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

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

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 human 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 capability 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.

Links: Project Page

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 Bai et al. (2025); Li et al. (2025); You et al. (2023). Yet one capability has lagged behind: general, multi-object 3D detection directly from monocular images. Existing monocular 3D detectors perform well within narrow domains, 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 Yao et al. (2024); Zhang et al. (2025), 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 Cho et al. (2024). 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

* *Work Done during an internship at NVIDIA.* † *Equal advising and corresponding authors: yxw@illinois.com, zhidingy@nvidia.com.*
Additional affiliations: [1] *University of Illinois Urbana-Champaign ,* [2] *The Hong Kong Polytechnic University.*

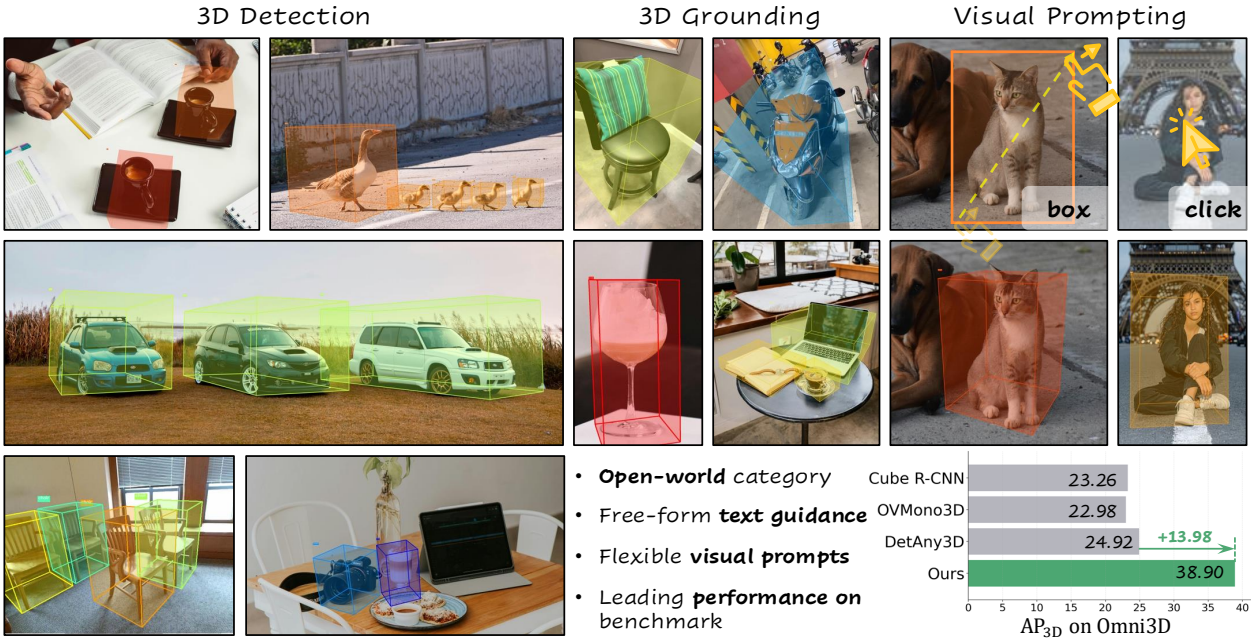

Figure 1: **LocateAnything3D** unifies 3D detection and grounding in a single vision-language model. It supports open-world categories with free-form text guidance and flexible visual prompts (*e.g.*, drag boxes, click points). All examples are zero-shot, highlighting strong out-of-domain generalizability.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

pose Marr (2010); Rock (1983); Von Helmholtz (1867). More 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 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 Brazil et al. (2023), our method attains state-of-the-art performance with **38.90** $\mathrm{AP}_{\mathrm{3D}}$, surpassing the previous best by **+13.98** absolute points even when the baseline is aided by ground-truth 2D boxes. 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→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, strong, and broadly applicable 3D perception within a VLM, closing a long-standing gap between open-vocabulary recognition and metric 3D understanding.

2. LocateAnything3D

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 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.

2.1. 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

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

where $\mathbf{t}_i = (X_i, Y_i, Z_i)^\top$ is the 3D center from the camera, $\mathbf{d}_i = (W_i, H_i, L_i)^\top$ are 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 previous equation is

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

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.

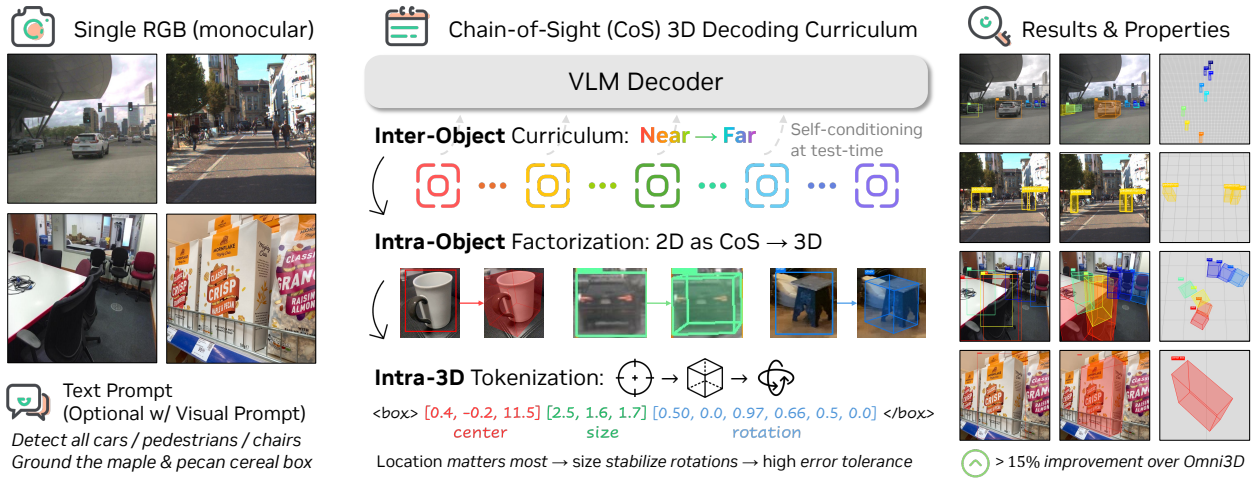

Figure 2: **Architecture of LocateAnything3D.** (1) Model input: a single RGB image with text and optional visual prompts (boxes/clicks). (2) Chain-of-Sight (CoS) decoding: a VLM decoder first emits 2D detections as an explicit visual evidence, then continues the sequence to 3D. Decoding follows three 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.

2.2. 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

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

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

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

where $\mathcal{S}_{<i}$ denotes all tokens emitted before step i. Compared to Eq. equation 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 Li et al. (2025); You et al. (2023). 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 Liu et al. (2024); Yao et al. (2024); Zhang et al. (2025). 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 Brazil et al. (2023), 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. equation 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 a 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.

3. 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. 2. 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 Baruch et al. (2021), SUN-RGBD Song et al. (2015), Hypersim Roberts et al. (2021), Objectron Ahmadyan et al. (2021), KITTI Geiger et al. (2013), and nuScenes Caesar et al. (2020) 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.

3.1. 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 element-wise 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.

3.2. Large-Scale Text Auto-Annotation

To supply rich referring expressions without manual labeling, we prompt strong VLMs Kavukcuoglu (2025); Team (2023) 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.

3.3. 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.

3.4. 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.

Table 1: **3D detection on the Omni3D benchmark.** Our LocateAnything3D achieves state-of-the-art results over all baselines, even outperform 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 Rukhovich et al. (2022)	23.5	23.4	21.5	-	-	-	-	-	-	9.4
SMOKE Liu et al. (2020)	25.9	20.4	20.0	-	-	-	-	-	-	10.4
OV-Uni3DETR Wang et al. (2023)	35.1	33.0	31.6	-	-	-	-	-	-	-
Cube R-CNN Brazil et al. (2023)	36.0	32.7	31.9	32.50	30.06	15.33	41.73	50.84	7.48	23.26
OVMono3D Yao et al. (2024)	-	-	-	25.45	24.33	15.20	41.60	58.87	7.75	22.98
DetAny3D	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	**39.8**	33.9	**36.1**	**43.75**	35.26	45.12	**59.89**	**71.90**	**18.12**	**38.90**

4. Experiments

Benchmarks and metrics. We evaluate on Omni3D Brazil et al. (2023), 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 (AP_{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. 2. All evaluations follow a *target-aware* protocol, as advocated in prior open-vocabulary works Yao et al. (2024); Zhang et al. (2025): 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 Brazil et al. (2023): the reference baseline released with Omni3D, a unified detector trained as a close-vocabulary model. (2) OVMono3D Yao et al. (2024): an open-vocabulary monocular 3D detector tailored to Omni3D. It "lifts" 2D detections to 3D by wiring an open-vocabulary 2D localizer Liu et al. (2024) to a 3D prediction head. (3) DetAny3D Zhang et al. (2025): 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 standard cross-entropy loss over the autoregressive token sequence. Additional training details, hyperparameters, and ablations are provided in the supplementary material.

Implementation details. Our work is built on SigLIP vision encoder Zhai et al. (2023) and Qwen2-8B backbone Bai et al. (2025) 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 for both vision and language, apply dynamic online packing to fill a 16,384-token context per sample, and optimize with AdamW 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.

4.1. 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

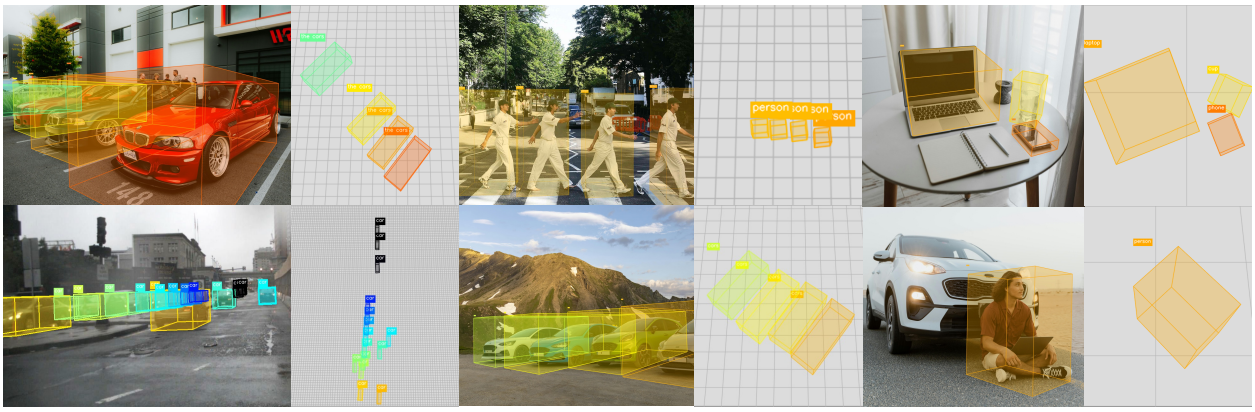

Figure 3: **Qualitative results of LocateAnything3D.** For each example, the left sub-figure overlays the projected 3D bounding boxes on the input image, while the right sub-figure shows the corresponding bird's-eye view with 1m×1m grids as the background. We use a turbo colormap based on depth, where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

(`Omni3D_OUT`), our method reaches **33.1** AP_{3D}, overperforming DetAny3D (32.2) and only behind 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 Yao et al. (2024); Zhang et al. (2025): 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 Geiger et al. (2013), **26.33** on SUN-RGBD Song et al. (2015), and **29.06** on ARK-itScenes Baruch et al. (2021), while competing approaches depend on an external 2D detector (Grounding DINO Liu et al. (2024)) 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.

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 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		
	$\mathrm{AP}^{\mathrm{kit}}_{3D}$	$\mathrm{AP}^{\mathrm{sun}}_{3D}$	$\mathrm{AP}^{\mathrm{ark}}_{3D}$
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

Table 3: **Indoor 3D Object Grounding Performance.** We compare LocateAnything3D against Cube-LLM trained on different data scales. Cube-LLM$_{\text{small}}$ is trained on the LV3D-small subset, while Cube-LLM$_{\text{large}}$ is trained on the full LV3D dataset containing approximately **9.6M images**. In contrast, our model is trained on a much smaller curated dataset of **1.7M images**. Despite this significant disparity in data scale, LocateAnything3D outperforms the best baseline by a large margin across all benchmarks. We report Average Precision (AP) prompted with either category names ($\text{AP}_{\text{3D}}^{\text{cat}}$) or category plus spatial location ($\text{AP}_{\text{3D}}^{\text{cat+loc}}$).

Method	Objectron		ARKitScenes		SUN-RGBD	
	$\text{AP}_{\text{3D}}^{\text{cat}}$	$\text{AP}_{\text{3D}}^{\text{cat+loc}}$	$\text{AP}_{\text{3D}}^{\text{cat}}$	$\text{AP}_{\text{3D}}^{\text{cat+loc}}$	$\text{AP}_{\text{3D}}^{\text{cat}}$	$\text{AP}_{\text{3D}}^{\text{cat+loc}}$
Cube-LLM$_{\text{small}}$ Cho et al. (2024)	56.7	36.1	21.6	28.3	25.5	25.5
Cube-LLM$_{\text{large}}$ *Choet al.* (2024)	69.8	45.4	23.5	31.8	29.7	28.8
LocateAnything3D (Ours)	**72.5**	**75.0**	**41.7**	**53.9**	**29.7**	**39.5**
Δ vs. Cube-LLM$_{\text{large}}$	+2.7	+29.6	+18.2	+22.1	+0	+10.7

4.2. Quantitative Evaluation of 3D Grounding

Problem setting. To further evaluate our LocateAnything3D's capability in following spatial language instructions, we conduct experiments on indoor 3D grounding benchmarks. We strictly follow the experimental protocol established by Cube-LLM Cho et al. (2024). Specifically, we repurpose the test sets of three standard indoor detection datasets: *Objectron* Ahmadyan et al. (2021), *ARKitScenes* Baruch et al. (2021), and *SUN-RGBD* Song et al. (2015) into grounding benchmarks. The task requires the model to localize particular objects based on text prompts that vary in specificity: (1) Category-only: The prompt contains only the object class name (*e.g.*, "chair"); and (2) Category+Location: The prompt includes the class name augmented with spatial descriptions derived from the object's position relative to the camera (*e.g.*, "chair on the left", "bookshelf close to camera"). The spatial qualifiers (left/right/center and close/medium/far) are generated based on the 2D image coordinates and depth thresholds defined in the baseline setting. We report the Average Precision (AP_{3D}) averaged over IoU_{3D} thresholds of $\tau \in \{0.15, 0.25, 0.50\}$. If multiple objects match the text description, the maximum IoU among them is used for evaluation.

Evaluation results. Table 3 summarizes the results on the benchmark. We copy the Cube-LLM numbers for models pre-trained on the "LV3D-small" and full "LV3D" corpora from their paper Cho et al. (2024), and add our LocateAnything3D model, which is trained using the Chain-of-Sight formulation on our unified 3D corpus. Across all three datasets and both metrics, LocateAnything3D substantially outperforms Cube-LLM, despite no task-specific architecture changes for indoor scenes.

From the table, we can also notice that Cube-LLM Cho et al. (2024) achieves lower performance for $\text{AP}_{\text{3D}}^{\text{cat+loc}}$ than $\text{AP}_{\text{3D}}^{\text{cat}}$, in two out of the three evaluation scenarios. On the contrary, LocateAnything3D achieves consistent performance improvement when location information is provided to the model as additional conditions. This different clearly highlights the higher capability of our model to interpret spatial descriptions and 3D understanding.

Problem with point-cloud grounding benchmarks. Existing indoor 3D grounding datasets such as ScanRefer Chen et al. (2020) and ReferIt3D Achlioptas et al. (2020) are explicitly built around point clouds rather than images, and are therefore ill-suited to our monocular 3D detection setting. Each scene in these benchmarks is represented by a single reconstructed point cloud but is associated with many RGB views that only partially observe the scene. Referring expressions are written to identify objects in the global 3D scene, not in a particular camera view, and a single object may be visible in multiple images with very different appearances and levels of occlusion. As a result, there is no unambiguous way to assign a unique image and 3D box pair to each language query, and any attempt to project the point-cloud annotations into 2D would depend on arbitrary choices of viewpoint and visibility thresholds. For this reason, we follow Cube-LLM Cho et al. (2024) and evaluate our model on indoor benchmarks derived from Objectron Ahmadyan et al. (2021), ARKitScenes Baruch et al. (2021), and SUN-RGBD Song et al. (2015), where each image already comes with camera-specific 3D boxes

and thus naturally supports monocular 3D detection and grounding.

4.3. 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 → far order with common alternatives degrades quality (Table 4). A random order performs worst (17.5), confirming that sequence position carries semantic load 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 (easy, high-evidence instances first) yields more stable and informative token prefixes.

Table 4: **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 $\mathrm{AP}_{\mathrm{3D}}^{\mathrm{out}}$. Our full design (highlighted) achieves the best performance, validating the importance of each design choice.

Design Component	Variant	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{out}}$ ↑
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**

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, even when predicted after the 3D signal.

Intra-3D token order. Within each object, decoding with the center, size, and rotation ordering performs best. Switching to Rotation-Size-Center harms results (32.9), and Center-Rotation-Size is slightly worse (28.8), 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.

4.4. Data Efficiency and Training Dynamics

To better understand the contributions of our Chain-of-Sight (CoS) formulation and the role of 2D pretraining, we conduct a detailed analysis of our model's performance under limited data regimes and different initialization strategies. Figure **??** visualizes these comparisons.

Impact of chain-of-sight on data efficiency. Figure **??** (left) compares our full 2D-3D CoS formulation against a "pure 3D" decoder trained without any explicit 2D step. On the horizontal axis we vary the fraction of our 3D training corpus from 10% to 100%; the dashed line marks the performance of DetAny3D Zhang et al. (2025) (32.2 $\mathrm{AP}_{\mathrm{3D}}$).

Across all data regimes, the CoS model is consistently stronger and markedly more data-efficient than the pure 3D variant. With only 10% of the data, CoS outperforms pure 3D prediction baseline by a large margin. As we further scale the data to 70% and 100%, the CoS curve continues to climb to 32.7 and 36.1 $\mathrm{AP}_{\mathrm{3D}}$, whereas pure 3D saturates at 19.5 and 22.7. This supports our central claim that explicitly factorizing 3D detection into a 2D grounding step followed by 3D lifting is not just more accurate, but also significantly more sample-efficient.

Impact of 2D pretraining on convergence. Figure **??** (right) studies the effect of the 2D grounding pretraining stage on our training dynamics. We plot $\mathrm{AP}_{\mathrm{3D}}$ as a function of CoS training steps, comparing models initialized with and without 2D pretraining, and again mark the DetAny3D performance with a dashed line.

Initializing from the 2D grounding stage yields a substantial head start. After only 1k CoS steps, the pretrained model already achieves 19.6 AP_{3D}, whereas the model trained from scratch is still at 7.3. As training proceeds, both curves improve, but the gap persists. At the final checkpoint, the model with 2D pretraining converges to 36.1 AP_{3D}, while the scratch model lags behind at 29.2. This indicates that robust 2D localization capabilities serve as a critical foundation for 3D perception, allowing the model to focus its capacity on lifting 2D features to 3D space rather than learning basic localization from scratch.

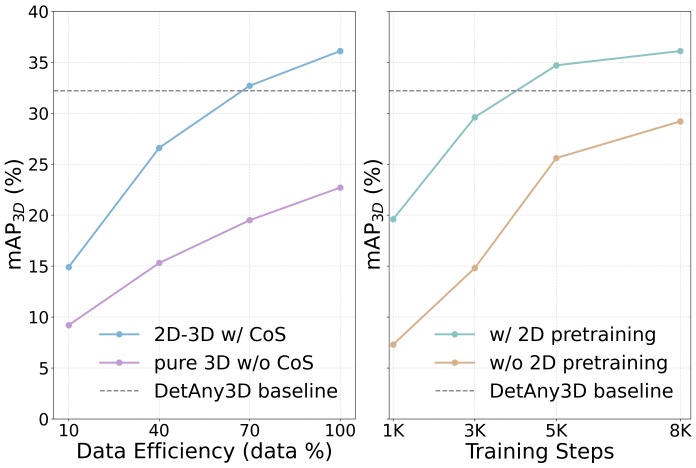

Figure 4: **Data efficiency and training dynamics analysis.** (1) The left figure shows data efficiency: We report AP_{3D} vs. percentage of training data used. Our Chain-of-Sight (CoS) formulation (blue) consistently outperforms direct 3D prediction (purple), achieving competitive performance with only 10% of the data. (2) The right figure shows training dynamics: We compare training curves with and without 2D detection pretraining. 2D pretraining (green) accelerates convergence significantly, surpassing the previous state of the art (dashed line) almost immediately, whereas training from scratch (orange) is slower and yields lower final accuracy.

4.5. 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.

5. 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.

5.1. 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 Radford et al. (2021), image-text pretraining for retrieval Li et al. (2022), and strong detectors for region-level grounding Kirillov et al. (2023); Liu et al. (2024); Ren et al. (2024). Building atop these capabilities, recent VLM systems either orchestrate experts as tools under a multimodal controller Liu et al. (2023); Wu et al. (2023) or pursue unified backbones that natively tackle a wide spectrum of perception tasks Liu et al. (2023); Team (2024,, 2023); Wang et al. (2023); Xiao et al. (2024); You et al. (2023). 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 Kazemzadeh et al. (2014); Liu et al. (2024); Mao et al. (2016); Plummer et al. (2015); Wang et al. (2024); Yu et al. (2016). 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 Bai et al. (2025); Google DeepMind (2024); Kavukcuoglu (2025); Peng et al. (2023), and point-based localization in MoLMo and RoboPoint Deitke et al. (2024); Yuan et al. (2024). 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 Xue et al. (2025), and 2D grounding as reasoning chain-of-thought Man et al. (2025); Shao et al. (2024). 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 Bai et al. (2025); Li et al. (2025).

5.2. 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 relied on prompting to elicit planning behaviors from VLM backbones Hu et al. (2023); Kim et al. (2024); Shin et al. (2024); Singh et al. (2022); Song et al. (2023), with code and API-centric tool interfaces further improving reliability Liang et al. (2022); Silver et al. (2024). Subsequent work introduces supervised finetuning, yielding compact yet capable agents for manipulation Huang et al. (2025); Kim et al. (2024); Lee et al. (2025); Liu et al. (2025); Lu et al. (2025); Zawalski et al. (2024); Zhang et al. (2025); Zhao et al. (2025) and household or procedural reasoning Chen et al. (2023); Ji et al. (2025); Wu et al. (2023). In parallel, spatial intelligence has emerged as essential competencies for open-world embodiment, with lines of work targeting distance/metric understanding and counting Cai et al. (2024); Chen et al. (2024); Du et al. (2024); Fu et al. (2024); Liao et al. (2024); Song et al. (2024); Yang et al. (2024); Zhou et al. (2025), as well as benchmark suites that synthesize complex 3D scenes and tasks Cheng et al. (2024); Ray et al. (2024). Embodied pointing and grounding further connect perception to action Hong et al. (2023); Li et al. (2024); Yuan et al. (2024), and recent efforts augment spatial reasoning with structured reasoning Liu et al. (2025); Yuan et al. (2025). Integrated frameworks exemplify the trend toward generalist embodied VLMs and unify perception, reasoning, and planning at scale Ji et al. (2025); Team et al. (2025), while curated benchmarks continue to expand the supervision landscape Chen et al. (2023); Cheng et al. (2025); Deitke et al. (2025); Luo et al. (2025); NVIDIA et al. (2025); Qiu et al. (2024); Qu et al. (2025); Team (2025); Yang et al. (2025,); Yuan et al. (2024). Beyond supervision fine-tuning (SFT), reinforcement-driven training also starts to revolutionize the role of reasoning traces in embodiment Chen et al. (2025); Yang et al. (2025); Yuan et al. (2025). 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.

5.3. 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 Caesar et al. (2020); Chen et al. (2021, 2016); Geiger et al. (2013); Huang et al. (2022); Li et al. (2024); Liu et al. (2020); Wang et al. (2021, 2022); Zhang et al. (2023); Zhou et al. (2021, 2019). Parallel lines of work study multi-sensor fusion and spatio-temporal reasoning to boost accuracy, yet typically inherit closed-set label space constraints Liang et al. (2022); Lin et al. (2022); Man et al. (2023). 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 Brazil et al. (2023). Subsequent efforts further explore bird-eye's-view formulations across indoor and outdoor settings Jhang et al. (2025); Li et al. (2024). 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 Cao et al. (2024,); Lu et al. (2022, 2023); Peng et al. (2025); Russakovsky et al. (2015); Wang et al. (2024); Zhang et al. (2024, 2025, 2022); Zhou et al. (2022); Zhu et al. (2023). Closer to our setting, OVMono3D lifts open-vocabulary 2D detections (*e.g.*, from Grounding DINO) into 3D with a unified head Liu et al. (2024); Yao et al. (2024). DetAny3D proposes a promptable 3D foundation model that transfers knowledge from 2D foundation models to monocular 3D via feature aggregation Zhang et al. (2025). 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.

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 believe 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.

A. Erratum

We would like to acknowledge an incorrect implementation of the evaluation in the initial report, which has been corrected in this revision. A root cause of the mistake was that mAP was computed with cross-image matches, instead of being computed on a per-image basis in standard protocol. The updated results are lower than those in the initial report, but the claims and conclusions remain unchanged. Problem definition, model training, inference, and visualization are not affected.

B. Acknowledgments

The team would like to thank the valuable discussions and input from Tianyi Xiong, Shaokun Zhang, Guo Chen, Di Zhang, Guilin Liu, Xiaolong Li, Paris Zhang, Yilin Zhao, Subhashree Radhakrishnan, Sifei Liu, Hongxu (Danny) Yin, Valts Blukis, Jonathan Tremblay, Bowen Wen, Yan Chang, Wei Liu, Yan Wang. We would also like to acknowledge the following teams: Metropolis, VILA, LPR Robotics, GEAR Lab, AV Research, and ISSAC Robotics. We would also like to thank the NVIDIA infrastructure team for their prompt and helpful assistance.

C. Additional Experiments and Analysis

C.1. Impact of Token Serialization Strategy

To further validate our Chain-of-Sight (CoS) design of interleaving 2D and 3D tokens on the per-object level ($\mathrm{2D_i} \rightarrow \mathrm{3D_i}$), we compare it against a "clustered" decoding strategy. In the clustered setting, the model is trained to predict all 2D bounding boxes for the scene first, followed by all corresponding 3D bounding boxes ($\mathrm{2D_{1...N}} \rightarrow \mathrm{3D_{1...N}}$). This ablation tests whether the tight coupling of 2D visual evidence with its corresponding 3D geometry is necessary, or if the model can simply learn two separate phases of detection. We report results trained for 1 epoch on three distinct datasets to analyze performance across different scene complexities. As shown in Table 5, our interleaved default setting consistently outperforms the clustered strategy. The magnitude of this performance gap is strongly correlated with scene clutter and object density.

Table 5: **Ablation of Token Serialization Strategy.** We compare our default *Interleaved* Chain-of-Sight strategy ($2D_i \rightarrow 3D_i$) against a *Clustered* strategy where all 2D boxes are predicted before all 3D boxes ($\mathrm{2D_{1...N}} \rightarrow \mathrm{3D_{1...N}}$) with average precision ($\mathrm{AP_{3D}}$). Models are trained for 1 epoch. The results show that the interleaved strategy is significantly more robust, especially in cluttered scenes where associating separated 2D and 3D sequences becomes difficult.

Serialization Strategy	Average Precision ($\mathrm{AP_{3D}}$)		
	Objectron	KITTI	Hypersim
Clustered ($\mathrm{2D_{1...N}} \rightarrow \mathrm{3D_{1...N}}$)	61.5	17.4	4.7
Interleaved ($\mathrm{2D_i} \rightarrow \mathrm{3D_i}$, Ours)	**63.0**	**22.1**	**11.2**
Performance Gap	+1.5	+4.7	+6.5

Object-centric scenes. On Objectron, which typically contains only 1 or 2 prominent objects per image, the performance gap is minimal (61.5 vs. 63.0). The additional effort for the model to associate the i-th 3D box with the i-th 2D box is negligible.

Structured outdoor scenes. KITTI scenes contain more objects with large depth range, but they follow a structured distribution (cars on a road) with relatively clear depth ordering. While the gap widens, the model can still maintain reasonable 2D-3D association in the clustered setting.

Highly cluttered scenes. The most significant drop occurs on Hypersim which is characterized by chaotic indoor scenes with dozens of objects and frequent occlusions. In these scenarios, the clustered strategy fails catastrophically. The model struggles to maintain the implicit alignment between the k-th 2D box generated early in the sequence and the k-th 3D box generated much later, resulting in a big difference between the two setting.

C.2. Runtime Analysis

Although LocateAnything3D is primarily designed as a general 3D perception VLM rather than a real-time perception system, we report its end-to-end inference latency for completeness. On average, processing a single image-query pair with LocateAnything3D takes **683 ms** under our evaluation setup with a single H100 GPU. This wall-clock time consists of three main components: (1) vision encoding of the input image, (2) LLM pre-filling with the textual prompt, and (3) autoregressive generation of the mixed 2D/3D box tokens produced by the Chain-of-Sight decoder.

To isolate the cost of the Chain-of-Sight factorization, we compare our full 2D-3D CoS model with a pure-3D variant that directly predicts 3D boxes without emitting intermediate 2D boxes. Introducing the 2D step increases the average latency by only **121 ms** (from roughly 562 ms to 683 ms), yet enables the substantial accuracy gains and data-efficiency improvements, as reported in the Section 5 of the main paper. In other words, CoS adds a modest computational overhead while making 3D detection both easier to learn and significantly more accurate.

We emphasize that LocateAnything3D is not meant to replace highly optimized real-time detectors used in latency-critical loops (*e.g.*, onboard obstacle avoidance). Instead, our goal is to endow a general-purpose VLM with strong 3D grounding capabilities so that it can serve as a foundation for downstream tasks such as offline planning, scene understanding, and multimodal agent reasoning. In this context, a sub-second per-image latency is well within an acceptable range, especially given the unified interface and performance benefits brought by the Chain-of-Sight formulation.

D. Implementation Details

D.1. Models, Tokenization, and Prompting

Model designs. (1) vision encoder. We use SigLIP Zhai et al. (2023) with FlashAttention 2 Dao et al. (2022) enabled. (2) Language model. We use a Qwen2 8B causal LM Team (2024) with FlashAttention 2, trained end-to-end (no freezing). (3) Multimodal connector. We use an MLP projector, which maps SigLIP tokens to the LLM hidden space with two-layer MLP.

Image tokenization. A tiling-based tokenization where we decompose images into patches of a forced image size of 448. The total image tokens scale linearly with the number of tiles.

Conversation format. Qwen2-chat template. Image tokens are inserted by replacing each `<image>` placeholder with `<IMG_START>` followed by repeated `<IMG_CONTEXT>` tokens and `<IMG_END>`. The repeat count equals per-tile tokens times the number of tiles for that image; we assert a strict match between precomputed and actual counts.

Labels. Only assistant spans are supervised; all instruction tokens are masked. Truncation safety checks keep training targets valid.

D.2. Dynamic Tiling and Packing

Tiling and image processing. Images are decomposed into an adaptive grid of 448-pixel tiles, min 1 and max 12 tiles, plus an optional global thumbnail. Tiling policy selects the closest aspect ratio while favoring large area coverage for stability.

Sequence construction and online packing. Our context length is 16,384 tokens per sample. We enable online packing to concatenate multiple short samples until the context budget is filled while tracking sub-sample boundaries in the attention mask. A dummy image is inserted only if the entire packed sample is text-only. Position ids respect packed boundaries; the model supports sequence parallel groups but we run with degree 1 in our experiments.

Category	Dataset
Captioning & Knowledge	ShareGPT4o OpenGVLab (2024), KVQA Shah et al. (2019), Movie-Posters skvarre (2024), Google-Landmark Weyand et al. (2020), WikiArt HugGAN (2024), Weather-QA Ma et al. (2024), Coco-Colors hazal karakus (2024), music-sheet EmileEsmaili (2024), SPARK Yu et al. (2024), Image-Textualization Pi et al. (2024), SAM-Caption PixArt-alpha (2024), Tmdb-Celeb-10k Ashraq (2024)
Mathematics	GeoQA+ Cao and Xiao (2022), MathQA Yu et al. (2023), CLEVR-Math/Super Li et al. (2023); Lindström and Abraham (2022), Geometry3K Lu et al. (2021), MAVIS-math-rule-geo Zhang et al. (2024), MAVIS-math-metagen Zhang et al. (2024), InterGPS Lu et al. (2021), Raven Zhang et al. (2019), GEOS Seo et al. (2015), UniGeo Chen et al. (2022)
Science	AI2D Kembhavi et al. (2016), ScienceQA Lu et al. (2022), TQA Kembhavi et al. (2017), PathVQA He et al. (2020), SciQA Auer et al. (2023), Textbooks-QA, VQA-RAD Lau et al. (2018), VisualWebInstruct TIGER-Lab (2024)
Chart & Table	ChartQA Masry et al. (2022), MMC-Inst Liu et al. (2023), DVQA Kafle et al. (2018), PlotQA Methani et al. (2020), LRV-Instruction Liu et al. (2023), TabMWP Lu et al. (2022), UniChart Masry et al. (2023), Vistext Tang et al. (2023), TAT-DQA Zhu et al. (2022), VQAonBD VQAonDB (2024), FigureQA Kahou et al. (2017), Chart2Text Kantharaj et al. (2022), RobuT-{Wikisql, SQA, WTQ} Zhao et al. (2023), MultiHiertt Zhao et al. (2022)
Naive OCR	SynthDoG Kim et al. (2022), MTWI He et al. (2018), LVST Sun et al. (2019), SROIE Huang et al. (2019), FUNSD Jaume et al. (2019), Latex-Formula OleehyO (2024), IAM Marti and Bunke (2002), Handwriting-Latex aidapearson (2023), ArT Chng et al. (2019), CTW Yuan et al. (2019), ReCTs Zhang et al. (2019), COCO-Text Veit et al. (2016), SVRD Yu et al. (2023), Hiertext Long et al. (2023), RoadText Tom et al. (2023), MapText Li et al. (2024), CAPTCHA parasam (2024), Est-VQA Wang et al. (2020), HME-100K TAL (2023), TAL-OCR-ENG TAL (2023), TAL-HW-MATH TAL (2023), IMGUR5K Krishnan et al. (2023), ORAND-CAR Diem et al. (2014), Invoices-and-Receipts-OCR mychen76 (2024), Chrome-Writting Mouchère et al. (2016), IIIT5k Mishra et al. (2012), K12-Printing TAL (2023), Memotion Ramamoorthy et al. (2022), Arxiv2Markdown, Handwritten-Mathematical-Expression Azu (2023), WordArt Xie et al. (2022), RenderedText wendlerc (2024), Handwriting-Forms ift (2024)
OCR QA	DocVQA Clark and Gardner (2018), InfoVQA Mathew et al. (2022), TextVQA Singh et al. (2019), ArxivQA Li et al. (2024), ScreencQA Hsiao et al. (2022), DocReason mPLUG (2024), Ureader Ye et al. (2023), FinanceQA Sujet AI (2024), DocMatrix Laurençon et al. (2024), A-OKVQA Schwenk et al. (2022), Diagram-Image-To-Text Kamizuru00 (2024), MapQA Chang et al. (2022), OCRVQA Mishra et al. (2019), ST-VQA Biten et al. (2019), SlideVQA Tanaka et al. (2023), PDF-VQA Ding et al. (2023), SQuAD-VQA, VQA-CD Mahamoud et al. (2024), Block-Diagram shreyanshu09 (2024), MTVQA Tang et al. (2024), ColPali Faysse et al. (2024), BenthamQA Mathew et al. (2021)
General VQA	LLaVA-150K Liu et al. (2023), LVIS-Instruct4V Wang et al. (2023), ALLaVA Chen et al. (2024), Laion-GPT4V LAION (2023), LLAVAR Zhang et al. (2023), SketchyVQA Tu et al. (2023), VizWiz Gurari et al. (2018), IDK Cha et al. (2024), AlfworldGPT, LNQA Pont-Tuset et al. (2020), Face-Emotion FastJobs (2024), SpatialSense Yang et al. (2019), Indoor-QA keremberke (2024), Places365 Zhou et al. (2017), MMinstruct Liu et al. (2024), DriveLM Sima et al. (2023), YesBut Nandy et al. (2024), WildVision Lu et al. (2024), LLaVA-Critic-113k Xiong et al. (2024), RLAIF-V Yu et al. (2024), VQAv2 Goyal et al. (2017), MMRA Wu et al. (2024), KONIQ Hosu et al. (2020), MMDU Liu et al. (2024), Spot-The-Diff Jhamtani and Berg-Kirkpatrick (2018), Hateful-Memes Kiela et al. (2020), COCO-QA Ren et al. (2015), NLVR Suhr et al. (2017), Mimic-CGD Laurençon et al. (2024), Datikz Belouadi et al. (2023), Chinese-Meme Contributors (2024), IconQA Lu et al. (2021), Websight Laurençon et al. (2024)
Text-only	Orca Lian et al. (2023), Orca-Math Mitra et al. (2024), OpenCodeInterpreter Zheng et al. (2024) MathInstruct Yue et al. (2023), WizardLM Xu et al. (2023), TheoremQA Chen et al. (2023), OpenHermes2.5 Teknium (2023), NuminaMath-CoT LI et al. (2024), Python-Code-25k flytech (2024), Infinity-Instruct BAAI (2024), Python-Code-Instructions-18k-Alpaca iamtarun (2024), Ruozhiba LooksJuicy (2024), InfinityMATH Zhang et al. (2024), StepDPO Lai et al. (2024), TableLLM Zhang et al. (2024), UltraInteract-sft Yuan et al. (2024)
2D Grounding & Counting	RefCOCO/+/g (en) Mao et al. (2016); Yu et al. (2016), Objects365 Shao et al. (2019), COCO Lin et al. (2014), EgoObjects Zhu et al. (2023), BLIP3-OCR Xue et al. (2024), BDD100K Yu et al. (2020), Nuimages Caesar et al. (2019), Flick30K Plummer et al. (2015), LVIS Gupta et al. (2019)

Table 6: **Summary of our extensive and diverse supervised fine-tuning datasets for 2D pretraining.** We use a comprehensive collection of numerous large-scale datasets spanning multiple domains and tasks to pretrain our model, ensuring broad coverage and robust performance across diverse visual and language understanding scenarios.

D.3. Optimization and Systems

We use a precision of bfloat16 across vision and language. For memory handling, gradient checkpointing is enabled for both the SigLIP encoder and the LLM; fused ops are used to reduce memory overhead. For the loss function, we fuse the linear cross-entropy with per-sample normalization using the number of valid answer tokens. For the optimizer and schedule, we use AdamW with a learning rate of 1e-5, a weight decay of 0.05, a cosine decay, and a warm-up of 3%.

Packing target. We use dynamic online packing to saturate the 16K context; the scripts set an iteration-level token target of 2^{17} $(= 128K)$ to govern accumulation and throughput.

Training scale. We train our model using 64 H100 GPUs. The whole training takes 46 hours with 37K steps, distributed with torchrun and DeepSpeed ZeRO-3.

D.4. 2D Grounding Pretraining

Dataset composition. We pretrain on a large-scale 2D grounding corpus covering four domains with different data mixture percentage: (1) **General detection:** Object365 Shao et al. (2019) (5 epochs), COCO Lin et al. (2014) (12 epochs), and LVIS Gupta et al. (2019) (3 epochs); (2) **Ego-centric & driving:** BDD100K Yu et al. (2020) (3 epochs), nuImages Caesar et al. (2020) (3 epochs), and EgoObjects Zhu et al. (2023) (3 epochs); (3) **Referring-expression grounding:** RefCOCO Yu et al. (2016) (3 epochs), RefCOCO+ Mao et al. (2016) (3 epochs), RefCOCOg Mao et al. (2016) (3 epochs), and Flickr30k Plummer et al. (2015) (3 epochs); (4) **Text grounding:** a BLIP3-OCR subset Xue et al. (2024) (≈ 1.0M samples). Overall, this results in over **15M** multi-turn dialogues in the grounding corpus, which we mix with an additional **8M** samples for general instruction tuning, as demonstrated in Table 6.

Annotation format. For each image, we construct a multi-turn dialogue where each turn follows the instruction template "Detect all the objects in the image that belong to the category set {c}." The response is either a comma-separated list of 2D bounding boxes in $[x_1, y_1, x_2, y_2]$ format (top-left to bottom-right, integer-quantized to $[0, 1000]$), or "None" if no instance exists. We include all positive categories present in the image and sample 10 absent categories as negatives, yielding per-image supervision that mixes existence and non-existence signals across multiple dialogue turns.

E. Limitations and Future Work

While LocateAnything3D establishes a strong foundation for VLM-native 3D perception, several avenues remain for future exploration. Our work primarily focuses on validating the Chain-of-Sight (CoS) decoding mechanism within a single-frame, end-to-end setting. Below, we outline key directions where our framework can be naturally extended to incorporate additional geometric signals and temporal contexts.

Integration of explicit depth priors. Currently, our model infers metric depth solely from monocular RGB cues and semantic context. While the near-to-far curriculum effectively regularizes this process, the model does not yet leverage explicit depth maps. Future work could introduce a depth encoder or use depth images as an additional conditioning input. This would allow the model to utilize output from state-of-the-art monocular depth estimators as a geometric prompt, potentially improving metric accuracy in texture-less or ambiguous scenes.

Explicit camera intrinsic conditioning. Our current approach normalizes 3D coordinates into a unified camera-centric space to maximize cross-dataset generalization. However, it implicitly relies on the vision encoder to handle variations in focal length and field of view. An extension is to explicitly tokenize camera intrinsic matrices (*e.g.*, focal length, principal point) and feed them as positional prompts. This would allow the decoder to mathematically adjust its size and depth predictions based on the specific camera optics, rather than learning an average projection model.

Extension to multi-frame and video settings. The current framework operates on single images. However, the autoregressive nature of our decoder is naturally suited for temporal sequences. Future iterations could extend the context window to include visual tokens from preceding frames. The model could learn to track objects over time, estimate velocity, and leverage multi-view consistency to resolve depth ambiguities that are ambiguous in a single frame.

F. Broader Impact

The development of LocateAnything3D represents a step toward unifying semantic understanding and metric perception within general-purpose foundation models. By enabling VLMs to perceive the physical world in 3D without specialized heads, we lower the barrier to entry for developing capable embodied agents and home robotics. This has positive implications for industries ranging from autonomous driving to assistive robotics.

However, we acknowledge potential risks associated with this technology. Like all deep learning models trained on web-scale data, our model may inherit biases present in the training corpora, such as geographic or cultural biases in object distributions. This could lead to uneven performance across different regions. We encourage the research community to prioritize the development of diverse, representative 3D datasets and to consider the ethical implications of spatial intelligence in deployment scenarios.

G. More Case Visualization

We provided more qualitative visualization in this section.

Failure case visualization. Figure 5 demonstrates some representative failure cases of our method. Despite the great performance, our model still suffers from the lack of diverse and high-quality 3D annotations compared

Figure 5: **Visualization of failure cases**. We show several failure cases of our model. Due to the lack of diverse 3D annotations, similar to the baselines Yao et al. (2024); Zhang et al. (2025), our model faces challenges when presented with scenes that exhibit very different focal length, spatial layouts, and textural details.

to the 2D scenario. Hence, similar to the baselines Yao et al. (2024); Zhang et al. (2025), it faces challenges when presented with scenes that exhibit very different focal length, spatial layouts, and textural details.

More successful visualization. Figure 6 demonstrates more successful cases.

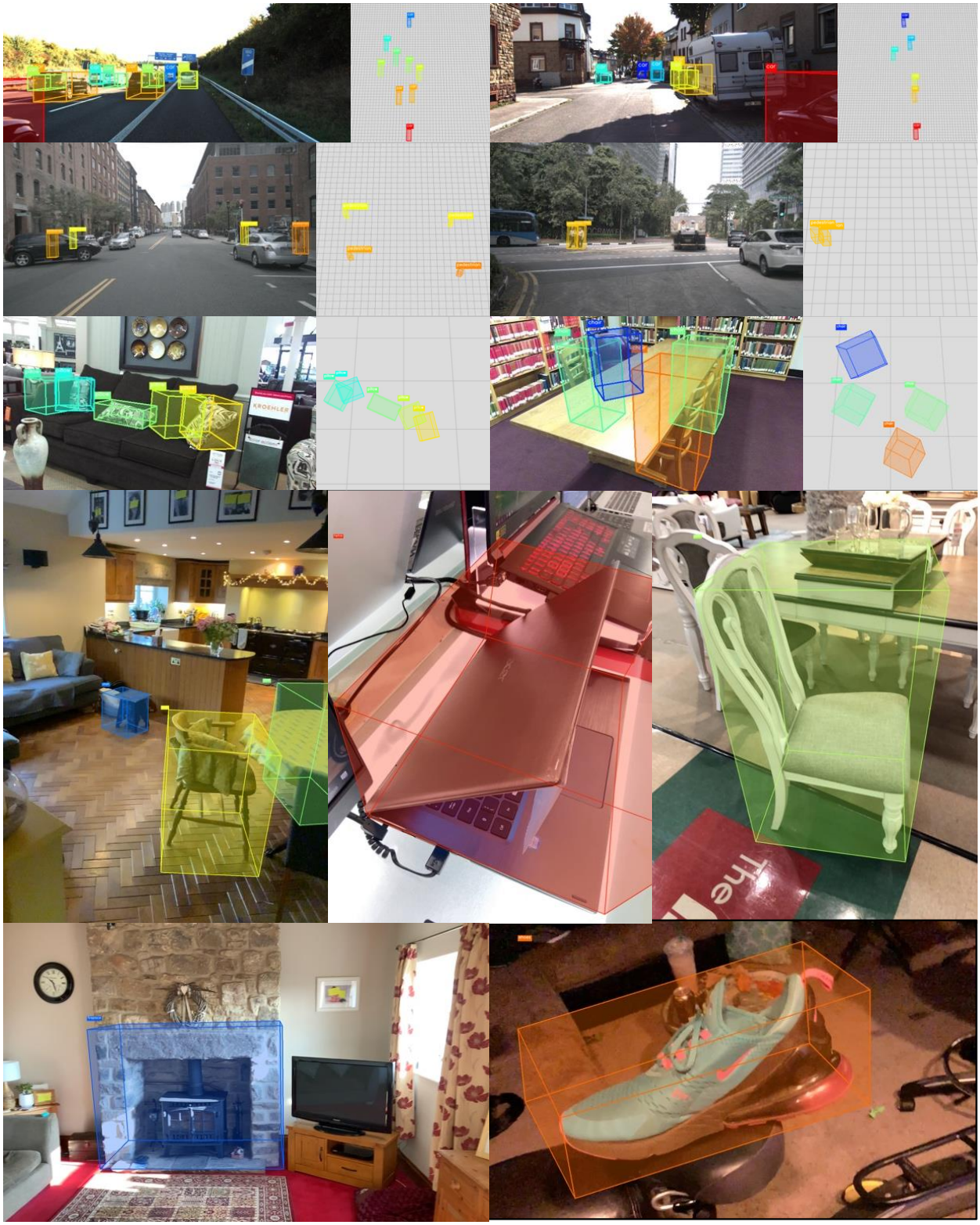

Figure 6: Visualization of more indoor and outdoor successful cases.

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] 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.

[3] aidapearson. Aida calculus math handwriting recognition dataset. https://www.kaggle.com/datasets/aidapearson/ocr-data, 2023.

[4] Ashraq. Tmdb-celeb-10k dataset. https://huggingface.co/datasets/ashraq/tmdb-celeb-10k, 2024.

[5] Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240, 2023.

[6] Azu. Handwritten-mathematical-expression-convert-latex. https://huggingface.co/datasets/Azu/Handwritten-Mathematical-Expression-Convert-LaTeX, 2023.

[7] BAAI. Infinity-instruct dataset. https://huggingface.co/datasets/BAAI/Infinity-Instruct, 2024.

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

[9] 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.

[10] Jonas Belouadi, Anne Lauscher, and Steffen Eger. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*, 2023.

[11] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019.

[12] 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.

[13] 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. *arXiv preprint arXiv:1903.11027*, 2019.

[14] 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.

[15] 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.

[16] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022.

[17] 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.

[18] 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.

[19] Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. Visually dehallucinative instruction generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5510–5514. IEEE, 2024.

[20] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

[21] 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.

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

[23] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.

[24] 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.

[25] 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.

[26] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.

[27] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, 2023.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pages 1571–1576, 2019.

[34] 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.

[35] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018.

[36] LLM-Red-Team Contributors. emo-visual-data: Emotion and visual data analysis project. https://github.com/LLM-Red-Team/emo-visual-data, 2024.

[37] 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.

[38] 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.

[39] 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.

[40] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S Oliveira. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsrc 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 779–784. IEEE, 2014.

[41] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer, 2023.

[42] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024.

[43] EmileEsmaili. sheet music clean ataset. https://huggingface.co/datasets/EmileEsmaili/sheet_music_clean, 2024.

[44] FastJobs. Visual emotional analysis dataset. https://huggingface.co/datasets/FastJobs/Visual_Emotional_Analysis, 2024.

[45] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

[46] flytech. Python codes 25k dataset. https://huggingface.co/datasets/flytech/python-codes-25k, 2024.

[47] 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.

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

[49] 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/, December 2024. Accessed 2025-04-28.

[50] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[51] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[52] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[53] hazal karakus. mscoco-controlnet-canny-less-colors dataset. https://huggingface.co/datasets/hazal-karakus/mscoco-controlnet-canny-less-colors, 2024.

[54] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *2018 24th international conference on pattern recognition (ICPR)*, pages 7–12. IEEE, 2018.

[55] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[56] 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.

[57] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[58] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*, 2022.

[59] 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.

[60] 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.

[61] 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.

[62] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction. In *International conference on document analysis recognition*, 2019.

[63] HugGAN. Wikiart dataset. https://huggingface.co/datasets/huggan/wikiart, 2024.

[64] iamtarun. Python code instructions 18k alpaca dataset. https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca, 2024.

[65] ift. Handwriting forms dataset. https://huggingface.co/datasets/ift/handwriting_forms, 2024.

[66] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[67] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

[68] 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.

[69] 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.

[70] 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.

[71] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018.

[72] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[73] Kamizuru00. Diagram image to text dataset. https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text, 2024.

[74] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022.

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

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

[77] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016.

[78] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pages 4999–5007, 2017.

[79] keremberke. Indoor scene classification dataset. https://huggingface.co/datasets/keremberke/indoor-scene-classification, 2024.

[80] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[81] 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.

[82] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, 2022.

[83] 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.

[84] 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.

[85] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9122–9134, 2023.

[86] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

[87] LAION. Gpt-4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023.

[88] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[89] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

[90] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[91] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.

[92] 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.

[93] 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.

[94] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

[95] 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.

[96] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024.

[97] Zekun Li, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu, and Nathalie Abadie. Icdar 2024 competition on historical map text detection, recognition, and linking. In *International Conference on Document Analysis and Recognition*, pages 363–380. Springer, 2024.

[98] 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.

[99] 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.

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

[101] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, 2023.

[102] W Lian, B Goodson, E Pentland, et al. Openorca: An open dataset of gpt augmented flan reasoning traces, 2023.

[103] 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.

[104] 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.

[105] 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.

[106] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[107] 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.

[108] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

[109] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[110] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023.

[111] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023.

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

[113] 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.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024.

[119] 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.

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

[121] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024.

[122] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Icdar 2023 competition on hierarchical text detection and recognition. In *International Conference on Document Analysis and Recognition*, pages 483–497. Springer, 2023.

[123] LooksJuicy. Ruozhiba dataset. https://huggingface.co/datasets/LooksJuicy/ruozhiba, 2024.

[124] 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.

[125] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

[126] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.

[127] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

[128] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35:2507–2521, 2022.

[129] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.

[130] 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.

[131] 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.

[132] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024.

[133] 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.

[134] Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

[135] Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. Chic: Corporate document for visual question answering. In *International Conference on Document Analysis and Recognition*, pages 113–127. Springer, 2024.

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

[137] 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.

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

[139] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

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

[141] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46, 2002.

[142] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022.

[143] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023.

[144] Minesh Mathew, Lluis Gomez, Dimosthenis Karatzas, and CV Jawahar. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3): 235–249, 2021.

[145] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022.

[146] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020.

[147] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012.

[148] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019.

[149] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024.

[150] Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 607–612. IEEE, 2016.

[151] mPLUG. Docreason25k dataset. https://huggingface.co/datasets/mPLUG/DocReason25K, 2024.

[152] mychen76. Invoices and receipts ocr v1 dataset. https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1, 2024.

[153] Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. *arXiv preprint arXiv:2409.13592*, 2024.

[154] 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.

[155] OleehyO. Latex formulas dataset. https://huggingface.co/datasets/OleehyO/latex-formulas, 2024.

[156] OpenGVLab. Sharegpt-4o dataset. https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o, 2024.

[157] parasam. Captcha dataset. https://www.kaggle.com/datasets/parsasam/captcha-dataset, 2024.

[158] 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.

[159] 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.

[160] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*, 2024.

[161] PixArt-alpha. Sam-llava-captions10m dataset. https://huggingface.co/datasets/PixArt-alpha/SAM-LLaVA-Captions10M, 2024.

[162] 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.

[163] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020.

[164] 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.

[165] 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.

[166] 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.

[167] Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.

[168] 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.

[169] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

[170] 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.

[171] 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.

[172] Irvin Rock. The logic of perception, 1983.

[173] 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.

[174] 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.

[175] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022.

[176] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015.

[177] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, 2019.

[178] 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.

[179] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019.

[180] 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.

[181] shreyanshu09. Block diagram dataset. https://huggingface.co/datasets/shreyanshu09/Block_Diagram, 2024.

[182] 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.

[183] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

[184] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.

[185] 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.

[186] skvarre. Movie posters-100k dataset. https://huggingface.co/datasets/skvarre/movie_posters-100k, 2024.

[187] 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.

[188] 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.

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

[190] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

[191] Hamed Rahimi Sujet AI, Allaa Boutaleb. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa, 2024. URL https://huggingface.co/datasets/sujet-ai/Sujet-Finance-QA-Vision-100k.

[192] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *ICDAR*, pages 1557–1562, 2019.

[193] TAL. Tal open dataset. https://ai.100tal.com/dataset, 2023.

[194] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

[195] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023.

[196] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

[197] 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.

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

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

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

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

[202] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

[203] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants. https://huggingface.co/datasets/teknium/OpenHermes-2.5, 2023.

[204] TIGER-Lab. Visualwebinstruct dataset. https://huggingface.co/datasets/TIGER-Lab/VisualWebInstruct, 2024.

[205] George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. Icdar 2023 competition on roadtext video text detection, tracking and recognition. In *International Conference on Document Analysis and Recognition*, pages 577–586. Springer, 2023.

[206] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

[207] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

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

[209] VQAonDB. Vqaondb dataset. https://ilocr.iiit.ac.in/vqabd/, 2024.

[210] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

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

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

[213] 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.

[214] 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.

[215] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10126–10135, 2020.

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

[217] 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.

[218] wendlerc. Renderedtext dataset. https://huggingface.co/datasets/wendlerc/RenderedText, 2024.

[219] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*, 2020.

[220] 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.

[221] Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv preprint arXiv:2407.17379*, 2024.

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

[223] 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.

[224] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European conference on computer vision*, pages 303–321. Springer, 2022.

[225] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024.

[226] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[227] 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.

[228] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint*, August 2024. URL https://arxiv.org/abs/2408.08872.

[229] 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.

[230] 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.

[231] 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.

[232] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2051–2060, 2019.

[233] 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.

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

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

[236] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

[237] 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.

[238] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

[240] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016.

[241] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

[242] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[243] Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. Icdar 2023 competition on structured text extraction from visually-rich document images. In *International Conference on Document Analysis and Recognition*, pages 536–552. Springer, 2023.

[244] Youngjoon Yu, Sangyun Chung, Byung-Kwan Lee, and Yong Man Ro. Spark: Multi-vision sensor perception and reasoning benchmark for large-scale vision-language models. *arXiv preprint arXiv:2408.12114*, 2024.

[245] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024.

[246] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34:509–521, 2019.

[247] 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.

[248] 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.

[249] 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.

[250] 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.

[251] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

[252] 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.

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

[254] Bo-Wen Zhang, Yan Yan, Lin Li, and Guang Liu. Infinitymath: A scalable instruction tuning dataset in programmatic mathematical reasoning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5405–5409, 2024.

[255] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019.

[256] 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.

[257] 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.

[258] 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.

[259] 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.

[260] 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.

[261] 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.

[262] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024.

[263] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *ICDAR*, pages 1577–1581, 2019.

[264] Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, et al. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*, 2024.

[265] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[266] 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.

[267] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022.

[268] Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. Robut: A systematic study of table qa robustness against human-annotated adversarial perturbations. *arXiv preprint arXiv:2306.14321*, 2023.

[269] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. *arXiv preprint arXiv:2402.14658*, 2024.

[270] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40 (6):1452–1464, 2017.

[271] 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.

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

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

[274] 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.

[275] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Chang, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[276] 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.

[277] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022.