WildDet3D: Scaling Promptable 3D Detection in the Wild

Anonymous CVPR submission

Paper ID *****

Figure 1. **Overview of WildDet3D.** Given a single RGB image and an optional depth map, WildDet3D performs open-vocabulary monocular 3D object detection with flexible prompt modalities—text, 2D point clicks, or 2D bounding boxes—and predicts full 3D bounding boxes for the specified objects. The model gracefully leverages additional geometric cues when available. To train WildDet3D for broad generalization, we also curate WildDet3D-Data, a large-scale in-the-wild dataset with approximately 1M human-verified samples spanning 13K categories.

Abstract

Understanding objects in 3D from a single image is a cornerstone of spatial intelligence. A key step toward this goal is monocular 3D object detection—recovering the extent, location, and orientation of objects from an input RGB image. To be practical in the open world, such a detector must generalize beyond closed-set categories, support diverse prompt modalities, and leverage geometric cues when available. Progress is hampered by two bottlenecks: existing methods are designed for a single prompt type and lack a mechanism to incorporate additional geometric cues, and current 3D datasets cover only narrow categories in controlled environments, limiting open-world transfer. In this work we address both gaps. First, we introduce ***WildDet3D****, a unified geometry-aware architecture that natively accepts text, point, and box prompts and can incorporate auxiliary depth signals at inference time. Second, we present* ***WildDet3D-Data****, the largest open 3D detection dataset to date, constructed through a* ***human-in-the-loop*** *pipeline in which crowd annotators select and rate the best 3D candidate per object, complemented by a VLM scorer aligned to*

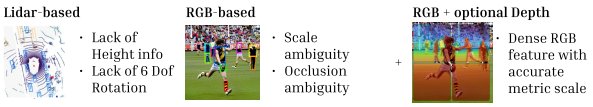


Figure 2. **Input modality comparison.** LiDAR lacks reliable height and 6-DoF rotation cues; RGB suffers from scale and occlusion ambiguity. RGB with *optional* depth retains rich semantics for open-vocabulary recognition while resolving metric scale when depth is present.

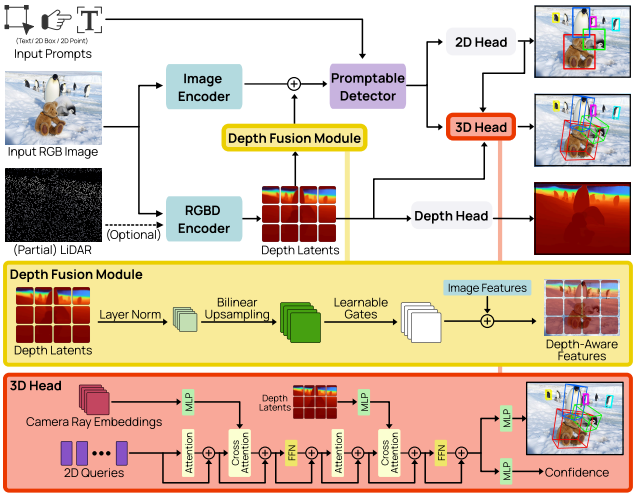


Figure 3. **Overview of WildDet3D.** Dual-vision encoders (Image + RGBD) extract visual and depth features in parallel. A depth fusion module injects depth latents into the image features, producing enriched queries that are combined with input prompts via the promptable detector. Outputs pass through cascaded 2D and 3D heads; depth latents are separately decoded by the auxiliary depth head.

these human judgments; this yields over 1M images across 13.5K categories in diverse real-world scenes. WildDet3D establishes a new state-of-the-art across multiple benchmarks and settings. In the open-world setting, it achieves 22.6/24.8 AP_{3D} on our newly introduced ***WildDet3D-Bench*** *with text and box prompts. On Omni3D, it reaches 34.2/36.4 AP_{3D} with text and box prompts, respectively. In zero-shot evaluation, it achieves 40.3/48.9 ODS on Argoverse 2 and ScanNet. Notably, incorporating depth cues at inference time yields substantial additional gains (+20.7 AP on average across settings).*

1. Introduction

Understanding objects in 3D is fundamental to spatial intelligence: robots, embodied agents, and AR/VR systems need not only to recognize objects but also to localize their extent, position, and orientation in 3D space. Despite rapid progress in open-vocabulary 2D recognition, monocular 3D object detection—recovering 3D boxes from a single RGB image—still lacks the generality needed for open-world use.

We argue that a truly general-purpose monocular 3D detector must satisfy three requirements [6, 52, 53, 57]: (i) generalize to long-tailed, open-ended categories in the wild; (ii) support multiple prompt modalities—text queries, 2D point clicks, and 2D bounding boxes—within a single architecture; and (iii) gracefully leverage extra geometric signals such as sparse LiDAR or partial depth when available. Prior work typically addresses only a subset: open-vocabulary methods [52, 53] focus on text queries, while oracle-prompt methods [45, 57] assume fixed geometric inputs. Neither provides a unified framework, nor accommodates additional depth at inference time.

On the model side, we introduce **WildDet3D**, a geometry-aware architecture that unifies text, point, box, and exemplar prompts. A key design choice is *RGB with optional depth* (Figure 2): LiDAR lacks height and 6-DoF cues, pure RGB suffers from scale and occlusion ambiguity, while RGB+depth preserves open-vocabulary semantics and degrades gracefully to monocular inference when depth is absent. WildDet3D reflects this through dual vision encoders, a depth fusion module, a promptable detector, and a 3D detection head trained with multi-task supervision (Section 2).

On the data side, we introduce **WildDet3D-Data**, a large-scale in-the-wild dataset built by applying multiple complementary methods to generate candidate 3D boxes for 2D annotations from diverse detection datasets [16, 26, 41, 46]. Crucially, final annotations are **chosen by humans**: crowd annotators select the best candidate per object and rate its quality, with quality control enforced by gold tasks (84–98% pass rate). Images beyond the human-annotated subset are handled by a VLM scorer *aligned to human judgments* (perfect Spearman correlation with human rejection rates), turning our pipeline into a scalable *human-in-the-loop* annotation system. This yields over 1M images across 13.5K categories—a 138× increase in category coverage over Omni3D—with broad open-world scene diversity (Section 3).

Empirically, WildDet3D achieves 22.6/24.8 AP_{3D} on WildDet3D-Bench (700+ categories) with text/box prompts, versus 2.3 AP for 3D-MOOD; 34.2/36.4 AP_{3D} on Omni3D with 10× fewer training epochs; and 40.3/48.9 ODS zero-shot on Argoverse 2 and ScanNet. Adding depth at inference time yields +20.7 AP on average (Section 4). Beyond benchmarks, WildDet3D powers real-world deployments including an iPhone app, Meta Quest 3 AR integration, robotic manipulation, and VLM-based spatial reasoning (Section 5).

2. WildDet3D

Given an RGB image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, optional intrinsics $\mathbf{K}$, optional partial or full depth $\mathbf{D}$, and a prompt $\mathcal{P}$, WildDet3D predicts 3D bounding boxes $\{\mathbf{B}_i\}$ with center $\mathbf{c}_i \in \mathbb{R}^3$ (meters), dimensions (w, h, l), rotation $\mathbf{R}_i \in \mathrm{SO}(3)$, and confidence $s_i \in [0, 1]$. The architecture (Figure 3) consists of three components: (i) *dual-vision encoders* with a *depth fusion module* for geometry-aware features (Section 2.1); (ii) a *promptable detector* unifying diverse prompt types (Section 2.2); and (iii) a *deeply-supervised 3D detection head* with unambiguous rotation normalization (Section 2.3). Auxiliary 2D detection and depth heads provide complementary supervision (Section 2.4).

2.1. Dual-vision encoder

A naive single encoder forces a trade-off between detection features (for open-vocabulary recognition) and metric depth features (for 3D reasoning). We instead use a *dual-encoder* design with a fusion stage.

Image encoder. A ViT-H with a SimpleFPN neck adopted from SAM 3 [8], initialized from a segmentation-pretrained checkpoint. The first 28 of 32 ViT blocks are frozen during training; SimpleFPN projects tokens to 256-channel feature maps.

RGBD encoder. A DINOv2 ViT-L/14 [34] taking 4-channel RGBD input at 686×686 (49×49 tokens), adopted from LingBot-Depth [44] (pretrained for metric depth). A ConvStack neck produces a 5-level pyramid from which depth latents $\mathbf{Z}_d \in \mathbb{R}^{C_d\times49\times49}$ ($C_d = 256$) are obtained via average pooling. The first 21 of 24 DINOv2 blocks are frozen. Training uses stochastic depth dropout (70% mono / 20% patch-masked / 10% full), enabling inference with or without depth. Using different backbones for the two encoders lets each specialize: segmentation-oriented for detection, depth-completion-oriented for geometry.

Depth fusion module. A ControlNet-style [58] zero-initialized residual that injects depth latents into visual features before the transformer encoder:

$$\mathbf{V}' = \mathbf{V} + \mathrm{Conv}_{1\times1}\big(\mathrm{LN}(\mathbf{Z}_d^{\uparrow})\big), \tag{1}$$

where $\mathbf{Z}_d^{\uparrow}$ is bilinearly upsampled to the visual resolution. The convolution is zero-initialized, so at training start $\mathbf{V}' = \mathbf{V}$; the depth branch is gradually learned without disrupting pretrained features.

2.2. Promptable detector

The promptable detector conditions the fused visual features on user-supplied prompts. It accepts four types:

- **Text**: a category name, selecting all instances.
- **Point**: one or more 2D coordinates with positive/negative labels, selecting the single object at that location [9, 12, 55].
- **Box**: a 2D bounding box, selecting the single object inside.
- **Exemplar**: a 2D box used as visual exemplar, detecting all similar objects.

Text is encoded with a CLIP-style tokenizer + 24-layer causal Transformer; box/point prompts use SAM 3's geometry encoder combining coordinate projection, ROI-aligned features, and sinusoidal positional encoding [8]. Encoded tokens serve as cross-attention memory for both the encoder and decoder. Training batches at the *per-prompt* level (every unique text category yields one batch entry), enabling fine-grained supervision and arbitrary categories per image. All four prompt types are jointly sampled. Full encoding details are in Appendix B.

2.3. Deeply-supervised 3D detection head

The 3D head lifts query features to 3D boxes through L Transformer decoder layers, each producing its own set of predictions supervised with equal-weight losses (*deep supervision*), which accelerates convergence and strengthens intermediate representations.

Multi-source aggregation. At each layer, query states $\mathbf{H}^l$ are enriched by two cross-attention modules. A *camera prompt* branch fuses per-pixel ray directions encoded via 8th-order real spherical harmonics: $\phi(\mathbf{r}) = \mathrm{RSH}_8(\mathbf{r}/\|\mathbf{r}\|) \in \mathbb{R}^{81}$, and a *depth prompt* branch fuses depth latents $\mathbf{Z}_d$. The two cross-attention modules have independent parameters per layer (see Appendix B).

3D box parameterization. A two-layer MLP regresses a 12-D encoding: 2D center offset $(\Delta c_x, \Delta c_y)$, log-depth $\hat{d}$, log-dimensions $(\hat{w}, \hat{h}, \hat{l})$, and a continuous 6D rotation $(r_1, \ldots, r_6)$ [60]. At inference, the 3D center is recovered by back-projecting the 2D offset at depth $d = \exp(\hat{d}/s_d)$. To resolve the inherent 4-fold rotation–dimension ambiguity of oriented 3D boxes, we apply an *unambiguous rotation normalization* (dimension ordering + yaw folding) to ground truth before loss computation; see Appendix B for the algorithm.

3D confidence. A parallel MLP predicts a 3D confidence $s_{3D} \in [0, 1]$ trained against a soft target that combines depth quality $q_{\text{depth}} = \exp(-|\log\hat{d} - \log d^*|)$ and IoU_{3D}. At inference the final score is $s = s_{2D} + \alpha \cdot s_{3D}$ with $\alpha = 0.5$, letting 3D confidence re-rank detections with similar 2D scores but different geometric quality.

2.4. Multi-task learning

During training each category in an image yields two branches: a *multi-target* query (50% text-only / 50% exemplar box) supervised against all instances, and a *single-target* geometric query (box or point) assigned to one selected instance, ensuring all prompt types receive supervision. We use one-to-many (O2M) matching [8] with top-k = 4 for denser supervision.

The total loss aggregates 3D regression, 3D confidence, auxiliary geometry (depth + camera rays), and 2D detection terms:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{conf}}}_{\text{3D detection}} + \underbrace{\mathcal{L}_{\text{geom}} + \mathcal{L}_{\text{2D}}}_{\text{auxiliary}}. \quad (2)$$

$\mathcal{L}_{\text{3D}}$ is an L1 on the 12-D encoding with per-component validity weights. $\mathcal{L}_{\text{conf}}$ is an IoU-aware focal BCE with adaptive soft targets. $\mathcal{L}_{\text{geom}}$ combines L1 + SILog [13] depth losses, an affine-invariant point-map loss [47], a confidence-mask BCE, and a camera-ray L2. $\mathcal{L}_{\text{2D}}$ combines IoU-aware BCE classification [8], L1+GIoU [38] box regression, and a per-category presence loss. Full formulations and loss weights are in Appendix B.

Ignore-region suppression. Because 3D annotations in Omni3D and WildDet3D-Data are not exhaustive (*e.g.*, heavily truncated or behind-camera objects are marked `IGNORE`), we suppress the negative classification loss for predictions whose 2D IoU with an ignore-annotated box exceeds 0.5. This aligns training with the evaluation protocol (ignored predictions counted as neutral), allowing confident detection regardless of 3D annotation availability.

3. WildDet3D-Data

Existing 3D detection datasets such as Omni3D [6] cover fewer than 100 categories in narrow domains, making open-world generalization difficult. Scaling 3D annotation is fundamentally harder than 2D: metric depth and calibrated intrinsics are costly to obtain at scale.

We introduce WildDet3D-Data, a large-scale open-vocabulary 3D detection dataset covering **1M images**, **3.7M annotations**, and **13.5K object categories**—a 138× increase over Omni3D. We repurpose mature 2D detection datasets (COCO [26], LVIS [16], Objects365 [41], V3Det [46]) by lifting their annotations into 3D through a three-stage pipeline (Figure 4): (i) five complementary models generate candidate 3D boxes for each 2D annotation; (ii) rule-based geometric and semantic filters remove implausible candidates; (iii) the best candidate is selected by human annotators or by a VLM-based scorer fine-tuned on Omni3D. Qualitative examples appear in Figure 5; Appendix C details each stage.

3.1. Pipeline overview

Candidate generation. For each image, we apply 4× super-resolution [56], estimate metric depth with MoGe-2 [47], and recover camera pose/intrinsics with PerspectiveFields [18] and WildCamera [61]. Five complementary methods then produce candidate 3D boxes per 2D annotation: **3D-MOOD** [52] (matched open-vocabulary predictions), **DetAny3D** [57] (direct regression), **SAM-3D** [45] (mesh reconstruction), **RANSAC-PCA** (geometric fitting with gravity alignment), and **LabelAny3D** [54] (analysis-by-synthesis). Each candidate is refined by depth-based translation optimization and PCA-based rotation alignment; details in Appendix C.

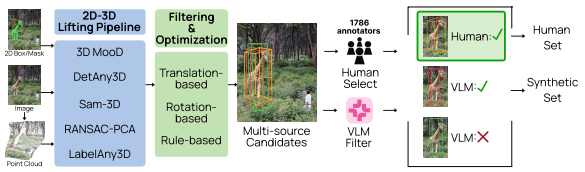


Figure 4. **WildDet3D-Data pipeline.** Five complementary models generate candidate 3D boxes from an image with 2D boxes and a depth-derived point cloud. After translation/rotation optimization and rule-based filtering, candidates enter two parallel selection branches: VLM scoring on six perceptual criteria and human annotation (select best + rate quality).

Filtering. Candidates pass through multi-stage filtering: (i) geometric rules (edge contact, occlusion, 3D-to-2D projection ratio); (ii) a VLM-based classifier (Qwen3.5-9B [2]) that removes *depicted* objects (pictures, reflections, screens) and composite images; (iii) GPT-4.1-mini-estimated [33] per-category physical dimensions filtering absolute size, depth-to-width ratio, and axis proportions; and (iv) a small-object upgrade that re-evaluates small objects via VLM criteria. Annotations failing any check are flagged `ignore3D=1` rather than deleted, preserving 2D recall. Full rules are in Appendix C.

Candidate selection. Final annotations come from two complementary paths. *Human selection* via Prolific [35]: annotators view four viewpoints per candidate (image overlay + three orthographic point-cloud views), *select the best candidate*, and *rate its quality* (`good_fit`/`acceptable`/`unacceptable`); quality control uses gold tasks, yielding 84–98% pass rates. *VLM selection* uses a Molmo2 [10] checkpoint fine-tuned on Omni3D-generated positive/negative pairs; it scores each candidate on six perceptual criteria (category, scale, translation, shape, rotation, vertical tilt) for a maximum total of 11, retaining the highest-scoring candidate above threshold. VLM scoring exhibits a monotonic correlation with human judgment (Spearman $\rho = -1.0$) and its top-2 covers 73.4% of human selections; detailed validation is in Appendix C.

3.2. Statistics

Table 1 summarizes WildDet3D-Data. The human-annotated portion (~103K images) has quality ratings: 35–48% `good_fit`, 33–50% `acceptable`, and 24–39% `unacceptable` (flagged as ignore). The VLM-filtered portion adds ~896K images. Images span three macro-scenes (Indoor 52%, Urban 32%, Nature 15%). Of 881 val categories, 826 (99.9%) have ≥ 1 training annotation

Figure 5. **Qualitative examples from WildDet3D-Data.** Each pair shows 3D annotations overlaid on the image (left) and rendered in the reconstructed point cloud (right). The dataset spans indoor, outdoor, and in-the-wild scenes with diverse categories, scales, and layouts.

Table 1. **WildDet3D-Data statistics.** Human annotations are rated by crowd-source workers; synthetic annotations are auto-selected by VLM scoring. Combined, WildDet3D-Data spans 13.5K categories—a 138× increase over Omni3D's 98.

Split	Source	Images	Ann.	Categories	Type	Scene	Max depth
Existing datasets							
Omni3D [6]	KITTI, nuSc., SUNRGBD, etc.	234K	3M+	98	Human	Driving, Furniture	67 m
COCO-3D [54]	COCO	18K	92K	80	Synthetic	In-the-wild	35 m
CA-1M [21]	ARKitScenes	3,500 (videos)	400K	Class-agnostic	Human	Indoor	5 m
WildDet3D-Data							
Train (Human)	COCO, LVIS, Obj365, V3Det	102,979	229,934	12,064	Human	In-the-wild	
Train (Synthetic)	COCO, LVIS, Obj365, V3Det	896,004	3,483,292	11,896	VLM filter	In-the-wild	
Val	COCO, LVIS, Obj365	2,470	9,256	785	Human	In-the-wild	
Test	COCO, LVIS, Obj365	2,433	5,596	633	Human	In-the-wild	
WildDet3D-Data (total)		**1,003,886**	**3,728,078**	**13,499**	Human + VLM	In-the-wild	**81 m**

and ~820 have ≥ 3. Candidate-model distribution, scene breakdown, and additional statistics appear in Appendix C.

4. Experiments

We evaluate WildDet3D on our proposed in-the-wild benchmark WildDet3D-Bench (Section 4.2), the standard Omni3D benchmark (Section 4.3), and zero-shot transfer to Argoverse 2 and ScanNet (Section 4.4), followed by ablation studies (Section 4.5). Evaluation on Stereo4D (real stereo depth), qualitative comparisons under text prompts, per-dataset ablation breakdowns, and training details appear in Appendix D.

4.1. Experimental setup

Datasets. We use (i) **WildDet3D-Bench** (ours), 700+ open-vocabulary categories with human-verified 3D annotations, split into rare (<5), common (5–20), and frequent (>20) groups; (ii) **Omni3D** [6], unifying KITTI [15], nuScenes [7], SUNRGBD [42], Hypersim [39], ARKitScenes [4], and Objectron [1]; and (iii) **zero-shot** evaluation on Argoverse 2 [50] and ScanNet [11] following 3D-MOOD [52].

Evaluation. We test in two modes: *text prompt* (category names as open-vocabulary queries) and *box prompt* (GT 2D boxes as geometric prompts, isolating 3D regression). For Omni3D we follow the standard protocol (AP_{3D} at 3D IoU $[0.05:0.50:0.05]$), treating ignored objects as neutral. For zero-shot we report ODS, the Open Detection Score of 3D-MOOD [52]. For WildDet3D-Bench (open-vocabulary, non-exhaustive), we report AP_{3D} using center-distance matching [52] with thresholds $[0.50:1.00:0.05]$, applying the LVIS federated protocol [16] so predictions matching 2D-only annotations are neutral.

Implementation. WildDet3D is trained in three stages with AdamW [29] on 32 GPUs (effective batch 128). Stage 1: 12 epochs on Omni3D. Stage 2: 12 epochs on a mixture of Omni3D, WildDet3D-Data, and supplementary 3D datasets (CA-1M [21], Waymo [43], 3EED [24], FoundationPose [49]), collectively "Others", for geometric diversity.

Stage 3: 3 epochs fine-tuning on Omni3D + WildDet3D-Data (human) with mask-guided point/box training. Inputs are resized to 1008×1008; inference uses per-category NMS at IoU 0.6. Full hyperparameters are in Appendix D.

4.2. In-the-wild evaluation on WildDet3D-Bench

Table 2 shows that WildDet3D trained on Omni3D alone already surpasses 3D-MOOD by 3.0× on text prompts (6.8 vs 2.3). Adding WildDet3D-Data raises performance to **22.6** (text) / **24.8** (box)—a 9.8× improvement over 3D-MOOD. GT depth at test time yields dramatic gains: text AP jumps from 6.8 to 20.7 for the Omni3D-only model, and from 22.6 to **41.6** for the full model, confirming that our architecture effectively leverages depth signals. Improvements are consistent across frequency splits, with the largest gains on rare categories (47.4 vs 2.4 for 3D-MOOD).

4.3. Results on Omni3D

Table 3 reports Omni3D results. WildDet3D achieves **34.2** AP with text prompts (+5.8 over 3D-MOOD) with 10× fewer training epochs (12 vs 120), and **36.4** in the box setting (+2.0 over DetAny3D with 6.7× fewer epochs). Gains are largest on indoor datasets (ARKitScenes, Objectron), reflecting stronger geometry estimation in cluttered scenes. Sparse depth further pushes box-prompt AP to **45.8** (+11.4 over DetAny3D), with dramatic gains on depth-equipped indoor datasets.

4.4. Zero-shot evaluation

Zero-shot results. WildDet3D reaches **40.3** ODS on AV2 and **48.9** on ScanNet, outperforming 3D-MOOD Swin-B by +16.5 and +17.4 ODS; AP is dramatically higher (43.4 vs 14.8 on AV2; 56.5 vs 28.8 on ScanNet). GT depth helps more on ScanNet (48.9 → 50.2) where indoor depth resolves scale ambiguity; the AV2 gain is marginal.

4.5. Ablation

Table 5 reports the ablation. The single most critical choice is joint 2D+3D prediction: removing the 2D head and predicting 3D directly collapses AP from 30.2 to 11.1 (−19.1), confirming that 2D detection provides essential spatial priors. Among training objectives, O2M matching is most impactful (−2.5), especially on dense driving scenes; explicit geometry supervision contributes −1.7, concentrated on indoor datasets. The 3D confidence head, deep supervision, and ignore-aware suppression each give smaller but consistent improvements.

4.6. Qualitative results

Figure 6 compares WildDet3D against OVMono3D [53] and DetAny3D [57] on four in-the-wild scenes using box prompts. WildDet3D produces tighter, better-oriented 3D boxes on outdoor animals, cluttered indoor desks, street scenes with varied depths, and fine-grained food items, where competing methods either hallucinate large boxes or mis-estimate orientation.

5. Applications

Beyond benchmark evaluation, we demonstrate WildDet3D across a range of real-world deployment scenarios spanning on-device mobile inference, AR headsets, robotic manipulation, and vision-language model integration.

WildDet3D in your pocket. We deploy WildDet3D on iPhone via a client-server architecture where the iPhone captures RGB frames and LiDAR depth via ARKit and streams them to a cloud-hosted inference endpoint (Figure 7a). The app supports open-vocabulary text queries, 2D bounding box prompts for geometric detection, and real-time camera-based inference; detected 3D boxes are rendered as AR overlays anchored via ARKit world tracking, enabling interactive 3D perception on consumer hardware. The app is publicly available on the App Store.

WildDet3D for Augmented Reality (AR). We integrate WildDet3D with Meta Quest 3 (Figure 7b). A Unity client captures passthrough camera frames with calibrated intrinsics and 6-DoF pose from the Quest's tracking system, sends them to the WildDet3D API, and renders detected 3D bounding boxes as overlays in the passthrough view. This enables spatial understanding for AR where users query objects in their environment by category and see metric 3D boxes anchored in physical space.

WildDet3D for Robotics. We apply WildDet3D to robotic manipulation with a Franka Emika Panda arm (Figure 7c). A third-view camera captures the scene, and WildDet3D produces open-vocabulary 3D bounding boxes that are transformed into the robot's coordinate frame. The predicted box centers and dimensions are directly consumed for grasp pose generation and fed to an IK-based interpolation planner—a zero-shot alternative to task-specific 3D perception modules that require per-object training or CAD models.

6. Related work

Monocular 3D object detection. Monocular 3D detection recovers 3D boxes from a single RGB image, an ill-posed problem due to scale ambiguity and missing geometric cues. Early work focused on closed-set driving [5, 28, 37, 48, 59] and indoor scenes [11, 22, 40, 42]. Omni3D [6] unified multiple datasets into a cross-domain benchmark, and UniMODE [25] further improved unified monocular 3D detection. Recent work extends toward open-set/open-vocabulary scenarios: OVMono3D [53], 3D-MOOD [52], OVM3D-Det [17], and LocateAnything3D [30] explore lifting open-vocabulary 2D detections into 3D, while DetAny3D [57] emphasizes promptable box-based prediction. These approaches specialize to a single prompt interface; in con-

Table 2. **WildDet3D-Bench evaluation.** (1) WildDet3D outperforms baselines when trained on Omni3D alone; (2) WildDet3D-Data improves performance substantially (6.8 → 22.6 text, 8.4 → 24.8 box); (3) depth input at test time nearly doubles performance.

Method	Training data	AP_{rare}	AP_{common}	$AP_{frequent}$	AP_{3D}
Text Prompt					
3D-MOOD [52]	Omni3D	2.4	2.1	2.6	2.3
WildDet3D	Omni3D	9.0	6.5	5.2	6.8
WildDet3D w/ depth	Omni3D	23.0	21.5	16.1	20.7
WildDet3D	Omni3D, Others, WildDet3D-Data	28.3	21.6	18.7	22.6
WildDet3D w/ depth	Omni3D, Others, WildDet3D-Data	**47.4**	**40.7**	**37.2**	**41.6**
Box Prompt					
OVMono3D-LIFT [53]	Omni3D	7.4	8.8	5.1	7.7
DetAny3D [57]	Omni3D, Others	9.9	7.4	6.3	7.8
WildDet3D	Omni3D	12.0	7.9	5.3	8.4
WildDet3D w/ depth	Omni3D	26.4	24.4	19.6	23.9
WildDet3D	Omni3D, Others, WildDet3D-Data	30.0	24.2	20.3	24.8
WildDet3D w/ depth	Omni3D, Others, WildDet3D-Data	**53.7**	**46.1**	**42.5**	**47.2**

Table 3. **Omni3D evaluation.** WildDet3D outperforms prior methods in both text and box settings and further improves with depth input.

Method	KITTI [15]	nuScenes [7]	SUNRGBD [42]	Hypersim [39]	ARKitScenes [4]	Objectron [1]	AP_{3D}
Text Prompt							
Cube R-CNN [6]	32.6	30.1	15.3	7.5	41.7	50.8	23.3
Uni-MODE* [25]	29.2	**36.0**	23.0	8.1	48.0	66.1	28.2
3D-MOOD Swin-T [52]	32.8	31.5	21.9	10.5	51.0	64.3	28.4
3D-MOOD Swin-B [52]	31.4	35.8	23.8	9.1	53.9	67.9	30.0
WildDet3D	**37.0**	31.7	38.9	16.5	64.6	60.5	34.2
WildDet3D w/ depth	36.1	32.0	**51.1**	**26.6**	**73.3**	**68.3**	**41.6**
Box Prompt							
OVMono3D-LIFT [53]	31.4	32.5	23.2	11.9	54.2	63.5	29.6
DetAny3D [57]	38.7	**37.6**	46.1	16.0	50.6	56.8	34.4
WildDet3D	**44.3**	35.3	43.1	17.3	66.6	60.8	36.4
WildDet3D w/ depth	42.8	35.9	**58.7**	**30.4**	**76.6**	**68.5**	**45.8**

Table 4. **Zero-shot evaluation** from Omni3D to Argoverse 2 and ScanNet. ODS [52] combines AP with translation/scale/orientation errors (higher is better). Full metric breakdown in Appendix D.

	AV2 [50]		ScanNet [11]	
Method	AP↑	ODS↑	AP↑	ODS↑
Cube R-CNN [6]	8.6	8.9	20.0	19.5
3D-MOOD Swin-T [52]	14.8	22.5	27.3	30.2
3D-MOOD Swin-B [52]	14.7	23.8	28.8	31.5
WildDet3D	43.4	40.3	56.5	48.9
WildDet3D w/ depth	**43.4**	**40.4**	**57.6**	**50.2**

Table 5. **Ablation** on Omni3D with box prompts. Joint 2D+3D detection is critical; O2M matching and geometry loss also contribute substantially. Per-dataset breakdown in Appendix D.

Configuration	AP_{3D}
Full model	30.2
w/o 2D head (3D only)	11.1 (−19.1)
w/o 3D confidence head	29.4 (−0.8)
w/o O2M matching	27.7 (−2.5)
w/o geometry loss	28.5 (−1.7)
w/o deep supervision	29.9 (−0.3)
w/o ignore-aware suppression	30.0 (−0.2)

trast, we target a unified setting accepting text, point, or box prompts, with optional depth cues available at test time.

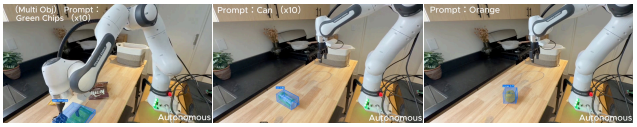
Open-vocabulary and promptable visual perception. Our work builds on progress in open-vocabulary 2D perception (GLIP [23], OWL-ViT [31, 32], Grounding DINO [27]) and promptable segmentation (SEEM [62], SAM 3 [8]), as well as multimodal LLMs with grounding capability (LISA [3, 20, 51], Molmo [9, 10, 12]). Existing 2D systems already support text, clicks, or boxes flexibly, but comparable flexibility is rare in monocular 3D detection, which typically exposes only category queries or externally provided 2D boxes. Our goal is to bring this prompt flexibility into 3D while preserving open-vocabulary recognition and enabling graceful improvement with depth.

3D annotation pipelines and open-world 3D data. Data is a major bottleneck for generalized 3D detection. Omni3D [6]

Figure 6. **Qualitative comparison on in-the-wild images with box prompts.** Each block shows a scene detected by three models with 2D box prompts: ground truth 3D boxes, WildDet3D, OVMono3D, and DetAny3D. WildDet3D produces more accurate 3D localization and tighter boxes across diverse categories (animals, vehicles, indoor electronics, small food items). Text-prompt comparisons are in Appendix D.

(a) Mobile APP (iPhone): text prompt detection in an office (left), 2D box prompt for geometric detection (middle), and open-vocabulary animal detection outdoors (right).

(b) Augmented Reality (Meta Quest 3): passthrough AR with 3D bounding boxes rendered in real time across three desk scenes.



(c) Robotics: Franka Emika Panda autonomously grasping objects specified by open-vocabulary text prompts (“Green Chips”, “Can”, “Orange”).

Figure 7. **Real-world deployment demos.** Each row shows three frames from a different deployment platform, demonstrating WildDet3D across diverse interaction modes and environments.

offers a multi-dataset benchmark but with limited vocabulary. LabelAny3D [54] introduces analysis-by-synthesis 3D box annotation and the COCO3D benchmark, and 3D-MOOD [52] and SAM-3D [45] demonstrate 2D-to-3D lifting and reconstruction-based annotation. Automatic annotations remain noisy on scale, rotation, and extent. Our pipeline combines multiple complementary candidate generators, VLM-based scoring, human selection, and geometry-aware filtering to provide large-scale in-the-wild open-vocabulary 3D supervision.

7. Conclusion

We presented WildDet3D, a geometry-aware monocular 3D detector unifying text, point, and box prompts with optional depth, and WildDet3D-Data, a human-in-the-loop 1M-image, 13.5K-category dataset (138× Omni3D). WildDet3D sets new state-of-the-art on Omni3D (34.2/36.4 AP_{3D} text/box) with 6–10× fewer epochs, reaches 40.3/48.9 ODS zero-shot on Argoverse 2 / ScanNet, and 22.6/24.8 AP on 700+ in-the-wild categories, with +20.7 AP on average when depth is provided.

References

[1] 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. *CVPR*, 2021. 5, 7, 17

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 4, 14

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *NeurIPS*, 2024. 7

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data, 2022. 5, 7, 17

[5] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, 2019. 6

[6] 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, 2023. 2, 4, 5, 6, 7

[7] 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, 2020. 5, 7, 17

[8] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. 3, 4, 7, 13, 14

[9] Christopher Clark, Yue Yang, Jae Sung Park, Zixian Ma, Jieyu Zhang, Rohun Tripathi, Mohammadreza Salehi, Sangho Lee, Taira Anderson, Winson Han, et al. Molmopoint: Better pointing for vlms with grounding tokens. *arXiv preprint arXiv:2603.28069*, 2026. 3, 7

[10] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. 4, 7, 15

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5, 6, 7

[12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, 2025. 3, 7

[13] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 4, 14

[14] Ziqi Gao, Jieyu Zhang, Wisdom Oluchi Ikezogwo, Jae Sung Park, Tario G. You, Daniel Ogbu, Chenhao Zheng, Weikai Huang, Yinuo Yang, Winson Han, Quan Kong, Rajat Saini, and Ranjay Krishna. Synthetic visual genome 2: Extracting large-scale spatio-temporal scene graphs from videos, 2026. 16

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5, 7, 17

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2, 4, 5

[17] Rui Huang, Henry Zheng, Yan Wang, Zhuofan Xia, Marco Pavone, and Gao Huang. Training an open-vocabulary monocular 3d detection model without 3d data. In *NeurIPS*, 2024. 6

[18] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration, 2023. 4, 14

[19] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos. In *CVPR*, 2025. 16

[20] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 7

[21] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection, 2024. 5

[22] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *CVPR*, 2025. 6

[23] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 7

[24] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, Youquan Liu, Dongyue Lu, Liang Pan, Lingdong Kong, Junwei Liang, and Ziwei Liu. 3eed: Ground everything everywhere in 3d, 2025. 5

[25] Zhuoling Li, Xiaogang Xu, SerNam Lim, and Hengshuang Zhao. Unimode: Unified monocular 3d object detection. In *CVPR*, 2024. 6, 7

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 4

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 7

[28] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPR Workshop*, 2020. 6

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[30] Yunze Man, Shihao Wang, Guowen Zhang, Johan Bjorck, Zhiqi Li, Liang-Yan Gui, Jim Fan, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Locateanything3d: Vision-language 3d detection with chain-of-sight, 2026. 6

[31] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 7

[32] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 2023. 7

[33] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah

Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 4, 14

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 3

[35] Prolific. Prolific academic online research, 2025. Accessed: March 20, 2026. 4, 14

[36] 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, 2021. 13

[37] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021. 6

[38] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019. 4, 14

[39] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 5, 7, 17

[40] 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. 6

[41] 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*, 2019. 2, 4

[42] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5, 6, 7, 17

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[44] Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026. 3

[45] SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images, 2025. 2, 4, 8

[46] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset, 2023. 2, 4

[47] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025. 4, 14

[48] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *CVPR*, 2021. 6

[49] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects, 2024. 5

[50] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting, 2023. 5, 7

[51] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 7

[52] Yung-Hsu Yang, Luigi Piccinelli, Mattia Segu, Siyuan Li, Rui Huang, Yuqian Fu, Marc Pollefeys, Hermann Blum, and Zuria Bauer. 3d-mood: Lifting 2d to 3d for monocular open-set object detection, 2025. 2, 4, 5, 6, 7, 8, 16

[53] Jin Yao, Hao Gu, Xuweiyi Chen, Jiayun Wang, and Zezhou Cheng. Open vocabulary monocular 3d object detection, 2025. 2, 6, 7, 16

[54] Jin Yao, Radowan Mahmud Redoy, Sebastian Elbaum, Matthew B. Dwyer, and Zezhou Cheng. Labelany3d: Label any object 3d in the wild, 2026. 4, 5, 8

[55] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. In *CoRL*, 2024. 3

[56] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion, 2025. 4, 14

[57] 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, 2025. 2, 4, 6, 7, 16

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[59] 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. 6

[60] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. 3, 13

[61] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *NeurIPS*, 2023. 4, 14

[62] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. 7

A. Limitations

WildDet3D has several limitations. First, predicted *camera intrinsics* are less accurate than ground-truth calibration, degrading metric 3D localization when intrinsics are unavailable. Second, monocular 3D detection is inherently ill-posed: distant or heavily occluded objects remain challenging without depth input, as evidenced by the substantial gains from sparse depth (Table 3). Third, *rotation prediction* remains the weakest component, particularly for near-symmetric objects (round tables, square boxes) where visual orientation cues are ambiguous. Fourth, the *dual-backbone* design increases memory and compute versus 2D-only detectors, limiting real-time on-device inference without distillation. Fifth, performance on rare long-tail categories still exhibits higher variance than frequent ones. Finally, the applications demonstrated in Section 5 are research prototypes—predictions may contain errors without guaranteed bounds—and WildDet3D is *not intended for safety-critical applications*.

B. Method details

B.1. Prompt encoding

We adopt SAM 3's prompt encoding [8]. *Text prompts* are tokenized with a CLIP-style [36] BPE tokenizer and encoded by a 24-layer causal text Transformer (width 1024, 16 heads), then linearly projected to $d = 256$. *Box and point prompts* are encoded by a geometry encoder that sums three complementary representations: (1) a direct linear projection of coordinates, (2) ROI-aligned features pooled from the image backbone (for boxes) or grid-sampled features (for points), and (3) sinusoidal positional encoding. A learnable positive/negative label embedding is added, and the result is refined by a 3-layer Transformer with cross-attention to image features. *Exemplar prompts* reuse the box encoding pipeline but are differentiated by a special text token ("visual") and a multi-target matching strategy that assigns all instances of the same category as ground truth. Encoded text and geometry tokens are concatenated into a single prompt sequence, which serves as cross-attention memory in both the encoder and decoder stages.

B.2. Multi-source aggregation in 3D head

For each decoder layer $l \in \{1, \dots, L\}$, hidden states $\mathbf{H}^l \in \mathbb{R}^{S\times d}$ are sequentially enriched with camera geometry and depth through two cross-attention modules. First, a *camera prompt* branch incorporates spatial ray features. Given intrinsics $\mathbf{K}$, we generate per-pixel rays $\mathbf{r}_{i,j} = \mathbf{K}^{-1}[u, v, 1]^\top$ and encode them with 8th-order real spherical harmonics $\phi(\mathbf{r}) = \mathrm{RSH}_8(\mathbf{r}/\|\mathbf{r}\|) \in \mathbb{R}^{81}$. Ray features are fused via cross-attention:

$$\tilde{\mathbf{H}}^l = \mathrm{FFN}\Big(\mathrm{CrossAttn}\big(\mathrm{SelfAttn}(\mathbf{H}^l),\ f_r(\phi(\mathbf{r}))\big)\Big), \quad (3)$$

where $f_r : \mathbb{R}^{81} \to \mathbb{R}^d$ projects the SH features. A *depth prompt* branch then fuses depth latents:

$$\hat{\mathbf{H}}^l = \mathrm{FFN}\Big(\mathrm{CrossAttn}\big(\mathrm{SelfAttn}(\tilde{\mathbf{H}}^l),\ f_d(\mathbf{Z}_d)\big)\Big), \quad (4)$$

with $f_d : \mathbb{R}^{C_d} \to \mathbb{R}^d$ projecting depth latents into the query embedding space. Both cross-attention modules use a single head, with independent parameters per decoder layer.

B.3. 3D box parameterization

Following Equation (5), the 12-D encoding is:

$$\mathbf{p}_{3\mathrm{d}} = \big(\underbrace{\Delta c_x, \Delta c_y}_{\text{center offset}},\ \underbrace{\hat{d}}_{\text{log depth}},\ \underbrace{\hat{w}, \hat{h}, \hat{l}}_{\text{log dims}},\ \underbrace{r_1, \dots, r_6}_{\text{rotation}}\big). \quad (5)$$

The components are: (i) center offset $(\Delta c_x, \Delta c_y)$—displacement between the 2D projection of the 3D center and the 2D box center, normalized by $s_c = 10$; (ii) log-depth $\hat{d} = s_d \cdot \log(d)$ with $s_d = 2.0$; (iii) log-dimensions $(\hat{w}, \hat{h}, \hat{l}) = s_{\mathrm{dim}} \cdot \log(w, h, l)$ with $s_{\mathrm{dim}} = 2.0$; (iv) the first two rows of the 3×3 rotation matrix in continuous 6D form [60], with Gram–Schmidt orthogonalization to recover the full rotation.

B.4. Unambiguous rotation normalization

Oriented 3D boxes have an inherent rotation ambiguity: a box with dimensions (w, h, l) rotated by yaw θ is geometrically identical to one with swapped (l, h, w) rotated by $\theta + 90^\circ$, and a 180° yaw flip yields the same box for symmetric objects. We resolve this ambiguity with a two-step normalization applied to ground-truth rotation and dimensions before loss computation:

1. **Dimension ordering.** If $w > l$, swap (w, l) and rotate by $R_y(90^\circ)$ so that $w \le l$ always holds.
2. **Yaw folding.** Fold the yaw angle into $[0, \pi)$ by applying $R_y(180^\circ)$ when yaw < 0 or yaw $\ge \pi$.

Together these reduce the 4-fold rotation–dimension ambiguity to a unique unambiguous form, yielding a one-to-one mapping between box geometry and regression target. The same normalization is applied to predictions at inference time before evaluation.

B.5. 3D confidence

The 3D confidence branch is a two-layer MLP predicting a scalar score $s_{3\mathrm{D}} \in [0, 1]$. The soft target combines depth quality and 3D IoU:

$$q^* = \beta \cdot q_{\mathrm{depth}} + (1-\beta) \cdot \mathrm{IoU}_{3\mathrm{D}}, \quad q_{\mathrm{depth}} = \exp(-|\log \hat{d} - \log d^*|), \quad (6)$$

with $\beta = 0.7$ emphasizing depth accuracy (the primary bottleneck in monocular 3D detection). At inference the final score is

$$s = s_{2\mathrm{D}} + \alpha \cdot s_{3\mathrm{D}}, \quad \alpha = 0.5, \quad (7)$$

letting 3D confidence re-rank detections with similar 2D scores but different geometric quality.

B.6. Loss formulations

3D regression loss $\mathcal{L}_{\text{3D}}$. An L1 loss on the 12-D encoding:

$$\mathcal{L}_{\text{3D}} = \frac{1}{N_{\text{pos}}} \sum_{i \in \mathcal{M}} \sum_{k} w_k \left| p_k^{(i)} - p_k^{*(i)} \right|, \tag{8}$$

where $\mathcal{M}$ is the matched index set and w_k are per-component validity weights (set to zero when depth or dimensions are unavailable).

3D confidence loss $\mathcal{L}_{\text{conf}}$. An IoU-aware focal BCE with adaptive soft targets. For each matched prediction with raw logit c_i, the target is

$$t_i = \sigma(c_i)^{\alpha} \cdot q_i^{*\,1-\alpha}, \quad \alpha = 0.25, \tag{9}$$

and the total loss combines a positive term over matched queries and a focal-weighted negative term over unmatched ones:

$$\mathcal{L}_{\text{conf}} = \frac{1}{N_{\text{pos}}} \sum_{i \in \mathcal{M}} w_+ \cdot \text{BCE}(c_i, t_i) + \frac{1}{N_{\text{neg}}} \sum_{j \notin \mathcal{M}} \sigma(c_j)^{\gamma} \cdot \text{BCE}(c_j, 0), \tag{10}$$

with $w_+ = 5$ and focal exponent $\gamma = 2$ to down-weight easy negatives.

Auxiliary geometry loss $\mathcal{L}_{\text{geom}}$. The geometry backend aggregates the following losses on predicted depth map and camera intrinsics: (i) **L1 metric depth** at valid pixels; (ii) **scale-invariant logarithmic (SILog) depth** [13] at valid pixels,

$$\mathcal{L}_{\text{SILog}} = \sqrt{\frac{1}{N} \sum_i g_i^2 - \frac{\lambda}{N^2} \left(\sum_i g_i \right)^2}, \quad g_i = \log \hat{d}_i - \log d_i^*, \tag{11}$$

with $\lambda = 0.85$; (iii) **confidence mask BCE** supervising per-pixel depth validity; (iv) **affine-invariant point-map losses** (global alignment, multi-scale local alignment, edge-aware) [47] on back-projected 3D point maps; (v) **camera ray L2 loss** supervising predicted intrinsics against ground truth.

Auxiliary 2D detection loss $\mathcal{L}_{\text{2D}}$. (i) **IoU-aware BCE** classification [8] with 2D IoU as soft target; (ii) **Box regression** combining L1 on center-size representation and generalized IoU [38]; (iii) **Per-category presence** loss supervising whether a queried category exists in the image; (iv) **One-to-many matching** [8]: each ground-truth object is paired with its top-$k = 4$ scoring predictions, providing denser gradient signals for both 2D and 3D heads.

C. Data pipeline details

C.1. Monocular depth and camera estimation

We first apply 4× image super-resolution [56] to increase spatial detail. MoGe-2 [47] then produces a metric depth map at 1024-long-edge resolution, while PerspectiveFields [18] and WildCamera [61] estimate camera pose (roll/pitch) and intrinsics (f_x, f_y, c_x, c_y) respectively. The depth map is reprojected into a 3D point cloud using the estimated camera parameters.

C.2. 3D box optimization

After initial prediction, each candidate undergoes two refinement steps: (i) *translation optimization*, which aligns the predicted depth to the estimated depth map using percentile-based scaling or anchor-based optimization; and (ii) *rotation optimization*, which corrects orientation using PCA-based gravity alignment and 2D projection constraints. The refined candidates from all five models are then merged into a unified 10-D format (center, dimensions, quaternion), yielding up to five candidates per 2D annotation.

C.3. Rule-based filtering

All candidates and annotations undergo multi-stage filtering. Failed annotations are *never deleted* but flagged as `ignore3D=1`, preserving the full 2D annotation set for 2D recall evaluation.

Geometric rules. Candidates are filtered by three geometric criteria: edge contact ratio ≥3%, occlusion ratio >15% (for RANSAC-PCA), and 3D-to-2D projection size ratio outside $[0.5, 1.5]$.

Depicted object filter. A Qwen3.5-9B [2] classifier identifies and discards annotations of *depicted* objects (pictures, posters, reflections, screen displays) that portray objects rather than real 3D instances.

Composite image filter. Images composed of multiple sub-images are detected with Qwen3.5-9B and removed, since they often yield inaccurate depth maps.

LLM-estimated size and geometry filter. GPT-4.1-mini [33] estimates per-category physical dimensions (axis ranges, depth-to-width bounds, fixed/variable size class). We filter by (i) **absolute size**—each axis must fall within the LLM-estimated range; fixed-size categories (person, car) use 1.5× tolerance, variable-size categories (toy, sculpture) use 3.0×, relaxed to 2.5×/5.0× for fine-grained datasets; (ii) **depth-to-width ratio**—catches depth-stretching artifacts; (iii) **axis proportion**—for non-flat, non-elongated objects.

Small object upgrade. Objects initially filtered as "small" (2D area < 0.5% of image) are re-evaluated using the VLM criteria; qualifying small objects are upgraded to valid annotations, recovering long-tail supervision.

C.4. Candidate selection

Human selection (Prolific). Crowdsourced annotators on Prolific [35] evaluate up to five candidates per object, each visualized from four viewpoints (perspective overlay

+ three orthographic point-cloud views). Annotators *select the best candidate* and *rate its quality* as `good_fit`, `acceptable`, or `unacceptable`. Each batch contains 50 regular tasks plus 5 gold (quality-control) tasks with known-bad annotations; batches where annotators fail to identify ≥ 2/5 gold tasks are discarded and reassigned. Overall pass rates range 84–98% across splits.

VLM selection. A Molmo2 [10] checkpoint is fine-tuned for candidate selection with synthetically generated positive and negative candidate pairs from Omni3D. Each candidate (cropped image overlaid with the projected 3D box wireframe) is scored on six perceptual criteria: *category correctness* (0–1), *scale accuracy*, *translation accuracy*, *shape fidelity*, *rotation correctness*, and *vertical tilt alignment* (each 0–2), for a maximum total of 11. We retain the highest-scoring candidate when its total score exceeds 10.

C.5. Pipeline validation

To validate the annotation pipeline, we analyze the human-annotated train split (230K accepted annotations).

Table 6. **Candidate model selection share and human rejection rate** on the human-annotated train set. Rejection rates vary by > 3× across models.

Model	Sel. Share	Rej. Rate
SAM-3D	40.4%	17.3%
RANSAC-PCA	28.2%	12.5%
DetAny3D	14.5%	42.9%
LabelAny3D	13.0%	21.3%
3D-MOOD	3.8%	25.7%
Overall	—	22.0%

Table 7. **VLM composite score vs. human rejection rate.** Scores correlate perfectly with human judgment (Spearman $\rho = -1.0$).

VLM Score	Rej. Rate	n
< 7	71.9%	1,992
7	67.4%	13,670
8	45.3%	18,665
9	36.1%	83,882
10	16.7%	310,329
11	9.2%	52,684

VLM Top-2 Coverage: **73.4%**

Candidate model quality. Table 6 shows the selection share and rejection rate of each candidate model. SAM-3D accounts for the largest share of accepted annotations (40.4%), followed by RANSAC-PCA (28.2%), DetAny3D (14.5%), LabelAny3D (13.0%), and 3D-MOOD (3.8%). Rejection rates vary by more than 3×: RANSAC-PCA achieves the lowest (12.5%) while DetAny3D is rejected most frequently (42.9%). This disparity confirms that candidate quality differs substantially across models and can only be reliably distinguished through human evaluation.

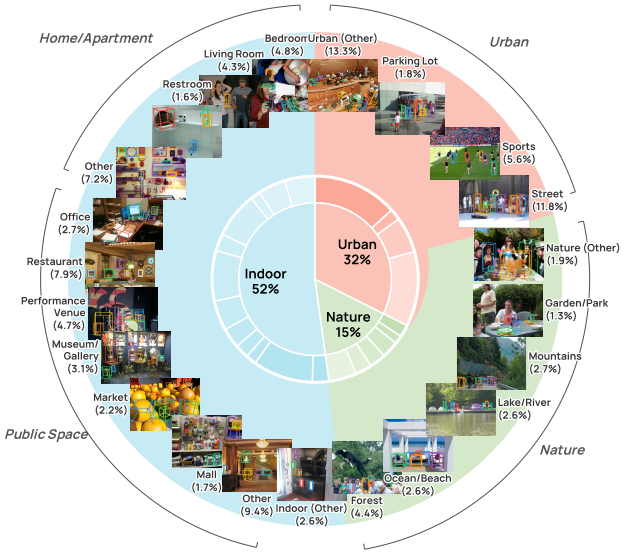


Figure 8. **Scene category distribution of WildDet3D-Data.** Images span three macro-categories: Indoor (52%), Urban (32%), and Nature (15%).

VLM-human correlation. VLM scores exhibit a perfect monotonic correlation with human rejection rates (Spearman $\rho = -1.0$; Table 7): rejection decreases steadily from 71.9% at score < 7 to 9.2% at score 11 (AUC = 0.66, point-biserial $r = 0.30$, $p < 10^{-100}$, $n = 481$K). Among the six VLM dimensions, *scale* (AUC = 0.60) and *shape* (AUC = 0.56) are the strongest predictors. VLM top-2 ranked candidates cover **73.4%** of human selections. Despite strong correlation, VLM scoring alone cannot substitute for human judgment: even at score 10 (64.5% of all candidates), the human rejection rate is 16.7%. This motivates our two-stage design: VLM scoring as pre-filter, followed by human verification for quality-critical subsets.

C.6. Additional statistics

Val/test sampling strategy. For validation and test sets, we use a three-phase balanced sampling algorithm: (1) greedy set cover for 100% category coverage, (2) multi-objective balanced fill optimizing category rarity, scene diversity, depth distribution, and source balance, and (3) targeted patching to ensure ≥ 3 samples per category.

Candidate model distribution. Among valid synthetic annotations, SAM-3D contributes ~55% of selected boxes, RANSAC-PCA ~28%, and LabelAny3D ~17%, reflecting the complementary strengths of mesh-based reconstruction,

geometric fitting, and single-image 3D reconstruction.

Filtering impact. The multi-stage filtering pipeline removes 15–20% of annotations via size/geometry checks, with the absolute size filter contributing the largest share. The depiction filter catches ~2% of annotations across human and synthetic splits.

D. Additional experiments

D.1. Stereo4D evaluation with real depth

To further validate generalization with real depth, we evaluate on Stereo4D [19], a video dataset with real stereo depth maps (383 images, 78 categories after filtering), with 2D annotations collected using the SVG2 pipeline [14]. Categories are split into rare (<5), common (5–10), and frequent (≥10) groups. AP is computed with center-distance matching. All models are evaluated zero-shot (not trained on Stereo4D).

Table 8. **Stereo4D zero-shot evaluation** with box prompts.

Method	AP_{rare}	AP_{common}	$AP_{frequent}$	AP_{3D}
OVMono3D-LIFT [53]	<u>12.3</u>	7.1	<u>11.4</u>	<u>9.9</u>
DetAny3D [57]	8.3	<u>8.2</u>	4.9	7.1
WildDet3D	8.1	6.3	8.5	7.5
WildDet3D w/ depth	**26.2**	**31.1**	**24.6**	**27.7**

Without depth (Table 8), our monocular model (7.5 AP) is competitive with DetAny3D (7.1 AP), while OVMono3D-LIFT achieves the highest monocular AP (9.9) due to stronger monocular depth estimation on this low-resolution stereo domain. When real depth is provided, performance improves dramatically to **27.7 AP**, a 2.8× improvement over OVMono3D-LIFT, demonstrating effective use of real depth signals.

D.2. Per-dataset ablation

Table 9 expands the ablation from Section 4.5 with per-dataset breakdowns. The most dramatic drop from removing the 2D head is on indoor datasets (SUNRGBD: 33.9→5.1, Objectron: 56.8→10.9). The O2M matching effect is most pronounced on driving datasets (KITTI: 27.9→23.2, nuScenes: 28.2→23.9). Geometry loss contributes most on indoor scenes (SUNRGBD: 33.9→28.6, Hypersim: 13.2→11.1). We expect ignore-aware suppression to have a larger effect on WildDet3D-Bench than Omni3D, since partial 3D annotations are far more common there.

D.3. Qualitative comparison with text prompts

Figure 9 compares WildDet3D against 3D-MOOD [52] on four in-the-wild scenes under text-prompt evaluation. WildDet3D consistently detects more categories and produces more realistic 3D placements.

Table 9. **Full ablation with per-dataset breakdown.** Evaluated in the oracle (box-prompt) setting on Omni3D.

Configuration	KITTI [15]	nuScenes [7]	SUNRGBD [42]	Hypersim [39]	ARKitScenes [4]	Objectron [1]	AP_{3D}
Full model	27.9	28.2	33.9	13.2	59.4	56.8	30.2
w/o 2D head (3D only)	18.3	15.6	5.1	9.7	28.5	10.9	11.1 (−19.1)
w/o 3D confidence head	28.0	27.9	32.1	13.0	58.2	56.9	29.4 (−0.8)
w/o O2M matching	23.2	23.9	30.8	12.2	56.8	53.5	27.7 (−2.5)
w/o geometry loss	28.3	27.7	28.6	11.1	57.0	56.4	28.5 (−1.7)
w/o deep supervision	28.1	28.3	32.4	12.6	58.7	56.6	29.9 (−0.3)
w/o ignore-aware suppression	28.2	29.4	33.2	13.0	59.2	56.4	30.0 (−0.2)

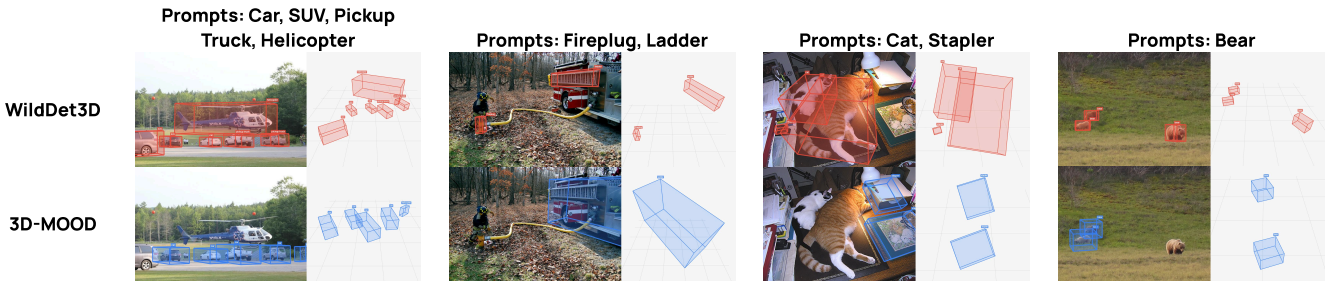


Figure 9. **Qualitative comparison with text prompts.** Each block shows the same scene detected by WildDet3D (top) and 3D-MOOD (bottom), prompted with text categories only. WildDet3D consistently detects more object categories with more realistic placement, orientation, and shape.