LocateAnything3D: Vision-Language 3D Detection with Chain-of-Sight

Yunze Man[1,3*], Shihao Wang[2], Guowen Zhang[2], Johan Bjorck[3]
Liang-Yan Gui[1], Jim Fan[3], Jan Kautz[3], Yu-Xiong Wang[1†], Zhiding Yu[3†]

[1]University of Illinois Urbana-Champaign [2]The Hong Kong Polytechnic University [3]NVIDIA

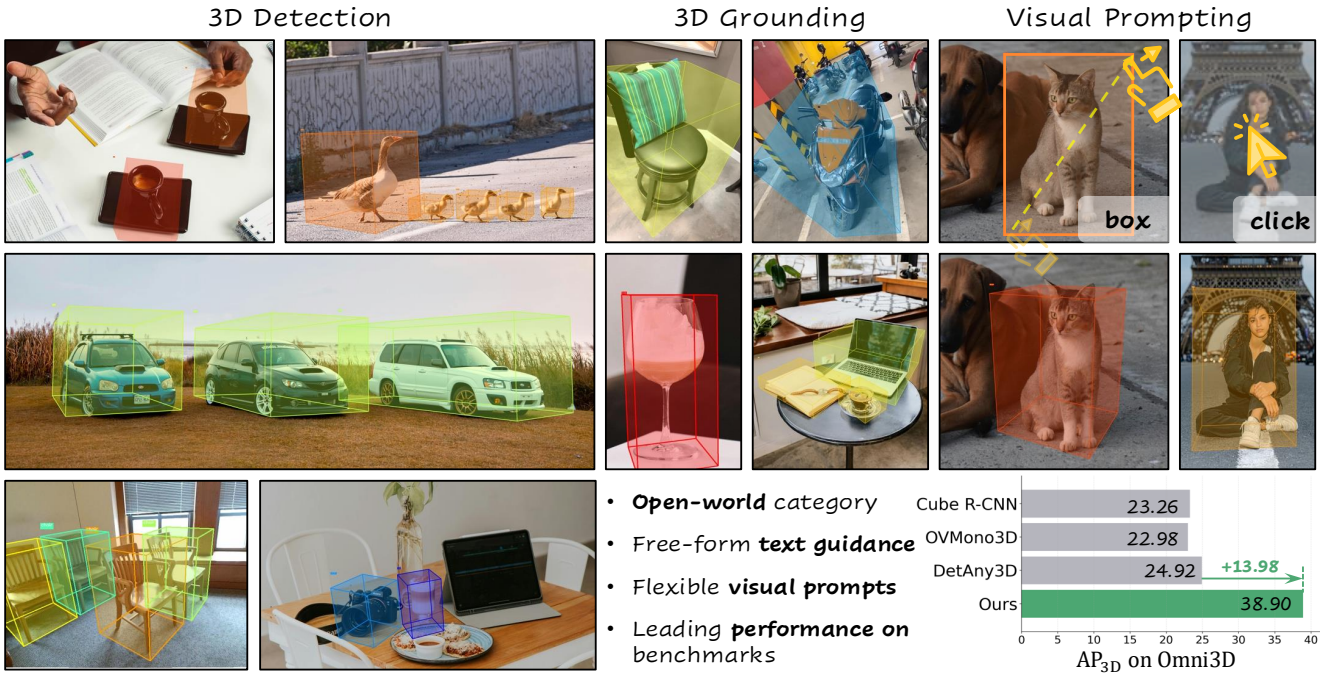

Figure 1. **LocateAnything3D** unifies 3D detection and grounding in a single vision-language model. It supports open-world categories with free-form text guidance and flexible visual prompts (*e.g.*, drag boxes, click points). All examples are zero-shot, highlighting strong out-of-domain generalizability. The bar chart (right) shows that LocateAnything3D achieves state-of-the-art $\mathrm{AP}_{\mathrm{3D}}$ on the Omni3D benchmark.

Abstract

To act in the world, a model must name what it sees and know where it is in 3D. Today's vision-language models (VLMs) excel at open-ended 2D description and grounding, yet multi-object 3D detection remains largely missing from the VLM toolbox. We present ***LocateAnything3D****, a VLM-native recipe that casts 3D detection as a next-token prediction problem. The key is a short, explicit Chain-of-Sight (CoS) sequence that mirrors how humans reason from images: find an object in 2D, then infer its distance, size, and pose. The decoder first emits 2D detections as a visual chain-of-thought, then predicts 3D boxes under an easy-to-hard curriculum: across objects, a near-to-far order reduces early ambiguity and matches ego-centric utility; within each object, a center-from-camera, dimensions, and rotation factorization ranks information by stability and learnability. This VLM-native interface preserves open-vocabulary and visual-prompting capabilities without specialized heads. On the challenging Omni3D benchmark, our model achieves state-of-the-art results, with* ***38.90*** $\mathrm{AP}_{\mathrm{3D}}$*, surpassing the previous best by* ***+13.98*** *absolute improvement even when the baseline is given ground-truth 2D boxes. It also generalizes zero-shot to held-out categories with strong robustness. By turning 3D detection into a disciplined next-token problem, LocateAnything3D offers a practical foundation for models to perceive in 3D.*

[*]Work done during an internship at NVIDIA.
[†]Equal advising.

1. Introduction

Vision-language models (VLMs) have rapidly advanced open-ended perception in 2D: with a single model and a single decoding interface, they localize, describe, and reason about arbitrary image content across diverse domains [3, 43, 107]. Yet one capability has lagged behind: general, multi-object 3D detection directly from monocular images. Existing monocular 3D detectors perform well within narrow domains [1, 26, 105], but rely on task-specific heads, closed label spaces, and carefully calibrated cameras; they do not inherit the versatility, compositionality, or instruction-following behavior that makes VLMs compelling. Recent work begins to bridge the gap by either coupling specialized 3D heads to open-vocabulary 2D detectors [105, 116], or by prompting foundation models with auxiliary geometric inputs, but they mostly address single-object grounding or require customized modules that break the simplicity of the VLM paradigm [18]. In short, we still lack a VLM that can *natively* perceive 3D and produce reliable, multi-object 3D boxes from a single image.

The strong motivation behind teaching VLMs to reason about 3D lies in the next frontier of the embodied intelligence: not just perception, but action. 3D boxes are a compact, metrically meaningful scene state: they connect recognition to interaction, make supervision verifiable, and enable calibration in diverse environments. Folding this capability into the same, token-based interface that already handles 2D grounding simplifies system design and makes scaling with data straightforward. The question we pursue is focused: *what is the most VLM-native recipe that makes multi-object monocular 3D detection just work*?

We answer this question with *Chain-of-Sight (CoS)*, a decoding and supervision scheme that teaches 3D the way humans often reason from pictures: first commit to what is visible in 2D, then infer distance, size, and pose [64, 75, 90]. Specifically, we cast detection as a short token sequence that interleaves 2D and 3D per instance: the decoder emits a 2D box, then the corresponding 3D box, and repeats until an end-of-sequence token. The explicit 2D step serves as a high-confidence visual chain-of-thought (CoT) that focuses the search on the right pixels, ties subsequent tokens to verifiable evidence, and reduces hallucination. In an autoregressive model, this is not merely convenient formatting: early tokens should be easy, highly informative, and attributable. Committing to image space first provides strong conditioning for the rest of the sequence, shapes the likelihood landscape to be smoother for 3D tokens, and yields a natural interface for prompting. Because the same decoder accepts either text or visual cues, a user can supply text instruction or a box/click, and the model continues with the 3D state for that instance without switching heads or losses.

Beyond the 2D proxy, we align supervision to the natural curriculum of autoregressive decoding. *Across objects*, we serialize detections by depth, from near to far. This ordering matches ego-centric utility (near objects matter first), provides high-evidence tokens early, and sets geometric context that constrains scale and distance for later objects via relative size and occlusion. Placing ambiguous, far instances at the tail prevents them from derailing the prefix. *Within each object*, we factorize the 3D box into a semantically ordered tuple and decode *center* $\rightarrow$ *size* $\rightarrow$ *rotation*. This ranking mirrors the observability of monocular cues: "where is it?" before "how big is it?" before "how is it oriented?", and stabilizes learning by letting location constrain the latter properties. Compared to corner-based encodings that entangle all parameters and amplify early errors, this factorization is both more learnable and better calibrated.

To train CoS end-to-end, we curate a camera-centric corpus that presents supervision in exactly the sequence the model will decode: 2D $\rightarrow$ 3D and near $\rightarrow$ far. We unify heterogeneous data sources into a shared schema, retain intrinsics and a consistent camera-frame parameterization, and convert the data into VLM conversations with calibrated negatives for anti-hallucination. The resulting package is a high-quality dataset that comprises approximately 1.74M training examples spanning indoor and outdoor scenes and diverse camera rigs for 3D vision-language perception.

The results demonstrate the power of our 2D-as-proxy and easy-to-hard curricula. On the challenging Omni3D dataset [5], our method attains state-of-the-art performance with **38.90** $\mathrm{AP}_{\mathrm{3D}}$, surpassing the previous best by **+13.98** absolute points. The same model shows strong zero-shot generalization to held-out categories. Ablations corroborate the design: replacing near-to-far with a left-to-right scanline or random ordering drops performance by a large margin. Removing the 2D CoS also collapses accuracy. Qualitative results (Figs. 1 and 3) show depth-consistent ordering, scale stability across repeated objects, and coherent orientations under occlusion and truncation.

This paper makes three contributions:

- A **Chain-of-Sight** formulation that turns open-world monocular 3D detection into a native next-token prediction problem in a VLM. By coupling explicit 2D grounding with 3D decoding, CoS improves reliability while preserving text- or visual-prompting within one interface.
- A **curriculum and representation** tailored to autoregressive decoding: near$\rightarrow$far serialization across objects and an intra-object tokenization that yields consistent decoding, stronger performance, and robustness under camera and category shifts.
- A **camera-centric dataset** that unifies heterogeneous data sources into CoS-ready corpus, enabling scalable and systematic ablations without task-specific heads.

These elements deliver simple and strong 3D perception within a VLM, closing a long-standing gap between open-vocabulary recognition and metric 3D understanding.

2. Related Work

We situate our contributions at the intersection of three converging directions: VLMs for visual perception that couple recognition with fine-grained grounding; embodied vision beyond static understanding toward spatial reasoning and acting; and 3D object detection from closed-set training to unified open-vocabulary formulations.

2.1. Vision-Language Models in Visual Perception

The 2D perception task provides a perfect playground for vision-language models (VLMs) to learn localizing and reasoning. Classical pipelines rely on specialized "vision experts" for recognition and grounding, including contrastive pretraining for open-set recognition and matching [71], image-text pretraining for retrieval [40], and strong detectors for region-level grounding [37, 52, 73]. Building atop these capabilities, recent VLM systems either orchestrate experts as tools under a multimodal controller [51, 97] or pursue unified backbones that natively tackle a wide spectrum of perception tasks [48, 86, 87, 89, 93, 99, 107]. Within grounding, the long-standing line of referring-expression comprehension (REC) frames 2D localization from unstructured language, with RefCOCO/+/g, Flickr30k-Entities, and subsequent datasets pushing object-level grounding in everyday scenes [34, 50, 63, 68, 94, 108]. Recent VLMs further include grounding into training objectives, learning to output boxes or points directly – *e.g.*, bounding-box supervision in Kosmos-2, Qwen-VL, and Gemini [3, 25, 33, 67], and point-based localization in MoLMo and RoboPoint [20, 110]. Beyond static REC, task-conditioned and temporally aware grounding has emerged as a key frontier for embodied use, including task-driven pointing and procedure-aware grounding [100], and 2D grounding as reasoning chain-of-thought [62, 78]. Our formulation adopts this 2D-first perspective: we leverage 2D detections as explicit intermediate tokens to structure perception before lifting to full 3D inference, aligning with evidence that tightly coupling grounding with reasoning yields more reliable downstream behavior [3, 43].

2.2. Vision-Language Models for Embodied Vision

Foundation models are increasingly leveraged as embodied agents that perceive, reason, and act in long-horizon tasks. Early systems primarily rely on prompting to elicit planning behaviors from VLM backbones [27, 35, 79, 81, 82], with code and API-centric tool interfaces further improving reliability [44, 80]. Subsequent work introduces supervised finetuning, yielding compact yet capable agents for manipulation [28, 36, 38, 49, 57, 113, 118, 121] and household or procedural reasoning [14, 32, 98]. In parallel, spatial intelligence has emerged as essential competencies for open-world embodiment, with lines of work targeting distance/metric understanding and counting [7, 10, 22, 23, 46, 83, 102, 125], as well as benchmark suites that synthesize complex 3D scenes and tasks [16, 72]. Embodied pointing and grounding further connect perception to action [26, 39, 110], and recent efforts augment spatial reasoning with structured reasoning [54, 111]. Integrated frameworks exemplify the trend toward generalist embodied VLMs and unify perception, reasoning, and planning at scale [31, 85], while curated benchmarks continue to expand the supervision landscape [15, 17, 21, 60, 65, 69, 70, 88, 103, 104, 109]. Beyond supervision fine-tuning (SFT), reinforcement-driven training also starts to revolutionize the role of reasoning traces in embodiment [12, 101, 112]. Our work is synergistic with these directions: we target *multi-object 3D perception* as a VLM-native next-token problem by structuring 3D detection with intermediate 2D reasoning and curriculum design. Our formulation provides an explicit, language-aligned perception interface that can plug into embodied agents.

2.3. 3D Object Detection

Classical monocular 3D object detection has been driven by single-dataset optimization on benchmarks, yielding strong in-domain performance with task-specific architectures but limited robustness under distribution [6, 11, 13, 24, 29, 41, 55, 91, 92, 120, 122, 123]. Parallel lines of work study multi-sensor fusion and spatio-temporal reasoning to boost accuracy, yet typically inherit closed-set label space constraints [45, 47, 61]. To reduce dataset and camera bias, Omni3D unifies diverse sources and introduces Cube R-CNN, showing that multi-dataset training improves cross-scene generalization for monocular detectors [5]. Subsequent efforts further explore bird's-eye-view formulations across indoor and outdoor settings [30, 42]. Moving beyond closed vocabularies, open-vocabulary 3D detection seeks to recognize and localize categories beyond those seen during training. Much of the early progress assumes point clouds as input or supervision [8, 9, 58, 59, 66, 77, 96, 115, 117, 119, 124, 126]. Closer to our setting, OVMono3D lifts open-vocabulary 2D detections (*e.g.*, from Grounding DINO) into 3D with a unified head [53, 106]. DetAny3D proposes a promptable 3D foundation model that transfers knowledge from 2D foundation models to monocular 3D via feature aggregation [116]. We pursue a VLM-native decoding that treats multi-object 3D detection as disciplined next-token inference, leveraging explicit 2D-to-3D factorization to improve generalization across categories and camera configurations.

3. LocateAnything3D Methodology

Overview. We study the monocular, open-world 3D detection task in a VLM-native setting, as demonstrated in our architecture diagram in Fig. 2. A single RGB image

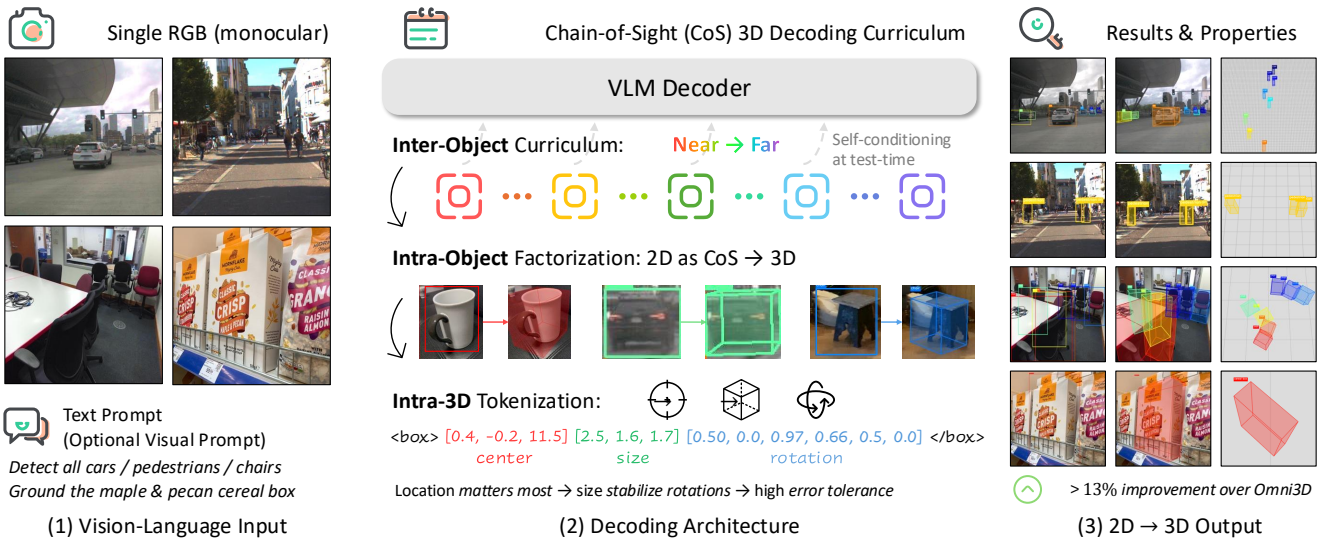

Figure 2. **Architecture of LocateAnything3D.** (1) Model input: a single RGB image with text and optional visual prompts (boxes/clicks). (2) Chain-of-Sight (CoS) decoding: a VLM decoder first emits 2D detections as an explicit visual evidence, then continues the sequence to 3D. Decoding follows three layers of design: inter-object curriculum ordering detections from near to far; intra-object factorization using 2D as CoS to robustly infer 3D; and intra-3D tokenization that outputs center, size, and rotation. (3) We output calibrated multi-object 3D boxes with open-vocabulary categories and flexible prompting, yielding strong results on Omni3D. We use turbo colormap for boxes to demonstrate their depth, where reddish and blueish colors indicate closer and farther objects, respectively.

and free-form text query drive an autoregressive (AR) decoder that emits a short, structured sequence comprising 2D proposals and their 3D counterparts. The core idea is our *Chain-of-Sight (CoS)* factorization, which makes 3D a native next-token prediction problem.

3.1. Preliminaries: Monocular 3D Detection

Let $I \in \mathbb{R}^{H\times W\times 3}$ be a monocular RGB image and let $c \in \Sigma^*$ denote a free-form textual description of a target category (*e.g.*, "car," "any cup," or "red chair"). The goal is to predict a variable-sized set of 3D bounding boxes of that category, denoted as $\mathcal{B}_c = \{\mathbf{b}_i\}_{i=1}^{N_c}$. We represent a 3D box in the camera coordinate frame as

$$\mathbf{b}_i = (\mathbf{t}_i, \mathbf{d}_i, \mathbf{R}_i) \quad \mathbf{t}_i \in \mathbb{R}^3; \mathbf{d}_i \in \mathbb{R}^3_+; \mathbf{R}_i \in \mathrm{SO}(3), \quad (1)$$

where $\mathbf{t}_i = (X_i, Y_i, Z_i)^\top$ is the 3D center from the camera, $\mathbf{d}_i = (W_i, H_i, L_i)^\top$ denotes the metric dimensions, and $\mathbf{R}_i$ is the object rotation. In scenes that admit the upright-world assumption (*e.g.*, autonomous driving), $\mathbf{R}_i$ can be parameterized by a single yaw angle; our formulation remains valid for the general case.

Given I and c, monocular 3D detection can be posed as set inference $\hat{\mathcal{B}}_c = \arg\max_{\mathcal{B}} P(\mathcal{B}|I, c)$ where $P(\mathcal{B}|I, c)$ is the conditional distribution of all 3D boxes of the queried category. With an autoregressive decoder, a standard factorization of the Eq. 1 is

$$P(\mathcal{B}|I, c) = \prod_{i=1}^{N_c} P(\mathbf{b}_i \mid I, c, \mathbf{b}_{<i}), \quad (2)$$

where $\mathbf{b}_{<i}$ denotes previously generated boxes and an end-of-sequence token handles the unknown cardinality N_c.

For later use, we also define the 2D bounding box of instance i as $\mathbf{q}_i = (x_i^{\min}, y_i^{\min}, x_i^{\max}, y_i^{\max}) \in \{0, 1, \ldots, 1000\}^4$ in normalized integer image coordinates, and a (known or estimated) pinhole projection operator Π mapping $(\mathbf{t}_i, \mathbf{d}_i, \mathbf{R}_i)$ to image space. We write $\Pi(\mathbf{b}_i) \Rightarrow \mathbf{q}_i$ when the 2D box is obtained by projecting the 3D cuboid.

3.2. Chain-of-Sight (CoS) Factorization

The key innovation is to interleave 2D and 3D in the token sequence so that 2D localization acts as a visual chain-of-thought, which we call chain-of-sight (CoS), that constrains 3D inference. Concretely, the decoder emits

$$\mathcal{S} = (\mathbf{q}_1, \mathbf{b}_1, \mathbf{q}_2, \mathbf{b}_2, \ldots, \langle\mathrm{eos}\rangle), \quad (3)$$

where each 2D box $\mathbf{q}_i$ is immediately followed by its 3D counterpart $\mathbf{b}_i$. The resulting conditional probability decomposes as

$$P(\mathcal{S} \mid I, c) = \prod_{i=1}^{N_c} \underbrace{P(\mathbf{q}_i \mid I, c, \mathcal{S}_{<i})}_{\text{2D localization}} \underbrace{P(\mathbf{b}_i \mid I, c, \mathcal{S}_{<i}, \mathbf{q}_i)}_{\text{3D estimation}} \cdot P(\langle\mathrm{eos}\rangle \mid I, c, \mathcal{S}_{\le N_c}), \quad (4)$$

where $\mathcal{S}_{<i}$ denotes all tokens emitted before step i. Compared to Eq. (2), this CoS factorization introduces a high-confident intermediate $\mathbf{q}_i$ that: (i) focuses the search on the right pixels, (ii) reduces hallucination by tying 3D tokens

to visible evidence, and (iii) aligns naturally with AR decoding, where early tokens should be both easy and highly informative. By committing to $\mathbf{q}_i$ first, the model learns to ground each instance before decoding its 3D state, mirroring how textual chain-of-thought stabilizes hard reasoning.

Inter-object curriculum. Conventional 2D detectors often impose a scanline or left-to-right ordering when serializing detections for AR decoders [43, 107]. Such policies are agnostic to 3D geometry: two boxes that are adjacent in 2D may be at very different depths. Hence, far-away instances that are intrinsically ambiguous in monocular views can appear adjacent and early in the sequence and derail subsequent decoding. We therefore adopt a *near-to-far* curriculum across objects. Placing nearer objects first improves three properties relevant to 3D: (1) *utility*: nearer instances matter most for interaction and safety; (2) *evidence quality*: close objects provide stronger monocular cues, yielding confident early tokens; and (3) *context*: once nearby geometry is established, it constrains the plausible size and depth of distant objects via relative scale and occlusion relationships. In practice, the depth-aware order leads to more stable, well-calibrated sequences than 2D scanline order.

Intra-object factorization (2D $\Rightarrow$ 3D). Although 3D detection does not *require* predicting 2D boxes, we deliberately ask the model to do so. Prior monocular methods commonly rely on an *external* 2D detector to propose boxes and then lift them to 3D with a specialized head [53, 106, 116]. In contrast, our VLM performs both 2D localization and 3D estimation *within the same decoder and the same interface*. This tight coupling is beneficial for the same reason textual chain-of-thought helps language problems: intermediate commitments break a hard prediction into easier, verifiable steps. Here, the 2D prediction serves as a *visual CoT* – our Chain-of-Sight – that anchors subsequent 3D tokens. The design also naturally supports visual prompting: when a user supplies a 2D cue (*e.g.*, a box or a click), the decoder can immediately continue with the corresponding 3D tokens for that instance, preserving the AR workflow.

Intra-3D tokenization. A 3D box can be represented in several ways. Corner-based encodings list eight projected or 3D vertices [5], but they are ambiguous to an AR decoder (which corner comes first?), and amplify early-token errors. Instead, we adopt the structured representation of Eq. (1) and, crucially, a *semantic ordering* for AR decoding: center $\mathbf{t} \rightarrow$ size $\mathbf{d} \rightarrow$ rotation $\mathbf{R}$. This order reflects information value and difficulty: "where is it?" before "how big is it?" before "how is it oriented?", and we find it substantially improves the robustness.

Coordinate and rotation systems. We predict boxes in the *camera* frame rather than the world frame. This avoids burdening the model with estimating scene-level coordinates (which vary across datasets and camera rigs) and improves cross-domain generalization. Projection to image space uses the usual pinhole model, $\Pi : (\mathbf{t}, \mathbf{d}, \mathbf{R}) \mapsto \mathbf{q}$, with intrinsics known or estimated. For rotation, our formulation supports either a full $\mathrm{SO}(3)$ rotation or a yaw-dominant parameterization when the upright assumption is reasonable (*e.g.*, driving scenes). The latter allocates most capacity to the most observable angle under monocular cues while retaining the general case when needed. Overall, the CoS factorization (Eq. 4), together with a near-to-far inter-object curriculum and center$\rightarrow$size$\rightarrow$rotation intra-object ordering, turns open-world monocular 3D detection into a compact sequence that is easy for a VLM to learn and robust to decode, all within a single, unified interface. Training uses standard cross-entropy losses over tokens; additional details follow in subsequent sections.

4. LocateAnything3D Data Curation at Scale

Goal. We construct a large, camera-centric corpus that natively supports our Chain-of-Sight decoding (Fig. 2) and the formulation in Sec. 3. The data are presented to the model exactly in the sequence it will decode at test time: first 2D, then 3D, and from near to far. We unify heterogeneous monocular 3D benchmarks into a single representation and package them as VLM conversations for both single-object grounding and multi-object detection.

Datasets and Unification. We leverage six public 3D detection datasets: ARKitScenes [4], SUN-RGBD [84], Hypersim [74], Objectron [2], KITTI [24], and nuScenes [6] into a shared JSONL format. Across datasets we retain camera intrinsics and adopt a camera-coordinate convention for 3D boxes to maximize cross-domain transfer.

4.1. Stage I: Canonical Multi-Box Normalization

Output unit. For each image and category we create one JSONL line containing all instances of that category, ordered by depth. Formally, each line corresponds to a tuple $(\texttt{image\_path}, \texttt{category\_name})$ and carries a list of per-instance fields aligned by index.

Geometry-based filtering and quality control. We drop instances that are behind the camera or entirely outside the image frustum relative to the camera frame. When the dataset provides visibility and truncation metadata, we keep items with visibility greater than 0.16 and truncation less than 0.84; otherwise we approximate these terms using 2D projections, depth ordering, and border intersection. These thresholds balance coverage and precision, removing ambiguous supervision that is particularly harmful early in autoregressive decoding.

2D and 3D representations. To represent 2D objects, we store both tight pixel boxes $\mathbf{q} = (x^{\min}, y^{\min}, x^{\max}, y^{\max})$ and normalized coordinates in $[0, 1000]$ (integers). The 2D representation can also be conveniently converted to

center point format for prompting variants (*e.g.*, points). For 3D reprensetation, we keep multiple redundant parameterizations to support ablations and alternative supervision choices for each instance: (i) the center in camera coordinates $\mathbf{t} = (X, Y, Z)$ (meters); (ii) dimensions $\mathbf{d} = (W, H, L)$ (meters); and (iii) rotation as a 3×3 matrix $\mathbf{R}$, Euler angles (ZYX) rescaled to $[0, 1]$, and their elementwise sine/cosine (mapped from $[-1, 1]$ to $[0, 1]$). Numeric fields are rounded to two decimals (zero preserved) to control entropy while retaining salient signal. Each line also stores image width/height and the intrinsic matrix $\mathbf{K}$. Within each grouped line we sort by increasing depth of the 3D center from the camera. This stage yields approximately **480K** single-image, multi-object training entries.

4.2. Large-Scale Text Auto-Annotation

To supply rich referring expressions without manual labeling, we prompt strong VLMs [33, 89] on images where exactly one target instance is highlighted at a time (a single tight 2D box overlay; the scene remains otherwise untouched). Prompts ask for concise, *unambiguous* descriptions that uniquely identify the target using semantic attributes, spatial layout (left/right/top/bottom; nearby objects), coarse pose, and contextual anchors.

We generate three paraphrases per target with mild sampling for lexical diversity, then conduct automated uniqueness checks: (i) contrastive A/B re-rendering on another instance of the same category; (ii) candidate-index selection tests; and (iii) rejection of hedged or unverifiable language. The resulting corpus contains $\sim$**1.0M** high-quality single-object grounding samples.

4.3. Negative Samples for Anti-Hallucination

We explicitly supervise *no-match* behavior. For each image we know the exact set of present categories from the canonical lines. We sample absent categories, including hard negatives chosen via semantic proximity (*e.g.*, car and van), and produce queries that should yield no detections. Negatives are capped at 10% of training examples (at most 2 per training image), so positives dominate while every batch carries calibrated rejection pressure. Packaging is identical to positives except the model must emit a sentinel token `<no_object/>`. This simple design significantly reduces false positives without harming recall.

4.4. Stage II: Packaging for VLM Training

We convert the canonical JSONL into conversational samples suitable for an autoregressive decoder.

Conversation record. Each example has a unique `id`, an `image` pointer, and a two-turn dialogue: a human prompt and a model response. The response concatenates one or more instance segments, each containing a 2D box immediately followed by its 3D counterpart (mirroring CoS). Multi-object examples preserve the near-to-far order inherited from Stage I.

Scale and generalization. The same processing applies to all datasets. The unified schema allows us to scale training without dataset-specific heads, and enables consistent ablations on ordering, representation choices, and instruction phrasing across all sources. Combining normalized detection entries, single-object grounding, and calibrated negatives yields approximately **1.74M** training conversations spanning diverse categories, camera rigs, and scene types, which will be made publicly available.

5. Experiments

Benchmarks and metrics. We evaluate on Omni3D [5], a large-scale monocular 3D detection suite covering both indoor and outdoor imagery. Omni3D provides official trainval and test splits. The test set is held out strictly: no images or labels from test are used during training or hyperparameter tuning. For evaluation metrics, unless otherwise stated, we adopt the benchmark metrics used in Omni3D. Reported scores are 3D Average Precision ($\mathrm{AP}_{\mathrm{3D}}$) computed over a sweep of 3D IoU thresholds ($\tau \in \{0.05, 0.10, \ldots, 0.50\}$). Intersections are measured volumetrically in the camera frame, consistent with Sec. 3. All evaluations follow a *target-aware* protocol, as advocated in prior open-vocabulary work [105, 116]: for each image, the detector is prompted only with the categories that actually occur in its annotations rather than an exhaustive vocabulary. This simple change alleviates naming inconsistencies and focuses the comparison on 3D localization quality rather than on taxonomy alignment.

Baselines. We compare against methods that are most compatible with our open-world, prompt-driven setup: (1) Cube R-CNN [5]: the reference baseline released with Omni3D, a unified detector trained as a close-vocabulary model. (2) OVMono3D [105]: an open-vocabulary monocular 3D detector tailored to Omni3D. It "lifts" 2D detections to 3D by wiring an open-vocabulary 2D localizer [52] to a 3D prediction head. (3) DetAny3D [116]: a promptable monocular 3D detector that accepts category text and outputs 3D boxes directly, designed for open-world settings.

Pretraining of 2D detection and grounding. Before training the full Chain-of-Sight model, we conduct a 2D detection and grounding pretraining phase to equip the model with strong 2D localization capabilities. This stage focuses exclusively on predicting 2D bounding boxes from text or visual prompts, establishing a robust foundation for the subsequent 2D-to-3D learning. After pretraining, we train the complete CoS sequence (2D $\rightarrow$ 3D) end-to-end using a standard cross-entropy loss over the autoregressive token sequence. Additional training details, hyperparameters, and ablations are provided in the supplementary material.

Table 1. **3D detection on the Omni3D benchmark.** Our LocateAnything3D achieves state-of-the-art results over all baselines, even outperforming DetAny3D with additional ground-truth 2D inputs on metrics. The first three columns (Omni3D_OUT) show outdoor-only results, while the remaining columns show results on the full unified dataset spanning indoor and outdoor scenes.

Method	Omni3D_OUT			Omni3D						
	$AP^{kit}_{3D}\uparrow$	$AP^{nus}_{3D}\uparrow$	$AP^{out}_{3D}\uparrow$	$AP^{kit}_{3D}\uparrow$	$AP^{nus}_{3D}\uparrow$	$AP^{sun}_{3D}\uparrow$	$AP^{ark}_{3D}\uparrow$	$AP^{obj}_{3D}\uparrow$	$AP^{hyp}_{3D}\uparrow$	$AP_{3D}\uparrow$
ImVoxelNet [76]	23.5	23.4	21.5	-	-	-	-	-	-	9.4
SMOKE [55]	25.9	20.4	20.0	-	-	-	-	-	-	10.4
OV-Uni3DETR [95]	35.1	33.0	31.6	-	-	-	-	-	-	-
Cube R-CNN [5]	36.0	32.7	31.9	32.50	30.06	15.33	41.73	50.84	7.48	23.26
OVMono3D [105]	-	-	-	25.45	24.33	15.20	41.60	58.87	7.75	22.98
DetAny3D [116]	35.8	33.9	32.2	31.61	30.97	18.96	46.13	54.42	7.17	24.92
DetAny3D$_{w/\text{ Ground-Truth 2D Box}}$	38.0	**36.7**	35.9	38.68	**37.55**	**46.14**	50.62	56.82	15.98	34.38
LocateAnything3D (ours)	**39.8**	33.9	**36.1**	**43.75**	35.26	45.12	**59.89**	**71.90**	**18.12**	**38.90**

Table 2. LocateAnything3D achieves the best zero-shot 3D detection performance, demonstrating strong generalization to unseen object classes. Notably, baseline methods rely on an external detector for 2D box as an additional input, while our method jointly predicts both 2D and 3D boxes end-to-end from a single image alone. Following existing methods, we report AP_{3D} using the target-aware metric (per-image existing categories for prompting).

Method	Novel Categories		
	AP^{kit}_{3D}	AP^{sun}_{3D}	AP^{ark}_{3D}
OVMono3D$_{w/\text{ Grounding-DINO 2D Boxes}}$	4.71	16.78	13.21
DetAny3D$_{w/\text{ Grounding-DINO 2D Boxes}}$	25.73	21.07	24.56
LocateAnything3D$_{\text{(single image, no external 2D)}}$	**25.87**	**26.33**	**29.06**
Δ vs. DetAny3D	+0.14	+5.26	+4.50

Implementation details. Our work is built on SigLIP vision encoder [114] and Qwen2-8B backbone [3] coupled by a lightweight MLP projector. Images are decomposed into up to 12 adaptive tiles plus a global thumbnail, each with 448 pixel size, and the resulting visual tokens replace repeated `<IMG_CONTEXT>` tokens in a Qwen2-style chat template. We train with bfloat16 and FlashAttention 2 [19] for both vision and language, apply dynamic online packing to fill a 16,384-token context per sample, and optimize with AdamW [56] and a learning rate of 1e-5, a weight decay of 0.05, a cosine scheduler, and a 3% warm-up, under ZeRO-3 with gradient checkpointing. Training uses 64 H100 GPUs for 46 hours over 37K steps. Please refer to the supplementary material for complete implementation details.

5.1. Main Performance

Overall evaluation and protocal. Table 1 summarizes results on the Omni3D benchmark. Our **LocateAnything3D** attains the best score on every metric and every split. On the outdoor-only training/evaluation track (`Omni3D_OUT`), our method reaches **36.1** AP_{3D}, overperforming DetAny3D (32.2) and DetAny3D aided by *ground-truth* 2D boxes (35.9). When trained and evaluated on the full unified indoor and outdoor dataset, we also markedly lead across all domains against existing methods, and we even outperform previous method with ground-truth 2D visual prompts in 5 out of 7 metrics. Our overall mean AP_{3D} reaches **38.90**, surpassing the prior best with privileged 2D boxes by **+4.52**. These improvements reflect the benefits of CoS decoding: accurate 2D grounding as a first-class step simplifies monocular 3D inference without relying on any auxiliary detectors or oracle boxes.

Following Omni3D, the `Omni3D_OUT` setting trains solely on outdoor driving data (KITTI+nuScenes) and reports per-domain and aggregated outdoor metrics; the full Omni3D setting trains on the entire corpus and evaluates across both indoor and outdoor domains. Importantly, all training excludes the official `test` images and labels. Notably, even methods that receive *ground-truth* 2D boxes at inference lag behind our end-to-end approach, which highlights that learning 2D and 3D jointly within a single autoregressive interface is more effective than bolting a 3D head onto externally supplied 2D proposals.

Zero-shot novel categories. We follow the evaluation protocol used by prior open-vocabulary methods [105, 116]: images are prompted only with the categories present in their annotations, and the held-out classes are never seen during training. As shown in Table 2, **LocateAnything3D** delivers the strongest zero-shot performance on all benchmarks, with **25.87** on KITTI novel classes [24], **26.33** on SUN-RGBD [84], and **29.06** on ARKitScenes [4], while competing approaches depend on an external 2D detector (Grounding DINO [52]) to supply proposals. Relative to DetAny3D+2D, we gain +0.14, +5.26, and +4.50 points on the three metrics, respectively; compared to OVMono3D+2D, our margins widen further. These results support our motivation that predicting 2D and 3D *together* – rather than lifting from external 2D – improves transfer to unseen categories.

5.2. Analysis and Ablations

In this section, we conduct ablation study to verify the design choices of our Chain-of-Sight learning paradigm.

Inter-object ordering. Replacing our depth-aware near $\rightarrow$ far order with common alternatives degrades quality (Ta-

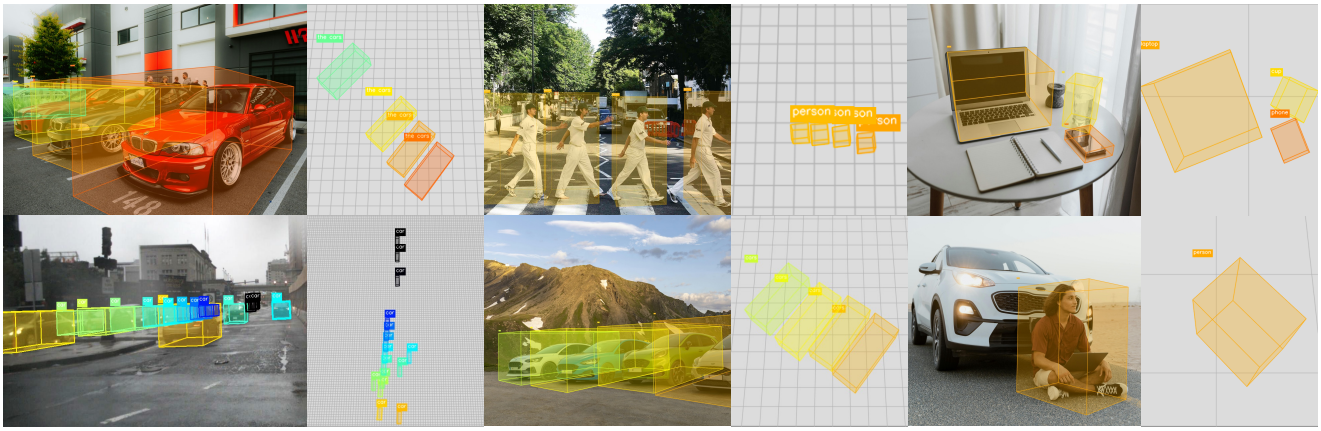

Figure 3. **Qualitative results of LocateAnything3D.** For each example, the left sub-figure overlays the projected 3D bounding boxes on the input image, while the right sub-figure shows the corresponding bird's-eye view with 1m×1m grids as the background. We use a turbo colormap based on depth, where redish colors indicate objects closer to the camera, and blueish colors indicate objects farther away.

Table 3. **Ablation study of Chain-of-Sight (CoS) design choices.** We evaluate each component of our three-layer decoding design on Omni3D_OUT. All results are reported using AP_{3D}^{out}. Our full design (highlighted) achieves the best performance, validating the importance of each design choice.

Design Component	Variant	AP_{3D}^{out} ↑
Inter-Object Curriculum	Random Ordering	17.5
	Left-to-Right Ordering	30.6
	Near-to-Far Ordering	**33.1**
Intra-Object Factorization	No 2D (Direct 3D)	22.7
	3D-then-2D	26.2
	2D-then-3D (CoS)	**33.1**
Intra-3D Tokenization	Rotation-Size-Center	28.8
	Center-Rotation-Size	32.9
	Center-Size-Rotation	**33.1**

ble 3). A random order performs worst (17.5), confirming that the sequence position carries semantic information in AR decoding. A left-to-right scanline policy is better (30.6) but still inferior to our near-to-far curriculum (**33.1**), indicating that 3D-aware serialization (high-evidence instances first) yields more stable and informative token prefixes.

Intra-object factorization. Removing the 2D step and predicting 3D directly drops performance to 22.7. Emitting 3D before 2D ("3D-then-2D") recovers some accuracy (26.2) but remains far from our CoS layout (**33.1**). These trends validate the role of 2D as a visual chain-of-thought: committing to image-space evidence makes the subsequent 3D tokens both easier to learn and better calibrated. And it shows that 2D is a helpful signal to learn, especially when predicted before the 3D signal.

Intra-3D token order. Within each object, decoding with the center, size, and rotation ordering performs the best. Switching to Rotation-Size-Center harms results (28.8), and Center-Rotation-Size is slightly worse (32.9), suggesting that anchoring location, then scale, before resolving orientation is the most learnable and robust schedule for monocular cues. The small but consistent gap between CSR and CRS further indicates that deferring rotation until after size stabilizes the pose estimate.

5.3. Qualitative Results

Figure 3 showcases representative predictions of LocateAnything3D. In each example, the left panel overlays the projected 3D cuboids on the RGB frame, while the right panel renders a bird's-eye-view with 1m×1m grids. The turbo colormap encodes depth, revealing a depth-consistent ordering that mirrors our CoS decoding: near instances are resolved first and anchor the subsequent geometry. The model handles moderate occlusion and truncation, maintains scale consistency across repeated objects, and preserves orientation structure even at distance.

Supplementary material. In the supplementary material we include more implementation details, extended analyses, limitations/failure cases, and broader impact.

6. Conclusion

We present LocateAnything3D, a VLM-native framework that turns monocular 3D detection into a concise next-token task via *Chain-of-Sight* decoding. By committing to 2D localization before 3D, ordering objects near-to-far, and factorizing each box as center, size, and rotation, our approach aligns supervision with the natural curriculum of autoregressive models. Coupled with a CoS-conformant, multi-domain corpus and a simple training recipe, the method delivers state-of-the-art results on Omni3D, both in-domain and zero-shot to novel categories. We hope that our CoS principle provides a practical route for scaling 3D perception within general-purpose VLMs and opens the door to future extensions in video, multi-view reasoning, and embodied planning.

Acknowledgments. Y. Man is supported by the NVIDIA Graduate Fellowship. Y. Man and Y.-X. Wang are supported in part by NSF under Grants 2106825 and 2519216, the DARPA Young Faculty Award, the ONR Grant N00014-26-1-2099, and the NIFA Award 2020-67021-32799.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 2

[2] 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

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and *et al.* Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 7

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 5, 7

[5] 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, 3, 5, 6, 7

[6] 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. 3, 5

[7] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 3

[8] Yang Cao, Zeng Yihan, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *NeurIPS*, 2024. 3

[9] Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3d object detection. *arXiv preprint arXiv:2406.00830*, 2024. 3

[10] 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. 3

[11] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. *arXiv preprint arXiv:2103.12605*, 2021. 3

[12] Hanyang Chen, Mark Zhao, Rui Yang, Qinwei Ma, Ke Yang, Jiarui Yao, Kangrui Wang, Hao Bai, Zhenhailong Wang, Rui Pan, Mengchao Zhang, Jose Barreiros, Aykut Onol, ChengXiang Zhai, Heng Ji, Manling Li, Huan Zhang, and Tong Zhang. ERA: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning. *arXiv preprint arXiv:2510.12693*, 2025. 3

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

[14] Yaran Chen, Wenbo Cui, Yuanwen Chen, Mining Tan, Xinyao Zhang, Dongbin Zhao, and He Wang. RoboGPT: an intelligent agent of making embodied long-term decisions for daily instruction tasks. *arXiv preprint arXiv:2311.15649*, 2023. 3

[15] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*, 2023. 3

[16] 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 model. *arXiv preprint arXiv:2406.01584*, 2024. 3

[17] Long Cheng, Jiafei Duan, Yi Ru Wang, Haoquan Fang, Boyang Li, Yushan Huang, Elvis Wang, Ainaz Eftekhar, Jason Lee, Wentao Yuan, Rose Hendrix, Noah A. Smith, Fei Xia, Dieter Fox, and Ranjay Krishna. Pointarena: Probing multimodal grounding through language-guided pointing. *arXiv preprint arXiv:2505.09990*, 2025. 3

[18] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, and Marco Pavone. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024. 2

[19] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359, 2022. 7

[20] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, and *et al.* Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024. 3

[21] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, 2025. 3

[22] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spa-

tial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. 3

[23] 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. 3

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 3, 5, 7

[25] Google DeepMind. Introducing gemini 2.0: Our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024. Accessed 2025-04-28. 3

[26] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023. 2, 3

[27] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look Before You Leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023. 3

[28] Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025. 3

[29] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 3

[30] Jin-Cheng Jhang, Tao Tu, Fu-En Wang, Ke Zhang, Min Sun, and Cheng-Hao Kuo. V-MIND: Building versatile monocular indoor 3d detector with diverse 2d annotations. In *WACV*, 2025. 3

[31] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025. 3

[32] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. RoboBrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*, 2025. 3

[33] Koray Kavukcuoglu. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, 2025. Accessed 2025-04-28. 3, 6

[34] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 3

[35] Byeonghwi Kim, Jinyeon Kim, Yuyeong Kim, Cheolhong Min, and Jonghyun Choi. Context-aware planning and environment-aware memory for instruction following embodied agents. *arXiv preprint arXiv:2308.07241*, 2024. 3

[36] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[38] 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. 3

[39] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. TopViewRS: Vision-language models as top-view spatial reasoners. *arXiv preprint arXiv:2406.02537*, 2024. 3

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[41] 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. 3

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

[43] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025. 2, 3, 5

[44] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022. 3

[45] 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. In *NeurIPS*, 2022. 3

[46] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. *arXiv preprint arXiv:2409.09788*, 2024. 3

[47] 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. 3

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[49] Huaping Liu, Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, and Hanbo Zhang. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2025. 3

[50] Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*, 2024. 3

[51] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023. 3

[52] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3, 6, 7

[53] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3, 5

[54] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, Helong Huang, Guangjian Tian, Weichao Qiu, Xingyue Quan, Jianye Hao, and Yuzheng Zhuang. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 3

[55] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPRW*, 2020. 3, 7

[56] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[57] Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025. 3

[58] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. *arXiv preprint arXiv:2207.01987*, 2022. 3

[59] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *CVPR*, 2023. 3

[60] Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, Shenglong Ye, Lewei Lu, Jingbo Wang, Wenhai Wang, Jifeng Dai, Yu Qiao, Rongrong Ji, and Xizhou Zhu. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025. 3

[61] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. BEV-guided multi-modality fusion for driving perception. In *CVPR*, 2023. 3

[62] Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *CVPR*, 2025. 3

[63] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 3

[64] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2

[65] NVIDIA, Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, Jingyi Jin, Brendan Johnson, Rizwan Khan, George Kurian, Elena Lantz, Nayeon Lee, Zhaoshuo Li, Xuan Li, Tsung-Yi Lin, Yen-Chen Lin, Ming-Yu Liu, Andrew Mathau, Yun Ni, Lindsey Pavao, Wei Ping, David W. Romero, Misha Smelyanskiy, Shuran Song, Lyne Tchapmi, Andrew Z. Wang, Boxin Wang, Haoxiang Wang, Fangyin Wei, Jiashu Xu, Yao Xu, Xiaodong Yang, Zhuolin Yang, Xiaohui Zeng, and Zhe Zhang. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 3

[66] Xingyu Peng, Yan Bai, Chen Gao, Lirong Yang, Fei Xia, Beipeng Mu, Xiaofei Wang, and Si Liu. Global-local collaborative inference with llm for lidar-based open-vocabulary detection. In *ECCV*. Springer, 2025. 3

[67] 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. *arXiv preprint arXiv:2306.14824*, 2023. 3

[68] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 3

[69] Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024. 3

[70] Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, Maoqing Yao, Haoran Yang, Jiacheng Bao, Bin Zhao, and Dong Wang. Embodiedonevision: Interleaved vision-text-action pretraining for general robot control. *arXiv preprint arXiv:2508.21112*, 2025. 3

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[72] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 3

[73] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

[74] 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

[75] Irvin Rock. The logic of perception. 1983. 2

[76] 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. 7

[77] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[78] 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. In *NeurIPS*, 2024. 3

[79] Suyeon Shin, Sujin jeon, Junghyun Kim, Gi-Cheon Kang, and Byoung-Tak Zhang. Socratic Planner: Self-qa-based zero-shot planning for embodied instruction following. *arXiv preprint arXiv:2404.15190*, 2024. 3

[80] Tom Silver, Soham Dan, Kavitha Srinivas, Joshua B Tenenbaum, Leslie Kaelbling, and Michael Katz. Generalized planning in pddl domains with pretrained large language models. In *AAAI*, 2024. 3

[81] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*, 2022. 3

[82] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *ICCV*, 2023. 3

[83] 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. *arXiv preprint arXiv:2411.16537*, 2024. 3

[84] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5, 7

[85] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, Yingbo Tang, Xiangqi Xu, Wei Guo, Yaoxu Lyu, Yijie Xu, Jiayu Shi, Mengfei Du, Cheng Chi, Mengdi Zhao, Xiaoshuai Hao, Junkai Zhao, Xiaojie Zhang, Shanyu Rong, Huaihai Lyu, Zhengliang Cai, Yankai Fu, Ning Chen, Bolun Zhang, Lingfeng Zhang, Shuyi Zhang, Dong Liu, Xi Feng, Songjing Wang, Xiaodan Liu, Yance Jiao, Mengsi Lyu, Zhuo Chen, Chenrui He, Yulong Ao, Xue Sun, Zheqi He, Jingshu Zheng, Xi Yang, Donghai Shi, Kunchang Xie, Bochao Zhang, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 3

[86] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3

[87] Emu3 Team. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[88] Gemini Robotics Team. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 3

[89] OpenAI Team. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 6

[90] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*. L. Voss, 1867. 2

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

[92] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*, 2022. 3

[93] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3

[94] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *CVPR*, 2024. 3

[95] Zhenyu Wang, Yali Li, Xi Chen, Hengshuang Zhao, and Shengjin Wang. Uni3DETR: Unified 3d detection transformer. In *NeurIPS*, 2023. 7

[96] 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. 3

[97] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT:

Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[98] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*, 2023. 3

[99] 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. 3

[100] Haotian Xue, Yunhao Ge, Yu Zeng, Zhaoshuo Li, Ming-Yu Liu, Yongxin Chen, and Jiaojiao Fan. Point-It-Out: Benchmarking embodied reasoning for vision language models in multi-stage visual grounding. *arXiv preprint arXiv:2509.25794*, 2025. 3

[101] Ganlin Yang, Tianyi Zhang, Haoran Hao, Weiyun Wang, Yibin Liu, Dehui Wang, Guanzhou Chen, Zijian Cai, Junting Chen, Weijie Su, Wengang Zhou, Yu Qiao, Jifeng Dai, Jiangmiao Pang, Gen Luo, Wenhai Wang, Yao Mu, and Zhi Hou. Vlaser: Vision-language-action model with synergistic embodied reasoning. *arXiv preprint arXiv:2510.11027*, 2025. 3

[102] 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. *arXiv preprint arXiv:2412.14171*, 2024. 3

[103] 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. 3

[104] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 3

[105] Jin Yao, Hao Gu, Xuweiyi Chen, Jiayun Wang, and Zezhou Cheng. Open vocabulary monocular 3d object detection. *arXiv preprint arXiv:2411.16833*, 2024. 2, 6, 7

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

[107] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2023. 2, 3, 5

[108] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 3

[109] 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. *arXiv preprint arXiv:2406.10721*, 2024. 3

[110] 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. *arXiv preprint arXiv:2406.10721*, 2024. 3

[111] Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv preprint arXiv:2505.08548*, 2025. 3

[112] Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-R1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*, 2025. 3

[113] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. 3

[114] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 7

[115] Dongmei Zhang, Chang Li, Renrui Zhang, Shenghao Xie, Wei Xue, Xiaodong Xie, and Shanghang Zhang. Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection. In *AAAI*, 2024. 3

[116] 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 *ICCV*, 2025. 2, 3, 5, 6, 7

[117] Hu Zhang, Jianhua Xu, Tao Tang, Haiyang Sun, Xin Yu, Zi Huang, and Kaicheng Yu. Opensight: A simple open-vocabulary framework for lidar-based object detection. In *ECCV*. Springer, 2025. 3

[118] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025. 3

[119] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 3

[120] 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. 3

[121] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, 2025. 3

[122] Dingfu Zhou, Xibin Song, Yuchao Dai, Junbo Yin, Feixiang Lu, Jin Fang, Miao Liao, and Liangjun Zhang. Iafa: Instance-aware feature aggregation for 3d object detection from a single image. *arXiv preprint arXiv:2103.03480*, 2021. 3

[123] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3

[124] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3

[125] Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. *arXiv preprint arXiv:2506.23088*, 2025. 3

[126] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*, 2023. 3