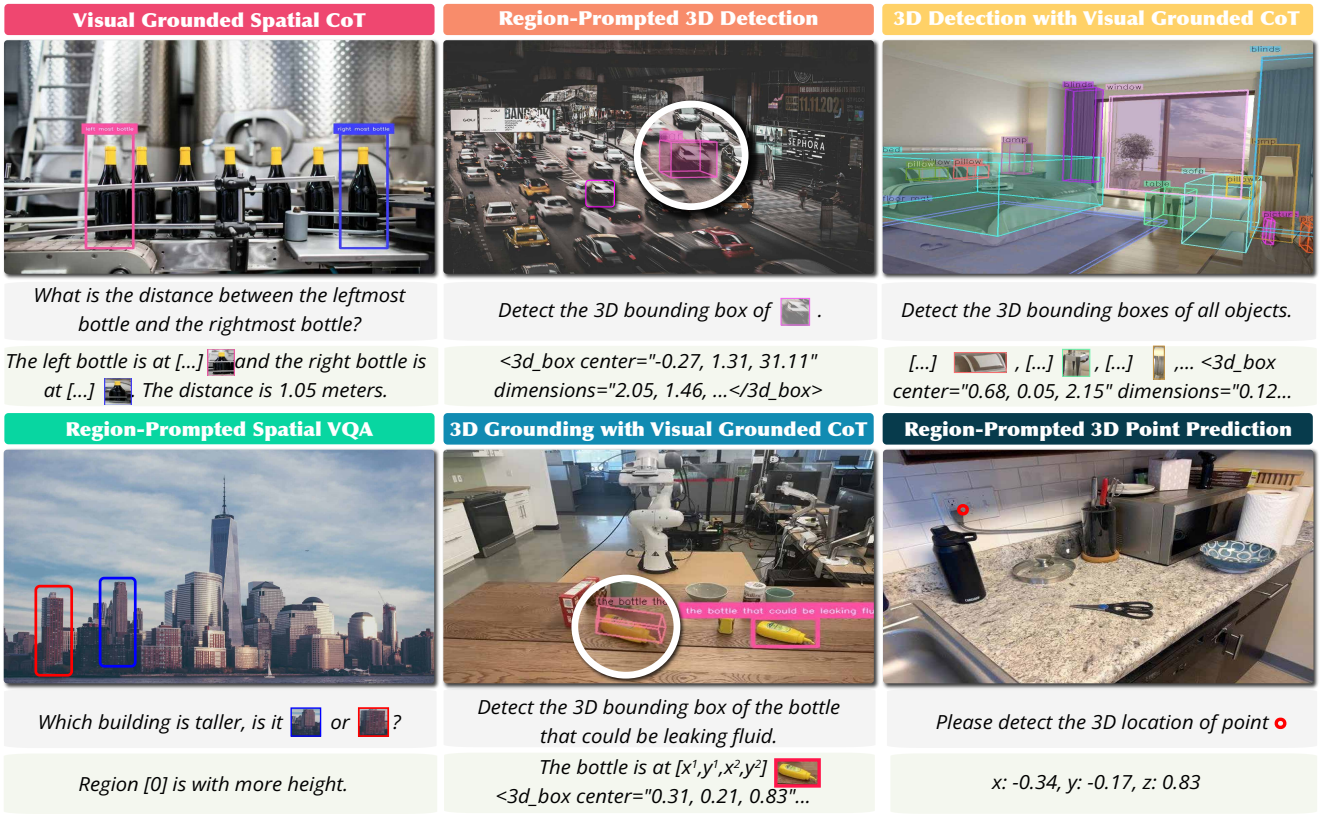

Figure 1. **Overview.** GR3D bridges 2D pixel-space and 3D metric-space by integrating multiple grounding capabilities into visual chain-of-thoughts, enabling complex spatial understanding through grounded 2D perception followed by 3D inference.

Abstract

We present GR3D, a spatial vision language model equipped with three complementary grounding capabilities—explicit 2D grounding, implicit 2D grounding, and monocular 3D grounding—within a single framework. GR3D introduces an implicit grounding mechanism that identifies entity mentions during generation and inserts the corresponding region tokens into the text stream, allowing the model to reference visual evidence on the fly when producing spatial chain-of-thought responses. In parallel, a region-prompted monocular 3D grounding design predicts 3D bounding boxes in the camera view from grounded region queries, supported by intrinsic-aware normalization and dense geometric supervision. Together, these grounding capabilities enable GR3D to decompose complex spatial understanding problems into grounded 2D perception followed by 3D inference. GR3D achieves consistent improvements across grounded and non-grounded spatial benchmarks, demonstrating grounding as an effective inductive bias for strengthening spatial understanding in VLMs. These grounding capabilities collectively enhance general spatial understanding beyond the grounding task itself.

*Work done during an internship at NVIDIA.

1. Introduction

Vision–language models (VLMs) have rapidly evolved into general-purpose perception–language systems [**? ? ? ? ? ? ? ?**], capable of understanding scenes, following open-ended instructions, and supporting diverse multimodal tasks. As these models begin to serve as the core of embodied agents that must act, manipulate, and navigate in the physical world [**? ? ? ? ? ? ? ? ?**], their spatial competence becomes crucial. Embodied intelligence requires models not only to recognize what is present, but also to understand where objects are and how they are arranged in space—capabilities essential for grounding language into actions such as where to reach, step, or orient [**? ? ?**]. Without reliable spatial grounding, the link between high-level instructions and physical interaction remains brittle, limiting the scalability of VLMs toward real-world embodied perception and control.

Rapid progress in spatial VLMs has substantially advanced 2D spatial understanding and even 3D perception [**? ? ? ? ? ? ? ? ?**]. Yet grounding—the ability to reliably associate linguistic mentions with concrete visual regions and connect 2D evidence with 3D structure—remains limited. Two challenges, in particular, are under-addressed. (i) Implicit 2D grounding is scarce: most systems support explicit "point to X" grounding but lack mechanisms or data for automatically detecting entities mentioned in free-form text and integrating their corresponding visual evidence during generation. Constructing such supervision is difficult, as it requires aligning textual mentions to latent visual regions and interleaving region information into the language stream. (ii) Monocular 3D grounding is inherently ill-posed: from a single view, object scale, depth, and intrinsics are entangled, and 3D prediction requires first identifying which instance the text refers to before estimating its 3D extent and pose. Existing approaches often bypass this intermediate localization step [**?**], rely on multi-view supervision [**?**], or are limited by the scarcity of 3D box annotations [**?**].

To address these limitations, we introduce **(GR3D)**, a spatial VLM that integrates grounding as a core mechanism for learning spatial representations. GR3D jointly supports three complementary grounding capabilities within a unified architecture: *explicit 2D grounding*, which predicts object regions through the language head in a structured textual format; *implicit 2D grounding*, which links linguistic mentions to visual evidence through dynamic region insertion; and *monocular 3D grounding*, which extends region understanding into 3D by predicting bounding boxes and camera intrinsics under dense geometric supervision. Together, these mechanisms establish a fine-grained alignment between language, image regions, and geometry, enabling consistent 2D and 3D spatial reasoning.

While explicit 2D grounding predicts the location of queried objects, it cannot handle free-form reasoning where spatial cues are implicit. Real-world spatial queries—*e.g.*, describing relations, distances, or navigation targets—require first recognizing and localizing the entities mentioned before reasoning about the query itself. GR3D bridges this gap with an implicit 2D grounding mechanism that performs *streaming region insertion*: as the model generates responses, it dynamically predicts the visual region corresponding to each mentioned entity, encodes the region into a token, and injects it directly into the ongoing language stream. This enables reasoning to evolve directly over grounded visual evidence, yielding coherent spatial predictions without any separate detection phase.

Inferring 3D structure from a single view introduces both linguistic and geometric ambiguities, such as determining which instance a description refers to and estimating its depth, scale, and pose without multi-view cues. GR3D addresses these challenges through a region-prompted 3D grounding formulation: each grounded 2D region is treated as a query for 3D inference, supported by intrinsic-aware normalization and dense geometric supervision derived from depth estimation. This design aligns semantic localization and geometric prediction within a consistent camera-view framework, enabling the model to infer coherent 3D structure directly from grounded 2D evidence and to generalize across diverse scenes and viewpoints. Crucially, by receiving region tokens produced by implicit 2D grounding, the 3D predictor naturally plugs into CoT-driven reasoning—allowing the model to first resolve "which object" via grounded language generation and then infer "what 3D structure" for that object. This decomposition makes monocular 3D grounding applicable to both instance-level referring tasks and open-set category-level 3D detection.

Integrating explicit 2D grounding, implicit 2D grounding, and monocular 3D grounding positions GR3D as a flexible spatial understanding framework spanning 2D/3D and single-/multi-view settings. Through this grounding-centered formulation, the model learns to localize, reference, and reason over spatial structure in a unified manner. Implicit grounding enhances CoT accuracy and spatial consistency on CVBench [**?**], ERQA [**?**] and SAT [**?**], while region-prompted 3D grounding with dense point supervision achieves state-of-the-art performance on Omni3D. Moreover, we observe key insights: (i) grounding improves general spatial understanding even without explicit localization; (ii) dense geometric supervision provides scalable structure cues; (iii) combining implicit grounding with region-prompted 3D inference unlocks a versatile decomposition pipeline that supports referring-instance 3D grounding, category-level 3D detection, and multi-object scene grounding. Together, these results show that embedding grounding within the model architecture strengthens

both spatial perception and grounded reasoning.

2. Method

GR3D is designed to address two major limitations of current Spatial VLMs: the lack of an implicit grounding mechanism that allows models to automatically associate linguistic mentions with visual evidence during reasoning, and the difficulty of performing monocular 3D grounding from a single image with entangled depth and scale cues. To overcome these, we first construct a **foundational Spatial VLM** (Sec. 2.1) that provides geometry-aware features for both single- and multi-view inputs. Building on this foundation, we introduce **explicit** and **implicit 2D grounding** (Sec. 2.2) to link linguistic expressions with visual evidence, and extend them to **monocular 3D grounding** through region prompts, intrinsic normalization, and dense geometric supervision (Sec. 2.3). Finally, we describe our **data construction pipeline** (Sec. 2.4) that generates large-scale implicit grounding annotations and balanced 2D–3D supervision to facilitate training.

2.1. Foundational Spatial VLM

Objective. We follow the design principle of SR-3D [**?**] to construct a foundational Spatial VLM based on the NVILA-8B-Lite [**?**] architecture. This model provides a unified spatial representation that supports both single- and multi-view spatial understanding, serving as the base for subsequent grounding modules. At this stage, no grounding capability is included; the focus is on building a geometry-aware representation layer compatible with language reasoning.

Single-view Setup. The base NVILA encoder extracts dense visual tokens from an RGB image for single-view inputs. To make these tokens spatially aware, we augment them with 2D positional embeddings derived from their pixel coordinates and relative depth cues. Each visual token therefore carries both appearance and geometric context. Unlike language tokens, which are processed sequentially, these enriched visual tokens retain their spatial arrangement within the image grid. They are passed through the multimodal projector before being fed to the language model. This projection preserves spatial locality while remaining compatible with NVILA's multimodal fusion pipeline.

In addition, we preserve the region-prompt design used in SR-3D: specific image regions can be encoded as individual query tokens by pooling features within a given bounding box. This structure allows downstream modules to reference localized spatial content directly, maintaining full alignment with the NVILA token structure and positional hierarchy. Overall, the single-view formulation provides a strong spatially-structured feature space for both region-level interaction and text-aligned representation.

Multi-view Setup. Our framework naturally extends from single-view to multi-view inputs by embedding all image tokens with depth- and pixel-based positional cues in a unified spatial feature space. The first view is processed exactly as in the single-view case, and all subsequent views are transformed into the first-frame coordinate system. Please see the supplementary materials for our multi-view results.

2.2. Grounding in the 2D Plane

Grounding on the 2D plane aims to teach the model to associate linguistic mentions with localized image evidence. We introduce both explicit and implicit forms of grounding, designed to strengthen the spatial reasoning capacity of the vision–language model.

2.2.1. Explicit 2D Grounding

For explicit 2D grounding, we adopt a simple and general formulation. Given a natural-language instruction, the model predicts 2D bounding boxes directly in HTML-style textual format (*e.g.*, <bbox>[x_1, y_1, x_2, y_2]</bbox>), using its standard language generation head without any additional detection branch. This unified design integrates grounding seamlessly into the vision–language interface, without introducing task-specific architectural components.

2.2.2. Implicit 2D Grounding

Consider a global spatial reasoning query such as: *"In the kitchen, how far is the second bottle on the shelf from the small brown teddy bear on top of the washing machine in the laundry room?"* Traditional spatial VLMs attempt to answer such questions directly from global image features, relying on large-scale question–answer pairs to memorize spatial relationships. However, this departs from how humans perceive scenes: we first identify where each mentioned object is before reasoning about their relations. Our implicit grounding mechanism explicitly introduces this intermediate step of *entity localization during generation*, aligning the model's behavior with human visual reasoning.

Streaming Region Insertion. Given an input instruction, the model generates its response in a chain-of-thought (CoT) fashion. When an entity (*e.g.*, "the second bottle on the shelf") is mentioned, the model first predicts its corresponding 2D bounding box coordinates, *e.g.*, $[x_1, y_1, x_2, y_2]$. Immediately after, the corresponding image region is encoded through the region encoder, and its embedding—a region token—is inserted directly into the text stream at that position. The generation then continues conditioned on both the textual and visual context. The same procedure repeats for subsequent entities, producing a temporally aligned reasoning trajectory that alternates between language and vision.

Training and Inference Paradigm. During training, the bounding box coordinates are directly predicted by the language head and optimized through teacher forcing, as they are treated as part of the textual output sequence. Once the

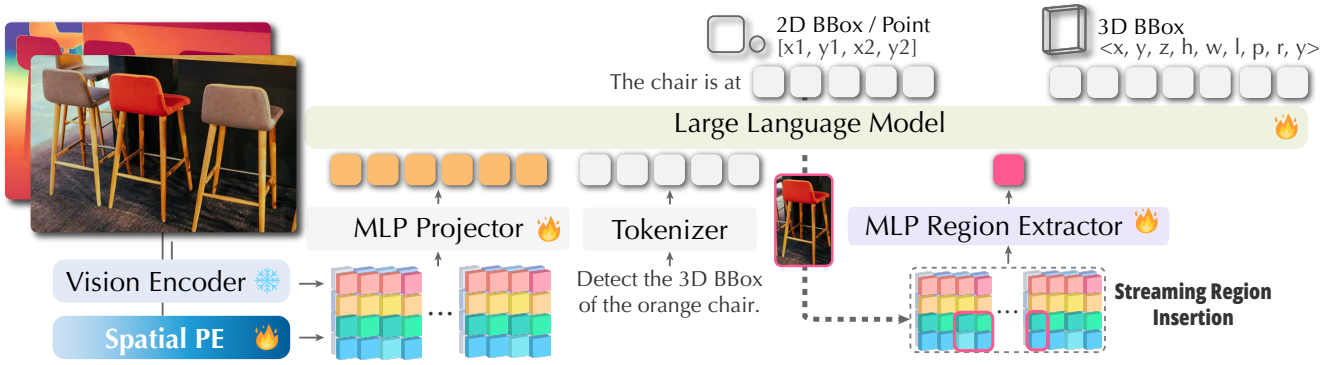

Figure 2. Method overview. GR3D builds on Region-VLMs by adding streaming region insertion for visual Chain-of-Thought reasoning. During CoT, the model repeatedly predicts a region, extracts its visual embedding, and reinserts a region token into the text sequence, enabling step-by-step spatial reasoning with dynamically refreshed visual cues.

coordinates are produced, the corresponding region token—derived from the ground-truth region—is inserted into the generation stream. This token is detached from the computation graph (i.e., no gradient flows through it) but serves as a strong conditional cue for subsequent token prediction. During inference, the process becomes fully autoregressive. The model first predicts coordinates, then encodes the predicted region to obtain its embedding, which is inserted back into the ongoing sequence before the next generation step. The subsequent reasoning, such as relational comparison or distance estimation, is thus conditioned on both the textual context and dynamically inserted region evidence.

Comparison and Interpretation. Our stream-based grounding can be viewed abstractly as analogous to a two-step process, *i.e.*, first grounding entities with a VLM, and then performing region-conditioned reasoning with a spatial VLM with a region encoder equipped. Unlike this staged formulation, our approach unifies both phases in a single generative stream. The model learns *when* and *what* to ground based on linguistic context, and its reasoning naturally unfolds on grounded evidence without explicit stage transitions. This results in a fluid, interpretable reasoning process that tightly couples perception and cognition while avoiding the discontinuities of discrete grounding modules.

2.3. Monocular 3D Grounding via Region Prompt

Monocular 3D grounding aims to enable single-view models to infer 3D structure from natural language and visual cues. This task faces two major challenges. First, linguistic ambiguity: textual references often under-specify which instance is being mentioned, requiring the model to implicitly identify the target entity before 3D reasoning. Second, geometric ambiguity: the coupling between object scale, depth, and camera intrinsics makes single-view estimation inherently uncertain. We address these through several components below that align semantic localization and geometric inference within a unified generative framework.

Region-prompt Formulation. Given a localized 2D region, the model treats this region as a spatial query for 3D reasoning. The region's visual features are pooled and encoded into a region token, which is fused into the text stream to guide 3D box prediction. Since the model already possesses implicit 2D grounding capability, this step focuses solely on extending that capacity from 2D to 3D—mapping a grounded region to its corresponding 3D representation. This formulation simplifies 3D grounding by conditioning inference on a given region, enabling the model to estimate position, scale, and orientation directly without performing explicit multi-step localization.

3D Box Representation. Each 3D bounding box is expressed in a unified, language-based format compatible with 2D HTML-style outputs, eliminating the need for task-specific heads. The box is parameterized by its center (x_c, y_c, z_c), size (w, h, l), and orientation $(\theta_p, \theta_r, \theta_y)$, where $(\theta_p, \theta_r, \theta_y)$ are *normalized* Euler angles (pitch/roll/yaw). To ensure consistency across datasets, we standardize orientations by selecting the rotation variant that minimizes the angular deviation between the local PCA axes of the region and the global coordinate axes (X, Y, Z)—that is, the variant closest to the identity basis rather than a mirrored alternative. This compact decomposition makes the representation transferable: the center term aligns naturally with depth-based supervision (see below), while the dimension and rotation terms capture view-invariant geometry. The format promotes stability, interpretability, and seamless integration into the generative language interface.

Intrinsic Normalization. To mitigate scale and depth ambiguity, we introduce an intrinsic-aware normalization strategy that rescales images according to focal length, yielding a consistent field of view across datasets. Concretely, given focal length f_x, we normalize the spatial scale by $W' = \frac{1000}{f_x} \cdot W$ and $H' = \frac{1000}{f_x} \cdot H$, aligning the apparent object size in the feature space and supporting robust 3D inference without explicitly regressing intrinsics.

Points and Direct Grounding Supervision. We supervise monocular 3D grounding with complementary signals be-

Method	SUN-RGBD [?]		ARKITSCENES [?]		OBJECTRON [?]		HYPERSIM [?]		KITTI [?]		NUSCENES [?]		AP_{3D} ↑
	AP_{15} ↑	mAP ↑	AP_{15} ↑	mAP ↑	AP_{15} ↑	mAP ↑	AP_{15} ↑	mAP ↑	AP_{15} ↑	mAP ↑	AP_{15} ↑	mAP ↑	
Vision Specialist Models													
ImVoxelNet [?]	-	-	-	-	-	-	-	-	-	-	-	-	9.4
SMOKE [?]	-	-	-	-	-	-	-	-	-	-	-	-	10.4
Cube R-CNN [?]	-	15.33	-	41.73	-	50.84	-	7.48	-	32.50	-	30.06	23.26
OVMono3D [?] *w/ Cube R-CNN*	-	15.20	-	41.60	-	58.87	-	7.75	-	25.45	-	24.33	22.98
DetAny3D [?] *w/ Cube R-CNN*	26.62	18.96	59.55	46.13	72.51	54.42	11.43	7.17	44.28	31.61	41.01	30.97	24.92
Vision Language Models													
Qwen3-VL-4B [?]	28.28	17.60	63.97	46.33	61.60	43.13	11.56	6.44	17.39	11.25	7.48	4.89	-
Qwen3-VL-8B [?]	28.28	17.77	62.32	45.23	61.63	43.59	11.62	7.23	5.23	3.32	11.52	7.56	-
GR3D-8B (Ours)	**43.49**	**31.64**	**67.49**	**52.52**	**71.68**	**54.32**	**16.42**	**10.87**	**22.18**	**14.75**	**22.98**	**16.59**	**25.40**

Table 1. Comparison on the Omni3D [?] benchmark between GR3D, vision specialists, and recent VLMs. We report AP_{15} and mAP for each dataset domain. GR3D outperforms all recent VLMs and vision specialists, especially on the indoor domain.

Method	AP_{2D}^{sun}	AP_{2D}^{ark}	AP_{2D}^{obj}	AP_{2D}^{hyp}	AP_{2D}^{kit}	AP_{2D}^{nus}
Cube R-CNN [?]	15.07	40.22	49.24	11.05	36.14	34.64
Qwen3-VL-8B [?]	8.06	22.44	30.06	3.08	1.54	2.56
GR3D-8B (Ours)	**38.86**	**46.17**	**51.66**	**28.53**	**20.49**	**22.16**

Table 2. 2D detection results on the Omni3D benchmark. We report the mean Average Precision (mAP) for each dataset domain.

yond sparse 3D-box labels. *(i) Region→3D:* when a 2D box is available, the model predicts its 3D box directly from the region prompt. *(ii) Pure text→3D:* when no 2D box exists, the model localizes the mentioned entity via its built-in textual grounding and regresses its 3D box, enabling coverage of text-only data. In addition, we construct an auxiliary dense region-to-3D supervision: from ground-truth or predicted depth maps, we randomly sample valid surface points per image (*e.g.*, 100 per image) and train the model to predict their 3D coordinates conditioned on the corresponding region prompt. This depth-driven signal scales supervision well beyond limited 3D-box annotations. Finally, to tolerate modest grounding noise, we apply lightweight 2D bounding-box augmentation (jitters in size and location), improving robustness while preserving semantic locality.

Together, region-prompt grounding, structured 3D box representation, intrinsic normalization, and scalable training signals address both linguistic and geometric ambiguities of monocular 3D grounding. These components jointly provide a camera-relative spatial understanding that generalizes across datasets and supports future extensions to multi-view and embodied reasoning tasks.

2.4. Data Construction and Composition

Data Construction for Grounding. To construct the implicit grounding corpus, we start from RefSpatial [?], which includes 2D samples from OpenImages [?], 3D video data from CA-1M [?], and synthetic scenes. RefSpatial contains diverse image–text pairs, but it lacks region-level annotations for all the mentioned entities. To obtain them, we use Florence-2 [?] to generate candidate 2D bounding boxes and class labels for each textual mention, producing dense but noisy region annotations.

Methods	Acc. (%)
Human	98.3
GPT-4V-Turbo [?]	66.9
GPT-4o [?]	64.5
LLaVA-v1.5-7B-xtuner [?]	50.8
CogVLM-7B [?]	50.8
LLaVA-v1.5-7B [?]	51.6
LLaVA-InternLM2-7B [?]	52.4
SpatialRGPT-8B* [?]	87.9
SR3D-8B* [?]	90.3
GR3D-8B (Ours)	**94.4**

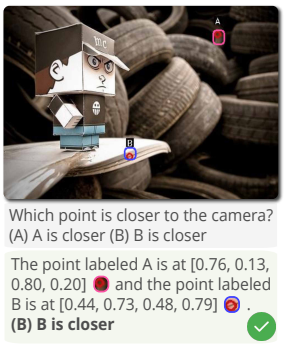

Figure 3. Results on the BLINK-Depth benchmark for point-level region spatial understanding. Left: comparison with VLM baselines. Right: visualization of one sample. Our method surpasses prior Region-VLMs (*), which require manual annotated masks.

We then refine these annotations through a VLM for verification and a rephrasing pipeline. This process (i) verifies one-to-one alignment between textual mentions and detected regions, removing unmatched or ambiguous cases, and (ii) rewrites generic class names into concise, instance-level descriptions based on image context. The resulting corpus provides high-quality implicit grounding supervision that links textual mentions to corresponding visual evidence with precise instance semantics.

For explicit grounding, we augment samples that contain ground-truth boxes by generating short instance-level referring expressions with a vision–language model and validating their existence in the image. Only verified matches are retained. Together, these procedures yield reliable implicit and explicit grounding data, while depth, point, and 3D-box supervision follow the setup in Sec. 2.3.

Data Composition and Distribution. Our training data is composed of publicly available sources: 97K grounded CoT samples, 780K 3D detection samples from Omni3D [?] and EmbodiedScan [?], and 272K pointmap reconstruction samples from DepthLM [?]. We do not use any proprietary or in-house data and the scale of our 3D detection data is comparable to prior works such as VST [?], ensuring performance gains are not simply due to data size.

Method	SPATIAL											GENERAL			
	BLINK [?]			CVBench [?]				RWQA [?]	ERQA [?]	SAT [?]	EMB [?]	ChartQA [?]	MME [?]	POPE [?]	AI2D [?]
	Dep.	Spa.	Avg.	Rel.	Dep.	Dis.	Avg.								
NVILA-Lite-8B	73.38	79.72	50.51	93.38	92.83	91.00	86.31	65.35	36.25	62.60	68.90	84.80	1692	88.10	91.01
GR3D-8B (Stage 1)	**87.90**	**83.21**	**54.35**	**96.92**	**98.16**	95.50	87.23	**68.75**	**40.25**	**76.00**	**81.01**	**84.48**	**1656**	**88.23**	**91.81**
GR3D-8B (Stage 2)	**87.90**	80.41	53.26	96.46	98.00	**96.00**	**87.26**	65.23	38.50	70.60	77.58	84.00	1626	87.00	91.54

Table 3. Performance comparison on general visual question answering and spatial reasoning benchmarks.

3. Experiments

In this section, we begin by describing the implementation details (Sec. 3.1), including the training stages and datasets used. We then present the main results of our model, highlighting its 3D detection performance in Sec. 3.2. Next, we assess whether the model preserves its general VLM and spatial capabilities in Sec. 3.3. In Sec. 3.4, we evaluate the visual grounded CoT enabled by our implicit grounding approach. Finally, Sec. 3.5 provides additional analysis and ablation studies of the model's 3D detection performance.

3.1. Implementation Details

Our model is trained in two stages as detailed in following.

Stage 1: Spatial Pretraining. The goal of this stage is to strengthen the model's spatial understanding and 2D grounding capabilities, which later improves its 3D detection performance, as shown in our analysis. We initialize the visual encoder, projector, and LLM from NVILA-Lite 8B, while the spatial positional encoding module is newly initialized. Training is performed on a data mixture similar to SR-3D, augmented with 2D grounding data and region-to-3D detection data from Sec. 2.4. During this stage, we freeze the visual encoder and train the remaining modules.

Stage 2: Detection CoT Finetuning. After pretraining, the model already possesses strong 2D grounding and basic 3D detection abilities. We then fine-tune it on CoT-oriented detection data, including detection data in CoT format (curated from Omni3D by first grounding in 2D and then predicting 3D boxes). Since the visual features are already well-formed after Stage 1, we fine-tune only the LLM to learn the reasoning and text-generation structure.

3.2. 3D Object Detection

We evaluate our model on the Omni3D test set, following the benchmark protocol and hyperparameters used in DetAny3D. The Omni3D benchmark reports Average Precision (AP), where predictions are matched to ground-truth using 3D IoU with thresholds ranging from 0.05 to 0.50.

For comparison, we include both vision-specialist baselines (e.g., ImVoxelNet [?], Cube R-CNN [?], OVMono3D [?], and DetAny3D [?]) and VLM-based baselines (e.g., Qwen3VL-4B [?] and Qwen3VL-8B [?]. Our main results are shown in Table 1 and Fig. 4, where our model outperforms all VLM baselines. Compared with vision specialists, our model achieves competitive results overall and delivers notably better performance on indoor datasets.

We further analyze why existing VLMs perform worse on 3D detection. First, unlike our approach, they do not disentangle 3D detection into a two-step process—2D grounding followed by 3D box prediction. As we show in the analysis (Sec. 3.5), 2D grounding provides a stable geometric anchor that leads to more reliable and consistent 3D predictions. Second, existing VLMs struggle with handling camera intrinsics. Qwen3VL is highly sensitive to input resolution, since pixel dimensions implicitly encode the focal length used in its geometric reasoning. This makes its 3D predictions unstable under changes in image size. VST [?] partially addresses this by normalizing focal length in a manner similar to ours. However, it still requires FoVs to be passed as text prompts. Representing metric geometric parameters in textual form is difficult for the model to parse and integrate reliably, which limits its 3D understanding across scenes and camera setups.

Since our method explicitly separates 2D grounding from 3D prediction, we also evaluate 2D grounding performance on the Omni3D benchmark. As shown in Table 2, our model exceeds region proposals generated by Cube R-CNN and the Qwen3-VL family. For Qwen3-VL models, which do not perform explicit 2D grounding, we evaluate using 2D boxes projected from their predicted 3D outputs.

3.3. Visual Question Answering

We investigate two key questions: (1) whether Stage 1 spatial pre-training effectively improves spatial reasoning performance, and (2) whether Stage 2 detection CoT finetuning negatively affects the model's general VQA capabilities. We evaluate two variants of our model: one after spatial pre-training and one after CoT finetuning. The results are presented in Table 3. After spatial pre-training, the model shows a clear improvement on spatial-related VQA benchmarks, confirming the effectiveness of this stage. In contrast, Stage 2 finetuning focuses on learning the structure of CoT reasoning, and the results indicate that it does not significantly reduce general VQA performance. Most benchmarks remain similar to the Stage 1 model, suggesting that the model maintains strong general-purpose abilities.

3.4. Implicit Grounding CoT

We aim to evaluate two aspects of our implicit grounding approach: (1) how accurate the grounding is, and (2)

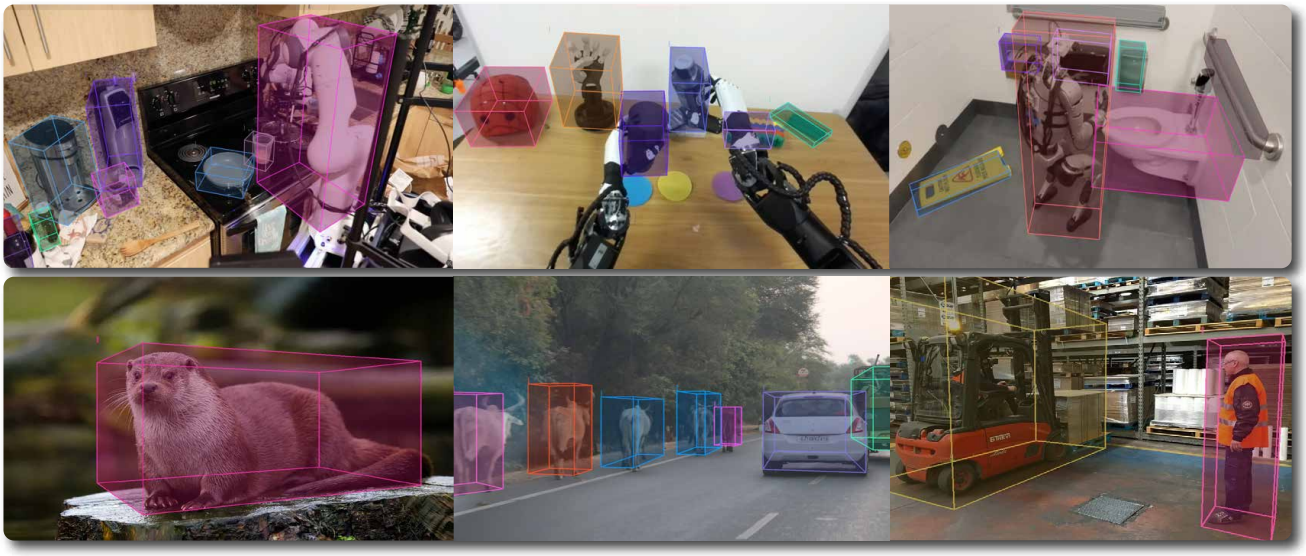

Figure 4. Qualitative results on 3D object detection. Our model produces accurate 3D bounding boxes on in-the-wild samples.

whether the grounding genuinely contributes to correct answers rather than producing hallucinated reasoning.

To study this, we evaluate our model on the MM-GCoT [**?**] benchmark, which provides three key metrics: answer accuracy (A-Acc), grounding accuracy (G-Acc), and answer–grounding consistency (Consist.). A-Acc measures the correctness of the textual answer. G-Acc follows the Acc@0.5 protocol, where a prediction is considered correct if its IoU with the ground-truth box exceeds 0.5. The consistency metric measures the percentage of predictions where both the answer and the grounding box are correct. We show results in Table 4, where our method outperforms baselines in all these metrics.

To further evaluate the performance in spatial reasoning scenarios, we conduct experiments on BLINK-Depth using the same grounding-based CoT formulation. As shown in Table 3, our method surpasses prior Region-VLMs, which are typically strong on this benchmark but require manually annotated masks as input. In contrast, our model achieves higher performance while performing grounding automatically. We additionally provide qualitative examples demonstrating that our model can accurately localize tiny regions and successfully handle point-level areas.

3.5. Analysis and Ablation Study

2D→3D vs Direct 3D Prediction. As shown in Table 5, first grounding the target region in 2D and then predicting its 3D bounding box leads to a clear improvement over direct 3D prediction. This two-step design is more vision-centric, as it explicitly forces the model to learn object-specific visual features before performing 3D reasoning. It also naturally decomposes the task into two subproblems—2D grounding and 3D inference—where the former benefits from significantly larger amounts of training data across generic detection and grounding datasets. Leveraging this abundant 2D supervision allows the model to establish stronger spatial priors, which in turn improves downstream 3D detection performance.

Do spatial pretraining help 3D detection? Table 5 further supports this assumption by showing that spatial pretraining noticeably improves performance in the outdoor domain. The Omni3D dataset is highly imbalanced [**?**], with far fewer outdoor training samples compared to indoor scenes. As a result, models trained from scratch struggle to generalize in outdoor settings. Spatial pretraining provides a strong remedy by injecting generic 2D spatial and grounding knowledge, enabling the model to better transfer its learned priors to the 3D detection task. This demonstrates that leveraging 2D supervision is especially beneficial when 3D data is limited or unevenly distributed.

Effect of Intrinsic Normalization. Intrinsic normalization yields a modest, yet consistent improvement. Although its impact is smaller than the two factors discussed above, normalizing intrinsics helps reduce systematic biases when the model encounters cameras with different focal lengths. Without this normalization, the model may lead to small but noticeable localization offsets in the predicted 3D boxes.

Contribution of Pointmap Reconstruction. We further analyze the effect of pointmap reconstruction as an auxiliary task for 3D detection. This supervision strengthens the model's ability to align region-level visual features with their corresponding 3D geometry. To isolate this effect from 2D grounding quality, we use ground-truth 2D boxes as our model input and evaluate only the 3D prediction. This separation is enabled by our disentangled pipeline and allows us to directly measure the reconstruction capability. As shown in Fig. 5, increasing the amount of pointmap supervision yields a clear scaling trend on SUN-RGBD: more pointmap

	Attribute			Judgement			Object			Average		
	Acc_{A} ↑	Acc_{G} ↑	Cons. ↑	Acc_{A} ↑	Acc_{G} ↑	Cons. ↑	Acc_{A} ↑	Acc_{G} ↑	Cons. ↑	Acc_{A} ↑	Acc_{G} ↑	Cons. ↑
Qwen2.5-VL-7B [?](AF)	73.6	72.5	59.8	87.9	56.3	51.5	57.8	64.1	59.1	73.1	64.3	56.8
Qwen2.5-VL-7B [?](GF)	48.8	82.6	45.7	80.6	72.8	62.4	26.7	62.3	32.6	52.0	72.6	46.9
LLaVA-7B [?](AF)	68.6	9.2	8.8	83.0	11.2	11.5	58.4	9.1	9.9	70.0	9.8	10.1
LLaVA-7B [?](GF)	59.7	6.3	5.6	82.5	0.5	0.6	35.9	5.5	9.7	59.4	4.1	5.3
LLaVA-GCoT-7B [?]	72.8	66.7	56.1	88.3	61.7	56.9	62.3	61.7	61.3	74.5	63.3	58.1
GR3D-8B (Ours)	**78.9**	**77.3**	**66.7**	85.0	**79.6**	**70.4**	**71.1**	**65.7**	**66.1**	**78.3**	**74.2**	**67.7**

Table 4. Results on the MM-GCoT benchmark. "AF" and "GF" correspond to answer-first and grounding-first prompting settings. Acc_{A}, Acc_{G}, and Cons. refer to answer accuracy, grounding accuracy, and consistency between them.

2D→3D	PT	Cam	$\text{AP}^{\text{sun}}_{15}$ ↑	$\text{AP}^{\text{sun}}_{3D}$ ↑	$\text{AP}^{\text{kit}}_{15}$ ↑	$\text{AP}^{\text{kit}}_{3D}$ ↑
–	–	–	30.19	20.27	10.08	6.22
✓	–	–	42.29	29.87	15.61	10.03
✓	✓	–	41.24	30.95	21.55	14.35
✓	✓	✓	**43.49**	**31.64**	**22.18**	**14.75**

Table 5. Ablation study on the key components of GR3D-8B. "PT" denotes pretraining, "2D→3D" denotes 2D grounding followed by 3D prediction, and "Cam" denotes using normalized intrinsics.

data consistently improves 3D detection performance.

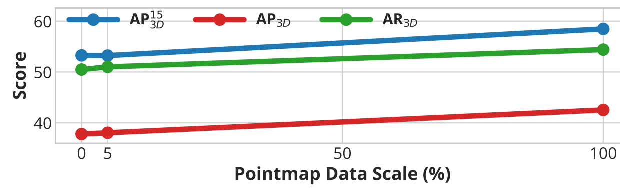

Figure 5. Scaling behavior when increasing pointmap prediction data on SUN-RGBD. More pointmap supervision leads to better 3D detection performance.

4. Related Work

Spatial Vision Language Models. Recent work has rapidly expanded the spatial capabilities of VLMs across 2D grounding, monocular spatial reasoning, and multi-view 3D scene understanding [**? ?**]. Representative 2D spatial VLMs such as SpatialVLM [**?**], SpatialPin [**?**], VST [**?**], and SpatialLadder [**?**] focus on image-plane relations including relative position, direction, and distance using explicit spatial cues. SRGPT [**?**] improves fine-grained single-view spatial perception by introducing a region branch for more precise region-level querying. SR-3D [**?**] extends this idea by preserving and enabling multi-view spatial reasoning through a unified visual tokens space. Other multi-view spatial VLMs [**? ?**] incorporate 3D cues or cross-view alignment for scene reasoning. Despite this progress, implicit 2D grounding and monocular 3D grounding from a single image remain underexplored [**? ? ?**]. In contrast, our approach jointly addresses both problems without requiring any spatial annotations at inference time.

Monocular 3D Grounding. Traditional 3D object detection has long focused on single-dataset, closed-set scenarios [**? ? ? ? ? ? ? ? ?**], achieving strong performance but suffering from poor generalization to new environments. Initial efforts to overcome this [**? ?**] utilized multi-dataset training to create universal detectors. The Omni3D [**?**], for instance, aggregated a wide variety of 3D datasets and proposed a universal model trained jointly on them. However, these models still confined to a predefined list of object classes seen during training. More recent work [**? ? ?**] have turned their focus to the open-vocabulary setting. OVMono3D [**?**] proposes a two-stage "detect-then-lift" pipeline: it first employs an off-the-shelf 2D open-vocabulary detector [**?**] to generate 2D proposals, then feeds these region features into a specialized 3D head to regress 3D parameters. DetAny3D [**?**] proposes a more integrated, promptable architecture that directly fuses features from 2D foundation models and uses 2D prompts, *e.g.*, points, boxes, or text to query the model for 3D outputs. Instead of treating 3D grounding as a standalone detection task, GR3D predicts 3D boxes as part of a unified VLM framework that also includes dynamic implicit 2D grounding. This unified approach allows GR3D to leverage grounding as a key driver to enhance general spatial alignment and geometry-consistent reasoning, improving performance on both grounded and non-grounded spatial tasks.

Thinking with Images. Our work is also realted to the recent line of "Thinking with Images" work [**? ? ? ? ? ? ? ?**]. Different from these approaches, GR3D avoids explicit visual thought processes and external tools, offering a more efficient and unified design through implicit 2D grounding and native 3D reasoning within the VLM's generative flow.

5. Conclusion

We introduced GR3D, a spatial VLM that integrates explicit 2D grounding, implicit grounding for CoT reasoning, and monocular 3D grounding within a single framework. By enabling the model to reference visual evidence during generation and by coupling region-grounded queries with 3D box prediction, GR3D decomposes spatial understanding into grounded 2D perception followed by 3D inference. GR3D delivers consistent gains across various benchmarks, showing that grounding serves as an effective inductive bias for better spatial understanding in VLMs.

Acknowledgements. This project was supported, in part, by NSF CAREER Award IIS-2240014, gifts from Amazon, Meta, and Qualcomm.

Grounded 3D-Aware Spatial Vision-Language Modeling

Supplementary Material

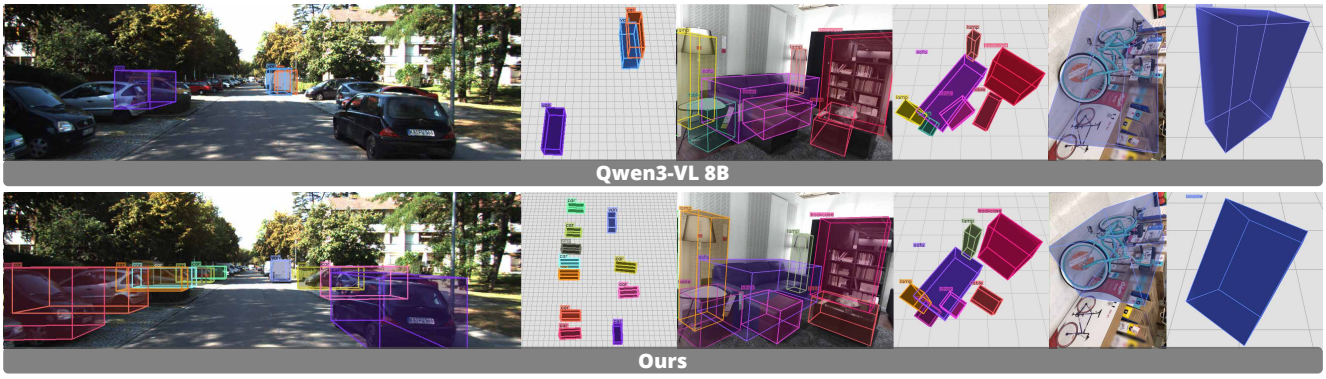

Figure 6. Qualitative comparison on 3D object detection between our model and Qwen3-VL 8B [**?**]. Our model produces more accurate 3D bounding boxes with fewer missed objects, demonstrating stronger spatial grounding and detection reliability.

Table of Contents

1. More Results on 3D Detection

We show a qualitative comparison on 3D object detection between GR3D and Qwen3-VL-8B [**?**]. 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 [**?**], 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 [**?**], demonstrating its ability to reason about complex spatial relations in 2D. We further evaluate on the widely used RefCOCO, RefCOCO+, and RefCOCOg datasets [**?** **?**] 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 [**?**] or Qwen2.5-VL [**?**], confirming that our strong 2D grounding ability generalizes well to both spatial and standard referring benchmarks.

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

Table 6. Performance comparison on RefSpatial [**?**].

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 [**?**]. 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 [?]	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0
Gemini-1.5 Flash [?]	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8	42.1
Gemini-1.5 Pro [?]	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.3
Open-source Models									
InternVL2-2B [?]	24.9	22.0	35.0	33.8	44.2				
InternVL2-8B [?]	31.3	29.0	48.9	44.2	38.0	33.4	28.9	46.4	37.5
InternVL2-40B [?]	41.3	26.2	48.2	27.5	47.6	32.7	27.8	44.7	37.0
LongVILA-8B [?]	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5	21.6
VILA-1.5-8B [?]	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8	28.9
VILA-1.5-40B [?]	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2
LongVA-7B [?]	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
LLaVA-Video-7B [?]	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6	35.6
LLaVA-Video-72B [?]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
LLaVA-OneVision-7B [?]	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4	32.4
LLaVA-OneVision-72B [?]	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 [? ?] following SR-3D [?], and then evaluate multi-view global spatial scene understanding on VSI-Bench [?]. 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 [?]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
UNINEXT-H [?]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
ONE-PEACE [?]	92.6	94.2	89.3	88.8	92.2	83.2	89.2	89.3
Vision Language Models								
InternVL3-1B [?]	85.8	90.1	81.7	76.6	84.1	69.2	82.8	82.6
InternVL3.5-1B [?]	85.4	89.7	80.2	77.7	85.5	69.5	81.9	81.6
InternVL3-2B [?]	89.8	92.6	86.4	84.0	89.2	76.5	87.6	87.2
InternVL3.5-2B [?]	88.7	91.6	84.8	82.7	88.4	76.6	85.6	85.5
Qwen2.5-VL-3B [?]	89.1	91.7	84.0	82.4	88.0	74.1	85.2	85.7
Shikra-7B [?]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
CogVLM-G [?]	92.8	94.8	89.0	88.7	92.9	83.4	89.8	90.8
Qwen2-VL-7B [?]	91.7	93.6	87.3	85.8	90.5	79.5	87.3	87.8
Qwen2.5-VL-7B [?]	90.0	92.5	85.4	84.2	89.1	76.9	87.2	87.2
TextHawk2 [?]	91.9	93.0	87.6	86.2	90.0	80.4	88.2	88.1
InternVL3.5-8B [?]	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 [? ?]. Baseline numbers are taken from [? ?]. GR3D achieves grounding accuracy comparable to vision specialists [? ? ?] models and performs on par with top VLMs such as InternVL3.5 [?].

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 [? ?] following SR-3D [?], and then evaluate multi-view global spatial scene understanding on VSI-Bench [?]. 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 [?] benchmark. However, ScanRefer assumes access

to a pre-aligned world coordinate space, which is not directly compatible with the settings of Qwen3-VL [**?**] 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 [**?**], MMSI-Bench [**?**] and SPAR-Bench [**?**], 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 [**?**], ScanQA [**?**], MMSI-Bench [**?**], and SPAR-Bench [**?**].

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 [**?**] 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* [**?**], 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 [**? ?**] or as "visual programmers" that generate code for custom analysis and edits [**? ? ?**]. Others generate intermediate visual representations to guide reasoning, often called Visual Chain of Thought (V-CoT) [**?**]. These V-CoT methods may interleave text with explicit visual groundings [**?**], sketch visual artifacts [**?**], generate subgoal images for robotics [**?**], or perform planning entirely through

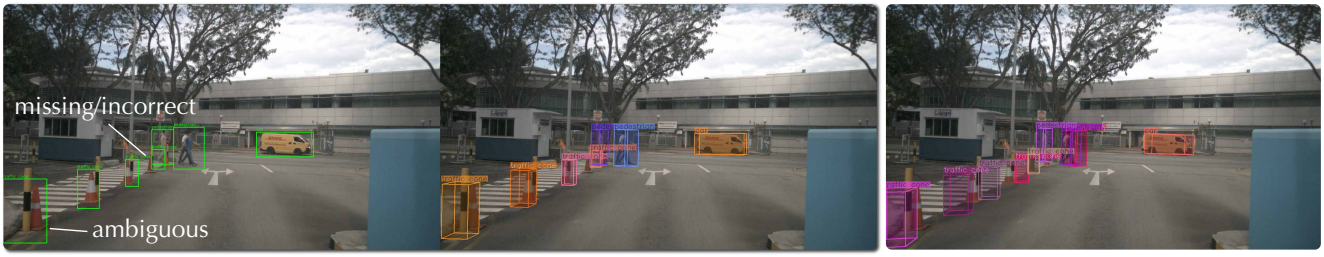

Figure 7. A failure case analysis of GR3D compared to Qwen3-VL-8B.

visual state sequences [**?**]. 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 [**?**].

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.