# 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

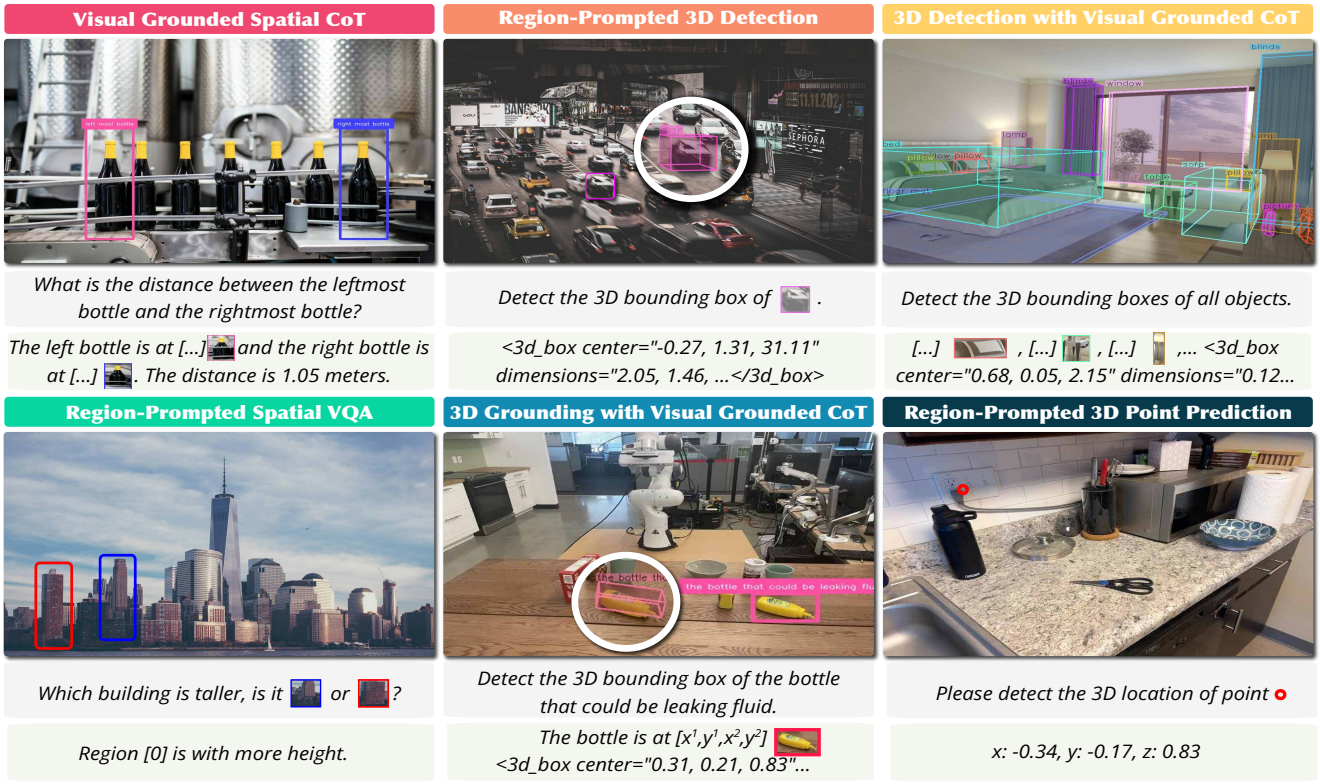


Figure 1. **GR3D overview.** Top-Left: Spatial CoT enabled by 2D implicit grounding. Top-Middle: Region-prompted 3D grounding predicts camera-relative 3D boxes. Top-Right: Grounded 3D detection performs multi-object 3D prediction. Bottom-Left: Region-prompted spatial QA. Bottom-Middle: Referring-based 3D grounding. Bottom-Right: Region-prompted point prediction.

## 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.*

[1]Work done during an internship at NVIDIA.

## 1. Introduction

Vision–language models (VLMs) have rapidly evolved into general-purpose perception–language systems [1–8], 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 [9–17], 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 [18–20]. 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 [21–29]. 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 [30], rely on multi-view supervision [31], or are limited by the scarcity of 3D box annotations [32].

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 [33], ERQA [30] and SAT [34], 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.

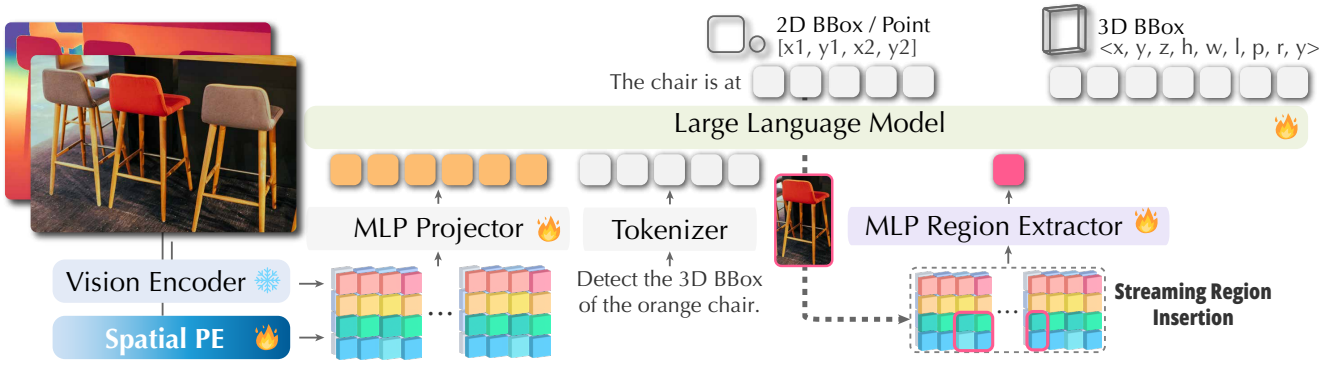


Figure 2. Method overview. GR3D builds on Region-VLMs by adding streaming region insertion for visual Chain-of-Thought reasoning. During CoT, the model repeatedly predicts a region, extracts its visual embedding, and reinserts a region token into the text sequence, enabling step-by-step spatial reasoning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

## 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 [22] to construct a foundational Spatial VLM based on the NVILA-8B-Lite [3] 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 multi-modal 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>[x1, y1, x2, y2]</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 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 [35] | | ARKITSCENES [36] | | OBJECTRON [37] | | HYPERSIM [38] | | KITTI [39] | | NUSCENES [40] | | $AP_{3D}$ ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $AP_{15}$ ↑ | mAP ↑ | $AP_{15}$ ↑ | mAP ↑ | $AP_{15}$ ↑ | mAP ↑ | $AP_{15}$ ↑ | mAP ↑ | $AP_{15}$ ↑ | mAP ↑ | $AP_{15}$ ↑ | mAP ↑ | |
| *Vision Specialist Models* | | | | | | | | | | | | | |
| ImVoxelNet [41] | - | - | - | - | - | - | - | - | - | - | - | - | 9.4 |
| SMOKE [42] | - | - | - | - | - | - | - | - | - | - | - | - | 10.4 |
| Cube R-CNN [32] | - | 15.33 | - | 41.73 | - | 50.84 | - | 7.48 | - | 32.50 | - | 30.06 | 23.26 |
| OVMono3D [43]$_{w/\ Cube\ R\text{-}CNN}$ | - | 15.20 | - | 41.60 | - | 58.87 | - | 7.75 | - | 25.45 | - | 24.33 | 22.98 |
| DetAny3D [44]$_{w/\ Cube\ R\text{-}CNN}$ | 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 [45] | 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 [45] | 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 [32] 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 | $AP^{sun}_{2D}$ | $AP^{ark}_{2D}$ | $AP^{obj}_{2D}$ | $AP^{hyp}_{2D}$ | $AP^{kit}_{2D}$ | $AP^{nus}_{2D}$ |
|---|---|---|---|---|---|---|
| Cube R-CNN [32] | 15.07 | 40.22 | 49.24 | 11.05 | 36.14 | 34.64 |
| Qwen3-VL-8B [45] | 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 [23], which includes 2D samples from OpenImages [53], 3D video data from CA-1M [54], 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 [55] 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 [7] | 66.9 |
| GPT-4o [58] | 64.5 |
| LLaVA-v1.5-7B-xtuner [59] | 50.8 |
| CogVLM-7B [60] | 50.8 |
| LLaVA-v1.5-7B [61] | 51.6 |
| LLaVA-InternLM2-7B [62] | 52.4 |
| SpatialRGPT-8B* [21] | 87.9 |
| SR3D-8B* [22] | 90.3 |
| **GR3D-8B** (Ours) | **94.4** |

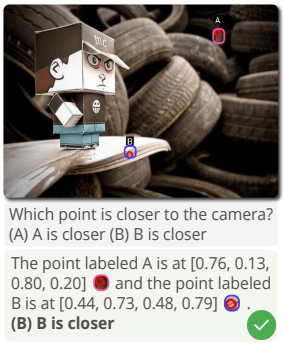


Figure 3. Results on the BLINK-Depth benchmark for point-level region spatial understanding. Left: comparison with VLM baselines. Right: visualization of one sample. Our method surpasses prior Region-VLMs (*), which require manual annotated masks.

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 [32] and EmbodiedScan [56], and 272K pointmap reconstruction samples from DepthLM [57]. 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 [26], ensuring performance gains are not simply due to data size.

| Method | SPATIAL | | | | | | | | | | | GENERAL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLINK [46] | | | CVBench [33] | | | | RWQA [47] | ERQA [30] | SAT [34] | EMB [48] | ChartQA [49] | MME [50] | POPE [51] | AI2D [52] |
| | 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 [41], Cube R-CNN [32], OVMono3D [43], and DetAny3D [44]) and VLM-based baselines (e.g., Qwen3VL-4B [45] and Qwen3VL-8B [45]. 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 [26] 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)

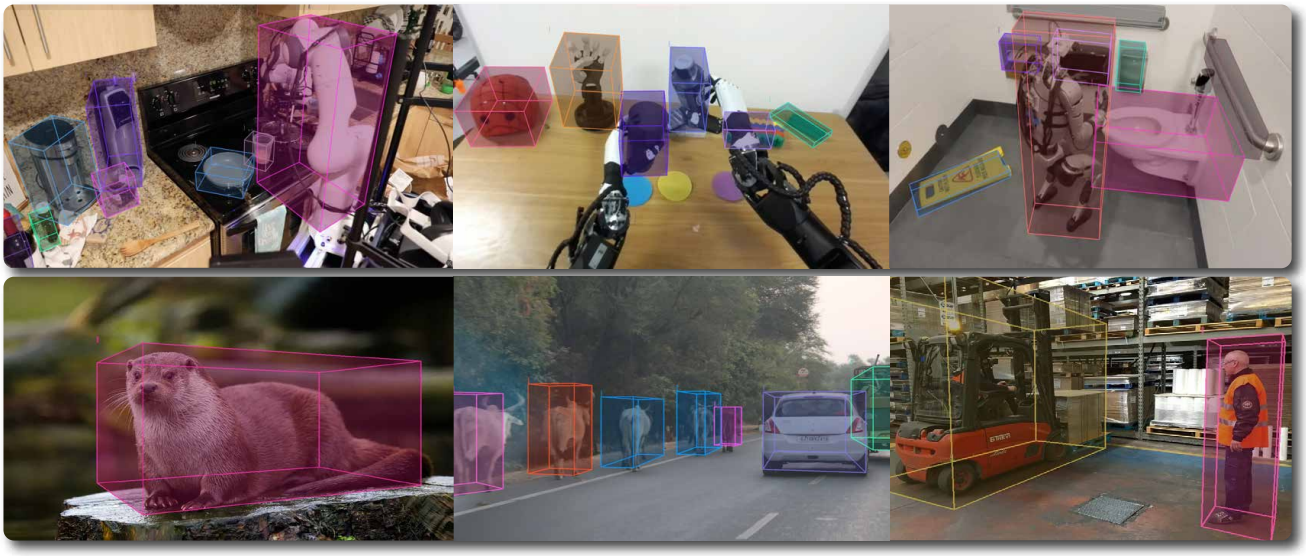

Figure 4. Qualitative results on 3D object detection. Our model produces accurate 3D bounding boxes on in-the-wild samples.

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

To study this, we evaluate our model on the MM-GCoT [63] 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 [44], 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 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\text{Acc}_\text{A}$ ↑ | $\text{Acc}_\text{G}$ ↑ | Cons. ↑ | $\text{Acc}_\text{A}$ ↑ | $\text{Acc}_\text{G}$ ↑ | Cons. ↑ | $\text{Acc}_\text{A}$ ↑ | $\text{Acc}_\text{G}$ ↑ | Cons. ↑ | $\text{Acc}_\text{A}$ ↑ | $\text{Acc}_\text{G}$ ↑ | Cons. ↑ |
| Qwen2.5-VL-7B [64](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 [64](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 [1](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 [1](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 [63] | 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. $\text{Acc}_\text{A}$, $\text{Acc}_\text{G}$, 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.

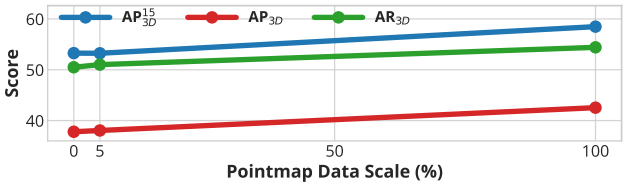


Figure 5. Scaling behavior when increasing pointmap prediction data on SUN-RGBD. More pointmap supervision leads to better 3D detection performance.

## 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 [25, 26, 34, 48, 65–79]. Representative 2D spatial VLMs such as SpatialVLM [65], SpatialPin [66], VST [26], and SpatialLadder [79] focus on image-plane relations including relative position, direction, and distance using explicit spatial cues. SRGPT [21] improves fine-grained single-view spatial perception by introducing a region branch for more precise region-level querying. SR-3D [22] extends this idea by preserving and enabling multi-view spatial reasoning through a unified visual tokens space. Other multi-view spatial VLMs [27, 28] 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 [26, 45, 80]. 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 [22, 41, 42, 81–86], achieving strong performance but suffering from poor generalization to new environments. Initial efforts to overcome this [32, 87] utilized multi-dataset training to create universal detectors. The Omni3D [32], 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 [43, 44, 88] have turned their focus to the open-vocabulary setting. OVMono3D [43] proposes a two-stage "detect-then-lift" pipeline: it first employs an off-the-shelf 2D open-vocabulary detector [89] to generate 2D proposals, then feeds these region features into a specialized 3D head to regress 3D parameters. DetAny3D [44] 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 [90–97]. 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.

## References


[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 8

[2] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 2

[3] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *CVPR*, 2025. 3

[4] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2024.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv:2308.12966*, 2023.

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 2024.

[7] OpenAI. Gpt-4 technical report, 2023. arXiv:2303.08774. 5

[8] Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 2

[9] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021. 2

[10] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

[11] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.

[12] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *RSS*, 2024.

[13] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *RSS*, 2025.

[14] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

[15] NVIDIA Research. GR00T N1.5: An Improved Open Foundation Model for Generalist Humanoid Robots. https://research.nvidia.com/labs/gear/gr00t-n1_5, 2025.

[16] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[17] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, and Ranjay Krishna. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025. 2

[18] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *CoRL*, 2022. 2

[19] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. In *RSS*, 2023.

[20] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *ICRA*, 2024. 2

[21] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024. 2, 5, 8

[22] An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, et al. 3d aware region prompted vision language model. *arXiv preprint arXiv:2509.13317*, 2025. 3, 5, 8, 1, 2

[23] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. 5, 1

[24] BAAI RoboBrain Team. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 1

[25] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. In *CoRL*, 2024. 8, 1

[26] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025. 5, 6, 8

[27] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. In *ICCV*, 2025. 8

[28] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, 2025. 8

[29] Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv:2507.23478*, 2025. 2

[30] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 2, 6

[31] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[32] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, 2023. 2, 5, 6, 8

[33] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 2, 6

[34] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Dynamic spatial aptitude training for multimodal language models. In *COLM*, 2025. 2, 6, 8

[35] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5

[36] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 5

[37] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*, 2021. 5

[38] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 5

[39] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 2013. 5

[40] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[41] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*, 2022. 5, 6, 8

[42] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPRW*, 2020. 5, 8

[43] Jin Yao, Hao Gu, Xuweiyi Chen, Jiayun Wang, and Zezhou Cheng. Open vocabulary monocular 3d object detection. *arXiv preprint arXiv:2411.16833*, 2024. 5, 6, 8

[44] Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and Zetong Yang. Detect anything 3d in the wild. In *CVPR*, 2025. 5, 6, 7, 8

[45] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 5, 6, 8, 1, 3, 4

[46] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024. 6

[47] xAI. RealWorldQA: A benchmark dataset for real-world spatial understanding. https://huggingface.co/datasets/visheratin/realworldqa, 2024. 6

[48] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *ACL*, 2024. 6, 8

[49] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *ACL*, 2022. 6

[50] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025. 6

[51] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 6

[52] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *ECCV*, 2016. 6

[53] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 5

[54] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *CVPR*, 2025. 5

[55] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. 5

[56] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai

Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 5, 2

[57] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025. 5

[58] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 5, 2

[59] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 5

[60] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, 2024. 5, 2

[61] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 5

[62] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 5

[63] Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025. 7, 8

[64] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv:2502.13923*, 2025. 8, 1, 2

[65] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 8, 1

[66] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In *NeurIPS*, 2024. 8

[67] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *ICRA*, 2025.

[68] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *ICCV*, 2025.

[69] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning. In *Findings of EMNLP*, 2025.

[70] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *CVPR*, 2025.

[71] Mingjie Xu, Mengyang Wu, Yuzhi Zhao, Jason Chun Lok Li, and Weifeng Ou. Llava-spacesgg: Visual instruct tuning for open-vocabulary scene graph generation with enhanced spatial relations. In *WACV*, 2025.

[72] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *CVPR*, 2025.

[73] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv:2501.10074*, 2025.

[74] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *EMNLP*, 2024.

[75] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025. 2

[76] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024.

[77] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Shawn Ma, Baoxiong Jia, and Siyuan Huang. Multimodal situated reasoning in 3d scenes. In *NeurIPS*, 2024.

[78] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv:2505.23747*, 2025.

[79] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. 8

[80] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *CVPR*, 2024. 8

[81] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 8

[82] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *TPAMI*, 2024.

[83] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *NeurIPS*, 2022.

[84] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, 2021.

[85] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

[86] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, 2023. 8

[87] Zhuoling Li, Xiaogang Xu, SerNam Lim, and Hengshuang Zhao. Unimode: Unified monocular 3d object detection. In *CVPR*, 2024. 8

[88] Zhenyu Wang, Yali Li, Taichi Liu, Hengshuang Zhao, and Shengjin Wang. Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In *ECCV*, 2024. 8

[89] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*. Springer, 2024. 8, 2

[90] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025. 8, 3

[91] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-ReAct: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 3

[92] Kaitao Chen, Shaohao Rui, Yankai Jiang, Jiamin Wu, Qihao Zheng, Chunfeng Song, Xiaosong Wang, Mu Zhou, and Mianxin Liu. Think twice to see more: Iterative visual reasoning in medical vlms. *arXiv preprint arXiv:2510.10052*, 2025. 3

[93] Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *ICCV*, 2023. 3

[94] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. In *ICLR*, 2026.

[95] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. In *ICML*, 2025. 3

[96] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *NeurIPS*, 2024. 3

[97] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024. 8

[98] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 2

[99] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2

[100] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 2

[101] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[102] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, 2025. 1

[103] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[104] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2

[105] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *ICLR*, 2025. 2

[106] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024. 2

[107] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 2

[108] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *TMLR*, 2025. 2

[109] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. 2, 3

[110] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 2

[111] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 2

[112] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2

[113] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023. 2

[114] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024. 2, 3

[115] Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *arXiv preprint arXiv:2404.09204*, 2024. 2

[116] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 2, 3

[117] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026. 3

[118] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *NeurIPS*, 2025. 3

[119] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, 2024. 3

[120] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. GRIT: Teaching mllms to think with images. In *NeurIPS*, 2025. 3

[121] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, 2025. 3

[122] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *ICLR*, 2026. 4