Detect Anything 3D in the Wild

Hanxue Zhang[1,2*], Haoran Jiang[1,3*], Qingsong Yao[4*], Yanan Sun[1], Renrui Zhang[5]
Hao Zhao[6], Hongyang Li[1], Hongzi Zhu[2], Zetong Yang[1,7]

[1] OpenDriveLab at Shanghai AI Laboratory [2] Shanghai Jiao Tong University [3] Fudan University
[4] Stanford University [5] CUHK MMLab [6] Tsinghua University [7] GAC R&D Center

https://github.com/OpenDriveLab/DetAny3D

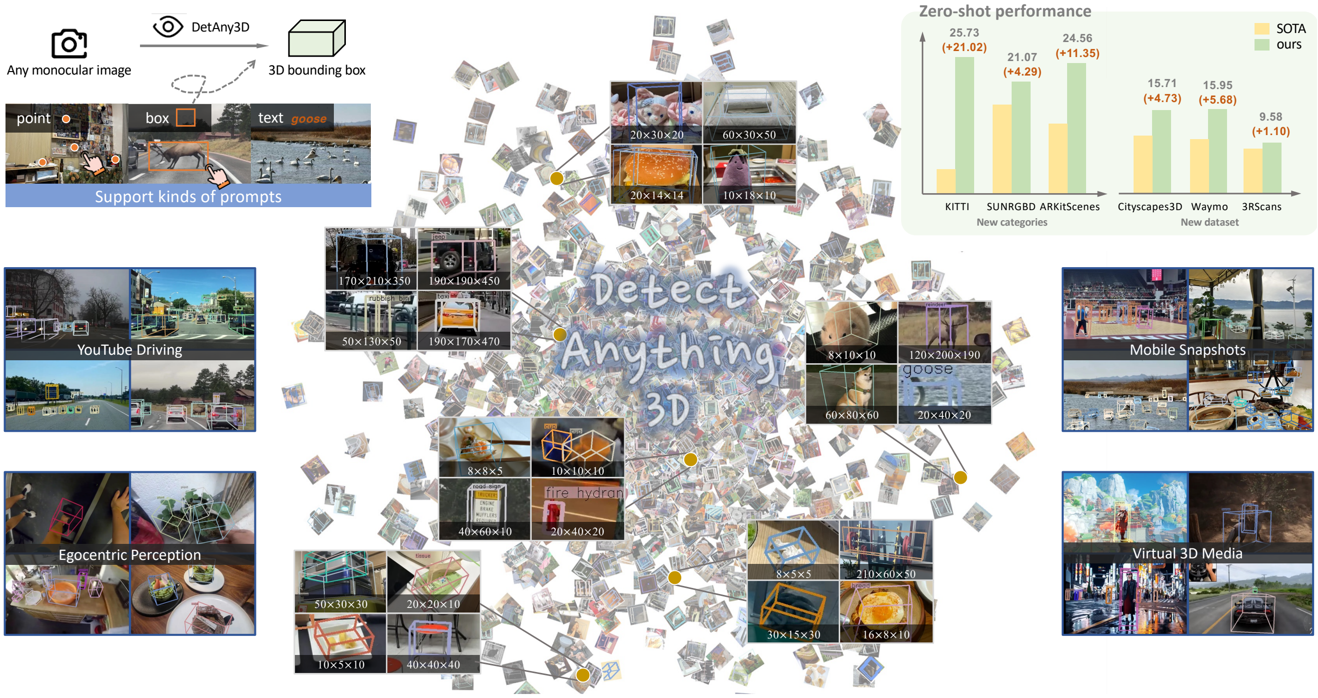

Figure 1. Introducing *DetAny3D*, a promptable 3D detection foundation model capable of detecting any 3D object with arbitrary monocular images in diverse scenes. Our framework enables multi-prompt interaction (*e.g.*, box, point, and text) to deliver open-world 3D detection results ($w \times h \times l$ in centimeter) for novel objects across various domains. It achieves significant zero-shot generalization, outperforming SOTA by up to 21.02 and 5.68 $\mathrm{AP}_{3\mathrm{D}}$ on novel categories and novel datasets with new camera configurations.

Abstract

Despite the success of deep learning in close-set 3D object detection, existing approaches struggle with zero-shot generalization to novel objects and camera configurations. We introduce DetAny3D, a promptable 3D detection foundation model capable of detecting any novel object under arbitrary camera configurations using only monocular inputs. Training a foundation model for 3D detection is fundamentally constrained by the limited availability of annotated 3D data, which motivates DetAny3D to leverage the rich prior knowledge embedded in extensively pre-trained 2D foundation models to compensate for this scarcity. To effectively transfer 2D knowledge to 3D, DetAny3D incorporates two core modules: the 2D Aggregator, which aligns features from different 2D foundation models, and the 3D Interpreter with Zero-Embedding Mapping, which stabilizes early training in 2D-to-3D knowledge transfer. Experimental results validate the strong generalization of our DetAny3D, which not only achieves state-of-the-art performance on unseen categories and novel camera configura-

*Equal contribution.

tions, but also surpasses most competitors on in-domain data. DetAny3D sheds light on the potential of the 3D foundation model for diverse applications in real-world scenarios, e.g.*, rare object detection in autonomous driving, and demonstrates promise for further exploration of 3D-centric tasks in open-world settings. More visualization results can be found at our code repository.*

1. Introduction

3D object detection is a fundamental technology for autonomous systems [12, 14, 15, 36, 48, 49], robotics [6, 67, 84], and augmented reality [43, 52]. 3D perception not only enables machines to perceive and interact with the physical world, but also serves as a foundational input for more advanced tasks, such as behavior decision [3, 11, 20, 31], world modeling [22, 23, 38] and 3D scene reconstruction [50, 73, 75]. For practical deployment, a generalizable 3D detector ideally should detect arbitrary objects from easily accessible inputs, such as monocular images, without relying on specific sensor parameters. Such a model would be highly adaptable and reliable for various downstream tasks in diverse and unpredictable environments [15, 36, 43, 84]. Also, accurate detection results provided by such a detector (*e.g.*, generating 3D bounding boxes for even images from the internet) make it a versatile tool, paving the way for scalable 3D systems that leverage Internet-scale data and advance toward open-world scenarios [22, 23, 38, 50, 73].

Previous research, exemplified by Omni3D [8], has attempted to improve the generalization of the 3D detection system through multi-dataset training [8, 35, 40, 68]. However, despite utilizing large datasets to train a unified detector [8, 40], these approaches provide limited generalization to novel camera configurations and cannot detect unseen object categories beyond predefined label spaces. Therefore, developing a 3D detection foundation model with strong zero-shot generalizability, which is capable of detecting any unseen object under arbitrary camera configurations, remains a crucial and unsolved problem.

While recent advances in 2D foundation models [33, 44, 51, 56] demonstrate remarkable zero-shot capabilities. Segment Anything Model (SAM) [33] features a promptable inference mechanism, supporting user-friendly prompts like points and boxes to segment user-specified objects. Their impressive generalization ability stems from training on billions of annotated images. However, in 3D object detection, the available labeled data is limited to only millions of samples—typically 3-4 orders of magnitude smaller than in 2D images. Such severe data scarcity [74, 86] poses a fundamental challenge, making it nearly infeasible to train a 3D foundation model from scratch.

In this work, we present DetAny3D, a promptable 3D detection foundation model designed for generalizable 3D object detection using only monocular images (see Figure 1). Given the inherent scarcity of 3D annotated data, we achieve strong generalization from two critical perspectives: model architecture and data utilization. The central insight of our approach is to leverage the extensive prior knowledge encoded within two broadly pre-trained 2D foundation models—SAM [33] and DINO [10, 51]—thus unlocking effective zero-shot 3D detection capabilities with minimal available 3D data.

Specifically, we adopt SAM as our promptable backbone, capitalizing on its versatile and robust object understanding capability derived from large-scale 2D data. Concurrently, we utilize DINO [51] depth-pretrained by UniDepth [54], to offer redundant 3D geometric priors [7, 76], which plays a pivotal role for accurate 3D detection in a monocular setting. To integrate the complementary features from SAM and DINO more effectively, we propose the 2D Aggregator, an attention-based mechanism that aligns these features and dynamically optimizes their contributions via learnable gating. 2D Aggregator fully exploits the strengths of each foundation model.

To further address the challenge of effectively transferring knowledge from 2D to 3D, we introduce the 3D Interpreter. Central to the 3D Interpreter is the Zero-Embedding Mapping (ZEM) mechanism, which ensures stable 2D-to-3D mapping by reducing early-stage interference and preserving pretrained 2D priors. By stabilizing the training process across diverse datasets with varying camera parameters, scene complexities, and depth distributions, the ZEM mechanism enables progressive zero-shot 3D grounding capabilities, significantly enhancing model generalization.

To leverage as much 3D-related data as possible, we aggregate a diverse range of datasets, including 16 datasets spanning depth with intrinsic data and 3D detection data, refered as DA3D. Experimental results, using prompts aligned with the baselines, demonstrate three key advantages: (1) Generalization to novel classes: achieves 21.0%, 4.3%, 11.3% higher zero-shot $\mathrm{AP}_{\mathrm{3D}}$ than baselines on novel categories on KITTI, SUNRGBD, and ARKitScenes. (2) Generalization to novel cameras: improves cross-dataset performance by 4.7%, 5.7% and 1.1% $\mathrm{AP}_{\mathrm{3D}}$ compared to baseline methods on zero-shot datasets Cityscapes3D, Waymo and 3RScan. (3) Performance on in-domain data: surpasses baseline by 1.6% $\mathrm{AP}_{\mathrm{3D}}$ on Omni3D. Core contributions are summarized in following:

- We develop DetAny3D, a promptable 3D detection foundation model capable of detecting any 3D object in real-world scenarios with arbitrary monocular inputs.
- DetAny3D introduces 2D Aggregator to effectively fuse the features from two 2D foundation models SAM and depth-pretrained DINO, which provide pivot shape and 3D geometric priors for various objects, respectively.
- In 2D-to-3D knowledge transfer, DetAny3D involves

Zero-Embedding Mapping in 3D Interpreter to achieve stable 2D-to-3D mapping, enabling the model to train robustly across datasets with diverse camera parameters, varying scenes, and different depth distributions.

- The experimental results demonstrate significant advantages of DetAny3D, particularly in accurately detecting unseen 3D objects with arbitrary camera parameters in the zero-shot setting, showcasing its potential across a wide range of real-world applications.

2. Related works

2.1. 3D Object Detection

Existing 3D object detection systems have predominantly focused on single-dataset optimization, achieving strong performance on benchmark datasets like KITTI [24] and nuScenes [9] through task-specific architectures [14, 18, 39, 41, 42, 45, 66, 80]. While effective in constrained scenarios, these approaches suffer from significant domain gaps when deployed in new contexts, primarily due to their reliance on limited sensor-specific data and closed-set assumptions. Recent works, exemplified by Omni3D [8], have demonstrated the potential of multi-dataset training. Models like Cube R-CNN [8] and UniMODE [40] train a universal monocular 3D detector across multiple datasets, achieving some level of robustness to camera parameters, but are still restricted to predefined classes. V-MIND [32] further addresses the data scarcity challenge by generating pseudo 3D training data from large-scale 2D annotations. Towards more general detection, OV-Uni3DETR [69] pioneers open-set detection that is able to detect with multimodal inputs, but it is trained separately for indoor and outdoor domains, thereby limiting its overall generalization. More recently, OVMono3D [74] leverages Grounding DINO's [44] 2D results with a 3D head on unified datasets. However, it does not fully exploit the priors contained in 2D foundation models, leading to performance constraints tied to the limited 3D data. In contrast, our approach fully capitalizes on the knowledge distilled in 2D foundation models while leveraging abundant 3D-related data, thereby enabling the detection of any 3D object from arbitrary monocular inputs.

2.2. Vision Foundation Models

Foundation models have demonstrated significant potential across various domains. For example, language foundation models such as GPT-4 [1] and DeepSeek [5, 26], trained on massive internet-scale corpora, have achieved impressive capabilities in natural language processing across diverse fields [1, 5, 60, 63, 81, 82]. Similarly, foundation models in the vision domain have made remarkable strides [29, 33, 37, 44, 51, 56, 79]. DINOv2 [51], trained on a vast range of curated data from diverse sources, is capable of producing general-purpose visual features that work seamlessly across different image distributions and tasks. SAM [33] has taken a step further in the vision domain by introducing promptability, enabling models to generalize to novel visual concepts through large-scale data training and continuous model refinement. In recent years, the development of foundation models in the 3D domain has started to take initial steps [13, 28, 55, 78, 83, 85]. Most existing 3D foundation models are often combined with vision-language models (VLMs) [13, 27, 55, 85], relying on point clouds as input to help the language models understand 3D [13, 85]. While these methods are valuable for scene understanding and semantic tasks, they do not directly provide precise 3D detection results. Moreover, point cloud inputs significantly restrict the use cases [72], as they are not always accessible in many practical scenarios. In contrast to these approaches, we aim to develop a foundation model specifically dedicated to 3D detection tasks with the most general inputs, monocular images. By leveraging the powerful priors from 2D vision foundation models, our approach enables the detection of any 3D object with arbitrary camera configurations, presenting a broad range of practical applications.

3. Detect Anything 3D in the Wild

3.1. Overview

As illustrated in Figure 2(a), DetAny3D takes a monocular RGB image and prompts (*e.g.*, boxes, points, text, intrinsic) as input. The box, point, and text prompts are used to specify objects, while the intrinsic prompts are optional. When not provided, the model predicts the intrinsic parameters and the corresponding 3D detection results. If intrinsic are available, the model can leverage them as geometric constraints to mitigate the ill-posed nature of monocular depth estimation and calibrate its detection results.

Specifically, the monocular image is embedded in parallel by two foundational models: SAM [33] for low-level pixel information, underpins the entire promptable architecture. And depth-pretrained DINO [51, 54], which provide rich high-level geometric knowledge, excels in depth-related tasks. These complementary 2D features are then fused through our proposed 2D Aggregator (see Figure 2(b)), which hierarchically aligns low-level and high-level information using cross-attention layers. The fused features are subsequently passed to the Depth/Camera Module, which extracts the camera and camera-aware depth embedding, collectively referred to as geometric embedding.

The geometric embedding and the 3D bounding box tokens with encoded prompt tokens are then fed into the 3D Interpreter (see Figure 2(c)), which employs a structure similar to the SAM decoder along with a specialized Zero-Embedding Mapping (ZEM) mechanism. 3D Interpreter injects 3D geometric features while ensuring stable 2D-to-

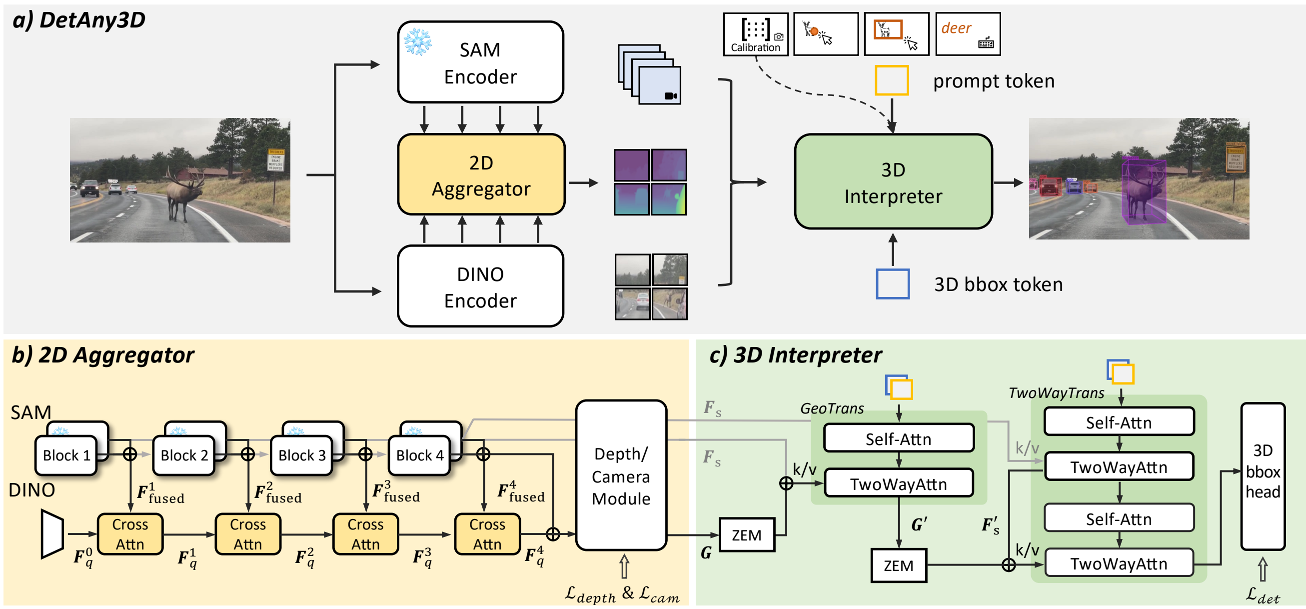

Figure 2. **Overview of DetAny3D.** It supports arbitrary monocular images as input and performs 3D object detection driven by prompts—box, point, and text to specify target objects and optional camera calibration to calibrate geometric projections. DetAny3D comprises two key modules: (b) 2D Aggregator, which employs a hierarchical cross-attention mechanism to dynamically fuse knowledge from SAM and DINO, with a learnable gate controlling each component's contribution to the geometric embedding; (c) 3D Interpreter, which introduces a Zero-Embedding Mapping (ZEM) strategy based on zero-initialized layers to gradually inject geometric priors, thereby enables zero-shot 3D grounding and avoids catastrophic forgetting during knowledge transfer.

3D knowledge transfer, enabling progressive 3D grounding across diverse data domains. Finally, the model predicts 3D boxes based on the hidden states of the 3D box tokens. Our DetAny3D is trained on selected seen classes and can detect any unseen classes in a zero-shot manner.

3.2. 2D Aggregator

To effectively fuse multiple foundation models , we propose 2D Aggregator to aggregate features from SAM and DINO, mitigating potential conflicts between their heterogeneous representations. As illustrated in Figure 2(b), the 2D Aggregator fuses features from SAM and DINO in a hierarchical manner, progressively integrating spatial and geometric information across four cascaded alignment units.

Feature Extraction. Given an input image, the SAM encoder extracts high-resolution spatial features $\mathbf{F}_s \in \mathbb{R}^{H_s \times W_s \times C}$, capturing fine-grained details and boundaries. Simultaneously, the DINO encoder outputs geometry-aware embeddings $\mathbf{F}_d \in \mathbb{R}^{H_d \times W_d \times C}$, which is depth-pretrained by UniDepth [54] and provides robust priors for depth and intrinsics. Following the design of ViT Adapter [16], we also employ a convolutional structure to produce preliminary image features, denoted as $\mathbf{F}_q^0$, serving as the initial query for subsequent attention-based fusion.

Hierarchical Fusion. Each of the four alignment units fuses SAM and DINO features via cross-attention. In the i-th unit, we first apply learnable gating weights α_i (initialized to 0.5) to combine the i-th block of SAM features $\mathbf{F}_s^i$ and DINO features $\mathbf{F}_d^i$ as follows:

$$\mathbf{F}_{\text{fused}}^i = \alpha_i \cdot \mathbf{F}_s^i + (1 - \alpha_i) \cdot \mathbf{F}_d^i. \tag{1}$$

We use $\mathbf{F}_{\text{fused}}^i$ as key and value, while the query feature $\mathbf{F}_q^{i-1}$ acts as the query in the cross-attention mechanism:

$$\mathbf{F}_q^i = \texttt{CrossAttn}\big(\mathbf{F}_q^{i-1}, \mathbf{F}_{\text{fused}}^i, \mathbf{F}_{\text{fused}}^i\big), \tag{2}$$

$$\hat{\mathbf{F}}_{\text{fused}}^i = \texttt{Norm}\big(\mathbf{F}_{\text{fused}}^i + \mathbf{F}_q^i\big). \tag{3}$$

This design enables the model to dynamically emphasize SAM's spatial details or DINO's semantic and geometric cues at different hierarchy levels while minimizing interference between the two representations.

Geometric Embeddings. The fused features $\hat{\mathbf{F}}_{\text{fused}}^i$, $i \in [1, 2, 3, 4]$, are subsequently processed by the depth and camera modules, following the UniDepth [54] architecture. Specifically, these modules predict the camera embedding $\mathbf{C}$ and camera-aware depth embedding $\mathbf{D}|\mathbf{C}$, referred as the geometric embedding $\mathbf{G} = \{\mathbf{D}|\mathbf{C}, \mathbf{C}\}$. These modules provide aligned depth and camera parameters under the monocular depth ill-posed problem. Further details can be found in the Supplementary material Section 2.1.

Overall, by progressively aligning multi-scale features and adaptively integrating their contributions, 2D Aggregator effectively leverages the strengths of both foundation models while minimizing potential conflicts.

3.3. 3D Interpreter

The diverse 3D object supervisions across various scenarios, depths, and camera intrinsics introduce challenges to model training. Our 3D Interpreter aims to progressively integrate geometric information while ensuring stable 2D-to-3D knowledge transfer. We introduce Zero-Embedding Mapping (ZEM) mechanism, which incrementally infuses 3D geometry into the decoder via zero-initialized layers—without disrupting the original 2D features. As Figure 2(c) shows, the 3D Interpreter comprises three main components: the Two-Way Transformer, the Geometric Transformer, and the 3D bounding box heads.

Two-Way Transformer. Following the SAM design, we first concatenate the 3D bounding box tokens with prompt-related tokens to form the query:

$$\mathbf{Q} = \Big[[\mathbf{T}_{\mathrm{3D},1}; \mathbf{T}_{\mathrm{p},1}], \cdots, [\mathbf{T}_{\mathrm{3D},N}; \mathbf{T}_{\mathrm{p},N}] \Big], \tag{4}$$

where $\mathbf{T}_{\mathrm{3D},i}$ denotes the 3D bounding box token for the i-th object, $\mathbf{T}_{\mathrm{p},i}$ is the prompt-related token, and $[\cdot;\cdot]$ denotes vector concatenation. The SAM encoder output $\mathbf{F}_s$ serves as both key and value for the first Two-Way Transformer layer, yielding:

$$\mathbf{F}'_s = \mathtt{TwoWayTrans}(\mathbf{Q}, \mathbf{F}_s, \mathbf{F}_s). \tag{5}$$

The initialized parameters of two-way transformer are copied using pre-trained SAM decoder.

Geometric Transformer. We then process the geometric embedding $\mathbf{G}$ (from the 2D Aggregator) through the zero-initialized 1×1 convolutional layer ZEM and add it to $\mathbf{F}_s$ for use as key and value in the Geometric Transformer:

$$\mathbf{G}' = \mathtt{GeoTrans}\big(\mathbf{Q}, \mathtt{ZEM}(\mathbf{G}) + \mathbf{F}_s, \mathtt{ZEM}(\mathbf{G}) + \mathbf{F}_s\big). \tag{6}$$

ZEM integrates the geometric embedding and avoids catastrophic forgetting in 2D features. Next, $\mathbf{G}'$ is again passed through ZEM and combined with $\mathbf{F}'_s$. This enriched representation is used as key and value in the second Two-Way Transformer layer to generate object features $\mathbf{O}$:

$$\mathbf{O} = \mathtt{TwoWayTrans}\big(\mathbf{Q}', \mathtt{ZEM}(\mathbf{G}') + \mathbf{F}'_s, \mathtt{ZEM}(\mathbf{G}') + \mathbf{F}'_s\big). \tag{7}$$

ZEM also helps stabilize parameter updates in the two-way and geometric transformer training, preventing conflicts arising from diverse 3D object supervision.

3D Bounding Box Heads. Finally, $\mathbf{O}$ is fed into the 3D bounding box heads to calculate the final predictions, which follows typical architectures from standard 3D detection frameworks [8, 66, 80]: $B_{\mathrm{3D}}(x, y, z, w, h, l, R, S)$ where x, y, z specify the 3D box center, w, h, l are its dimensions, R is the rotation matrix, and S is the predicted 3D Intersection over Union (IoU) score.

3.4. Loss

Our loss function comprises three components, the depth loss $\mathcal{L}_{\mathrm{depth}}$, the camera intrinsic loss $\mathcal{L}_{\mathrm{cam}}$, and the detection loss $\mathcal{L}_{\mathrm{det}}$. The overall loss is defined as the sum of these three components. For depth loss $\mathcal{L}_{\mathrm{depth}}$, we adopt the commonly used SILog loss [19, 64] to supervise depth prediction. For camera intrinsic loss $\mathcal{L}_{\mathrm{cam}}$, we follow the dense camera ray approach [30, 54] to represent intrinsics and also employ the SILog loss to measure deviations between predicted and ground-truth parameters. At last, for detection loss $\mathcal{L}_{\mathrm{det}}$, we use the smooth L1 loss [40, 66, 80] to regress 3D bounding boxes parameters and predicted IOU scores and the Chamfer loss [8, 74] for rotation matrices. Detailed formulations of these loss functions can be found in the supplementary material Section 2.3.

3.5. Prompt Interaction

DetAny3D supports point, box, and text prompts to detect 3D box for user-specified objects. To calibrate more precise depth for specific camera, DetAny3D allows users to specify the camera configuration via the intrinsic prompt.

Box and Point Prompts. Following SAM's methodology, both box and point prompts are encoded based on their respective positions and embeddings. For the box prompt, two points (top-left and bottom-right corners) are used. The point prompt is derived by combining the positional encoding of the point and the corresponding embedding.

Text Prompts. Recent 2D foundation models like Grounding DINO [44] are able to detect bounding box for the open-vocabulary object specified by users using text prompt. DetAny3D can further generate 3D bouding box using the prediction of Grounding DINO, which enables text as prompts in the zero-shot interface.

Intrinsic Prompts. Unlike most existing 3D detectors that employ a fixed virtual camera and rely on GT intrinsics to recover the true depth, inspired by Unidepth, we predict intrinsics for camera-aware 3D detection. When no intrinsic prompt is given, the model infers intrinsics for outputs:

$$\mathrm{Box}_{3D} = \mathtt{3DInterpretor}\big(\mathbf{Q}, \hat{\mathbf{G}}, \mathbf{F}_s\big), \tag{8}$$

where $\hat{\mathbf{G}} = \{\mathbf{D}|\hat{\mathbf{C}}, \hat{\mathbf{C}}\}$, $\hat{\mathbf{C}}$ is the predicted camera embedding, and $\mathbf{D}|\hat{\mathbf{C}}$ is the depth embedding conditioned on the predicted camera embedding. When intrinsic prompts are given, the model refines the 3D detection results based on the true intrinsic:

$$\mathrm{Box}_{3D} = \mathtt{3DInterpretor}\big(\mathbf{Q}, \mathbf{G}, \mathbf{F}_s\big), \tag{9}$$

where $\mathbf{G} = \{\mathbf{D}|\mathbf{C}, \mathbf{C}\}$. This boosts performance on both intrinsic prediction and 3D detection since the model continuously predicts and aligns the intrinsic with the 3D detection rather than estimating it solely from input image.

4. Experiment

4.1. Experimental Setup

DA3D Benchmark. We present DA3D, a unified 3D detection dataset that aggregates 16 diverse datasets for 3D detection and depth estimation. Building upon Omni3D's original datasets (Hypersim [57], ARKitScenes [4], Objectron [2], SUNRGBD [61], KITTI [24], and nuScenes [9]), we incorporate additional four outdoor detection datasets (Argoverse2 [70], A2D2 [25], Waymo [62], Cityscapes3D [21]), one indoor detection dataset (3RScan [65]), and five depth and intrinsic datasets (Scannet [17], Taskonomy [77], DrivingStereo [71], Middlebury [59], IBIMS-1 [34]). All data is standardized with monocular images, camera intrinsics, 3D bounding boxes, and depth maps. Following prior work [74], we select partial categories from KITTI, SUNRGBD, and ARKitScenes as zero-shot test classes. We select Cityscapes3D, Waymo, and 3RScan as our zero-shot datasets with novel camera configurations, where 3RScan also contains novel object categories. Depth supervision from LiDAR, RGB-D, and stereo sensors enhances 75% of training samples, while intrinsic parameters cover 20 camera configurations across 0.4 million frames (2.5× Omni3D's scale). Dataset statistics and splits are detailed in Supplementary material Section 1. All data are subject to their respective licenses.

Baselines. We choose Cube R-CNN [8] and OVMono3D [74] as our primary baselines, as their settings align most closely with our experimental protocol: Cube R-CNN is a benchmark provided by the Omni3D dataset. It is a unified detector capable of performing detection on predefined categories. OVMono3D is a recently available open-vocabulary 3D detector on the Omni3D dataset. It lifts 2D detection to 3D by connecting the open-vocabulary 2D detector Grounding DINO [44] with a detection head.

Metrics. We adopt the metrics in the Omni3D benchmark [8], which is Average Precision (AP). Predictions are matched to ground-truth by measuring their overlap using IoU3D, which computes the intersection-over-union (IoU) of 3D cuboids. The IoU3D thresholds range from $\tau \in [0.05, 0.10, \ldots, 0.50]$. For experiments using text prompts, we additionally employ target-aware metrics from OVMono3D [74]: Prompt the detector only with category names present in the per-image annotations instead of providing an exhaustive category list. This addresses severe naming ambiguity (*e.g.*, "trash can" *vs.* "rubbish bin") and missing annotation issues prevalent in indoor datasets like 3RScan (see Supplementary material Section 3.).

Implementation Details. We implement DetAny3D via PyTorch [53]. We use the pretrained ViT-L DINOv2 [51, 54] and ViT-H SAM [33] as our initial models, with SAM serving as the promptable backbone, where the encoder is frozen during training. All main experiments are conducted using 8 NVIDIA A100 machines with 8 GPUs for each and a batch size of 64. The model is trained for 80 epochs, taking approximately 2 weeks to complete. The training uses the AdamW [47] optimizer with an initial learning rate of 0.0001, adjusted according to the cosine annealing policy [46]. During box prompt training, we apply a 0.1 positional offset disturbance. For point prompt training, points are randomly selected from the mask. Text prompts are converted into box prompts via Grounding DINO SwinT [44]. For fair comparisons, all baseline-related experiments incorporate intrinsic prompts and use aligned prompt inputs.

4.2. Main Results

Zero-shot Category Performance. In this experiment, we use two sources for the prompt input: text prompt processed by Grounding DINO and box prompt from ground-truth 2D bounding box. We evaluate our model on KITTI, SUNRGBD, and ARKitScenes datasets with the same zero-shot categories as OVMono3D [74]. As shown in Table 1 (left), our DetAny3D demonstrates superior zero-shot adaptation performance compared to the OVMono3D baseline. When using Grounding DINO for text prompt input, our method achieves significant improvements of 21.02 AP_{3D} on KITTI, 4.29 AP_{3D} on SUNRGBD, and 11.35 AP_{3D} on ARKitScenes under the target-aware metric. When using 2D ground-truth as box prompt input, DetAny3D attains 28.96 AP_{3D} on KITTI, 39.09 AP_{3D} on SUNRGBD, and 57.72 AP_{3D} on ARKitScenes, showing 3.4×, 2.3×, and 4.1× gains over the baseline, respectively. This substantial performance gap highlights our method's enhanced ability to generalize to novel object categories.

Zero-shot Camera Performance. To assess robustness against novel camera parameters, we conduct cross-dataset evaluation as shown in Table 1 (right). For Cityscapes3D and Waymo, We use Cube R-CNN's 2D detections and ground-truth as box prompt and Grounding DINO processed text prompt for comparison. For 3RScan, due to namespace inconsistency with Cube R-CNN's predefined categories and the presence of novel classes, we only use text prompt and ground-truth box prompts, benchmarking against OVMono3D. DetAny3D exhibits strong adaptation to unseen camera configurations. When using Cube R-CNN-aligned prompts, our model achieves AP_{3D} scores of 10.33 and 15.17 on Cityscapes3D and Waymo, respectively, surpassing Cube R-CNN by +2.11 and +5.74. With text prompts, under identical settings as OVMono3D [74], our method improves AP_{3D} by +4.73 on Cityscapes3D, +5.68 on Waymo, and +1.1 on 3RScan under *target-aware metrics*. Both models show low scores on conventional metrics for 3RScan due to severe naming ambiguity and missing annotations. Using 2D ground-truth as box prompts, DetAny3D attains AP_{3D} of 16.88, 15.83, and 21.36 across the three datasets, outperforming OVMono3D by +6.82, +5.6,

Table 1. Zero-shot 3D detection performance comparison on novel categories (left) and novel cameras (right). Results report $\mathrm{AP}_{\mathrm{3D}}$ with different prompt strategies: (1) Cube R-CNN, (2) *Grounding DINO* outputs (traditional metric / target-aware metric) and (3) *Ground Truth*. Target-aware metric uses per-image existing categories for prompting.

Prompt	Method	Novel Categories			Novel Cameras		
		$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{kit}}$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{sun}}$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{ark}}$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{city}}$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{wym}}$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{3rs}}$
-	Cube R-CNN [8]	-	-	-	8.22	9.43	-
Cube R-CNN	OVMono3D [74]	-	-	-	4.97	10.89	-
	DetAny3D (ours)	-	-	-	**10.33**	**15.17**	-
	Δ	-	-	-	+5.36	+4.28	-
Grounding DINO	OVMono3D [74]	4.71 / 4.71	4.07 / 16.78	13.21 / 13.21	5.88 / 10.98	9.20 / 10.27	0.37 / 8.48
	DetAny3D (ours)	**25.73 / 25.73**	**7.63 / 21.07**	**24.56 / 24.56**	**11.05 / 15.71**	**15.38 / 15.95**	**0.65 / 9.58**
	Δ	+21.02 / +21.02	+3.56 / +4.29	+11.35 / +11.35	+5.17 / +4.73	+6.18 / +5.68	+0.28 / +1.10
Ground Truth	OVMono3D [74]	8.44	17.16	14.12	10.06	10.23	18.05
	DetAny3D (ours)	**28.96**	**39.09**	**57.72**	**16.88**	**15.83**	**21.36**
	Δ	+20.52	+21.93	+43.60	+6.82	+5.60	+3.31

Table 2. In-domain performance comparison between DetAny3D and baselines. The first three columns show results trained only on NuScenes and KITTI, while the next seven columns show results trained on the unified dataset. Two prompt sources are used: (1) *Cube R-CNN* 2D detections, (2) *Ground Truth*.

Method	Omni3D_OUT			Omni3D						
	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{kit}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{nus}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{out}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{kit}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{nus}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{sun}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{ark}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{obj}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}^{\mathrm{hyp}}\uparrow$	$\mathrm{AP}_{\mathrm{3D}}\uparrow$
ImVoxelNet [58]	23.5	23.4	21.5	-	-	-	-	-	-	9.4
SMOKE [45]	25.9	20.4	20.0	-	-	-	-	-	-	10.4
OV-Uni3DETR [68]	35.1	33.0	31.6	-	-	-	-	-	-	-
Cube R-CNN [8]	**36.0**	32.7	31.9	**32.50**	30.06	15.33	41.73	50.84	7.48	23.26
OVMono3D [74]$_{w/\ \text{Cube RCNN}}$	-	-	-	25.45	24.33	15.20	41.60	**58.87**	**7.75**	22.98
DetAny3D (ours)$_{w/\ \text{Cube RCNN}}$	35.8	**33.9**	**32.2**	31.61	**30.97**	**18.96**	**46.13**	54.42	7.17	**24.92**
OVMono3D [74]$_{w/\ \text{Ground Truth}}$	-	-	-	33.69	23.79	27.83	40.85	56.64	11.99	25.32
DetAny3D (ours)$_{w/\ \text{Ground Truth}}$	**38.0**	**36.7**	**35.9**	**38.68**	**37.55**	**46.14**	**50.62**	**56.82**	**15.98**	**34.38**

and +3.31, respectively. These results highlight the effectiveness of our architecture and its potential for real-world applications with arbitrary camera configurations.

In-domain Performance We also evaluate our model's in-domain detection capability using two prompt sources: 2D detections from Cube R-CNN and 2D ground-truth. Besides the unified model, we also train DetAny3D on Omni3D_out for comparison. As shown in Table 2, DetAny3D achieves competitive results with Cube R-CNN when provided with aligned input. Using GT prompts, DetAny3D outperforms OVMono3D by 9.06 $\mathrm{AP}_{\mathrm{3D}}$, indicating that Cube R-CNN may bottleneck performance, and stronger 2D prompts could further boost results.

4.3. Possible Applications of DetAny3D

Other than robustly detecting diverse corner cases in real-world tasks such as autonomous driving and embodied perception, DetAny3D's open-world detection results can further serve as inputs for advanced downstream tasks.

3D Bounding Box Guided Video Generation. We feed DetAny3D outputs into Sora for zero-shot, open-world 3D box guided video generation. As shown in Figure 3, we compare: (i) image + 3D box + text, (ii) image + 2D box +

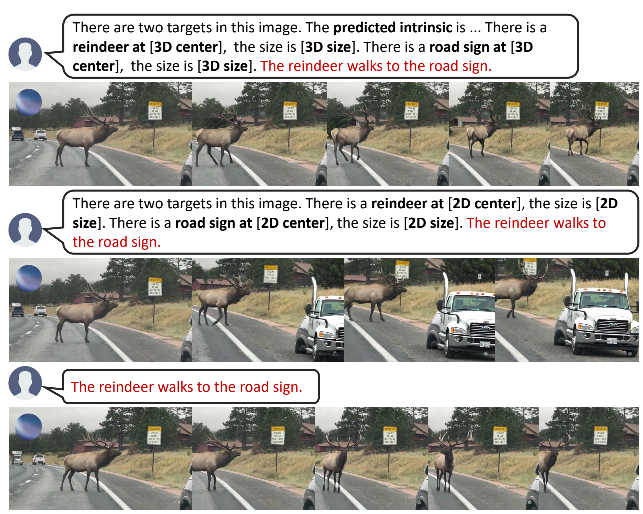

Figure 3. **Zero-Shot Transfer Video Generation via Sora.** We provide Sora with Internet-sourced images. As shown, when controlled with 3D bounding box, Sora can better capture the scene's geometric relationships. In contrast, with only controlled by 2D bounding box prompt, Sora respects pixel-level spatial cues but fails to generate accurate geometric offset.

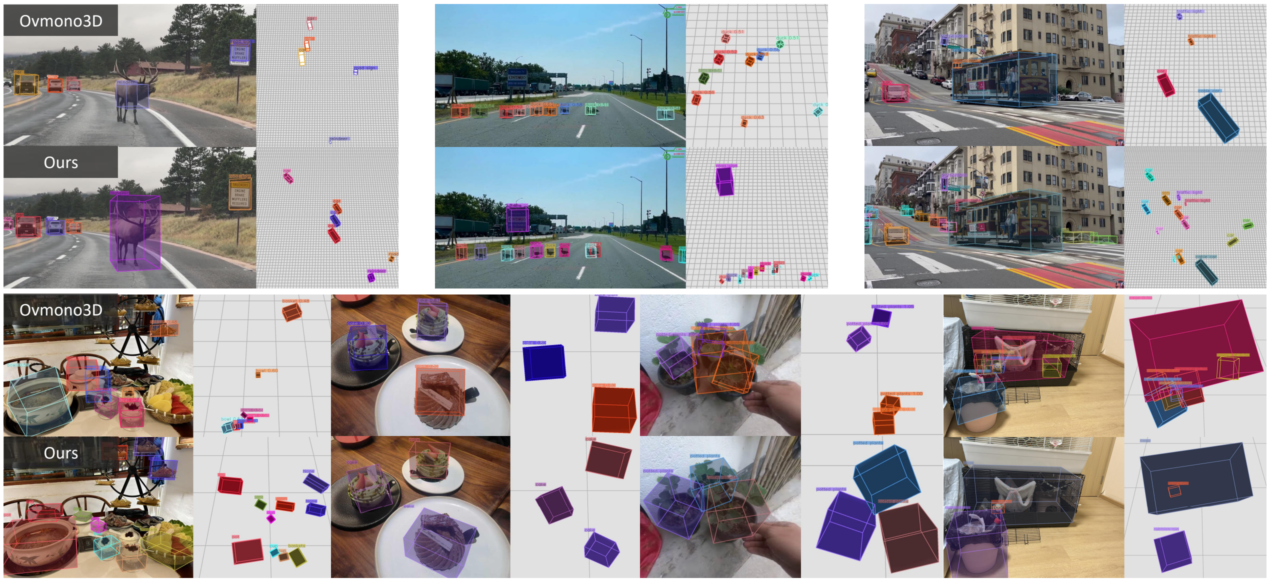

Figure 4. **Qualitative Results.** We present qualitative examples from open-world detection. In each pair of images, the top row is produced by OVMono3D, and the bottom row by DetAny3D. For each example, the left sub-figure overlays the projected 3D bounding boxes, while the right sub-figure shows the corresponding bird's-eye view with 1m×1m grids as the background.

Table 3. **Ablation study** of DetAny3D. The table shows the impact of different design choices on AP_{3D} performance. Each component is progressively added. To save resources, ablations are conducted on 10% of the full training dataset.

Depth&Cam.	Merge DINO	2D Agg.	ZEM	$AP_{3D}\uparrow$
-	-	-	-	5.81
√	-	-	-	10.10
√	√	-	-	20.20
√	√	√	-	23.21
√	√	√	√	25.80

text, and (iii) image + text. With 3D box constraints, Sora generates videos better aligned with intent.

4.4. Ablation Studies

As shown in Table 3, we ablate key components of DetAny3D, showing the evolution from a SAM-based baseline to DetAny3D with strong 3D generalization. The base model extends SAM with 3D box tokens and a 3D head for direct box prediction. Additional ablations, including backbone and prompt types, are in Supplementary Section 4.

- **Effectiveness of Depth & Camera Modules.** Depth map provides denser supervision, while camera configuration intrinsic help mitigate disruptions caused by multiple datasets training. Integrating both depth map and camera intrinsic yields improvement in 3D feature extraction and generalization across diverse datasets.
- **Effectiveness of Merging Depth-Pretrained DINO.** Incorporating depth-pretrained DINO yields remarkable improvements, demonstrating that the rich geometric information from DINO effectively compensates for SAM's limited geometric understanding.
- **Effectiveness of 2D Aggregator.** Compared to directly adding the features from two models, the 2D Aggregator reduces conflicts between different foundation models, further unleashing the performance gains from two foundation model integration.
- **Effectiveness of ZEM.** ZEM mechanism integrates geometric features through zero-initialized layers, which enables stable 2D-to-3D knowledge transfer during training across datasets with diverse camera parameters, scenes, and depth distributions.

4.5. Qualitative Results

We provide qualitative comparisons with OVMono3D. As shown in Figure 4, our model predicts more accurate intrinsics when the camera parameters are unknown and infers more consistent camera poses and 3D detections.

5. Conclusions

We propose DetAny3D, a promptable 3D detection foundation model that can detect arbitrary 3D objects from any monocular image input. DetAny3D exhibits significant zero-shot detection capabilities across diverse domains and effective zero-shot transfer across various tasks, highlighting its suitability for real-world deployment in dynamic and unstructured environments. Moreover, its flexible and robust detection ability opens the door to gathering large-scale, multi-source data for more 3D perception-guided tasks, paving the way toward open-world systems.

Acknowledgements

We sincerely thank Jiazhi Yang, Tianyu Li, Haochen Tian, Jisong Cai, and Li Chen for their invaluable discussions and constructive feedback throughout this project. Their insights and expertise have contributed significantly to the success of this work. We also appreciate the continuous support and encouragement from all the members of OpenDriveLab. This work is supported by the National Key Research and Development Program of China (2024YFE0210700) and the National Natural Science Foundation of China (NSFC) under Grants 62206172 and 62432008. The project is in part financially supported by Meituan Inc.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[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. 6

[3] Umar Asif, Jianbin Tang, and Stefan Harrer. Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices. In *IJCAI*, 2018. 2

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS Datasets*, 2021. 6

[5] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 3

[6] Georg Biegelbauer and Markus Vincze. Efficient 3d object detection by fitting superquadrics to range image data for robot's object manipulation. In *ICRA*, 2007. 2

[7] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2

[8] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *CVPR*, 2023. 2, 3, 5, 6, 7

[9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 3, 6

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[11] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *CVPR*, 2021. 2

[12] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE TPAMI*, 2024. 2

[13] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *CVPR*, 2024. 3

[14] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 2, 3

[15] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017. 2

[16] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023. 4

[17] 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. 6

[18] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *CVPR*, 2016. 3

[19] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 5

[20] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *CVPR*, 2020. 2

[21] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. *arXiv preprint arXiv:2006.07864*, 2020. 6

[22] Ruiyuan Gao, Kai Chen, Enze Xie, HONG Lanqing, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *ICLR*, 2023. 2

[23] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 2

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 3, 6

[25] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 6

[26] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024. 3

[27] Ziyu Guo*, Renrui Zhang*#, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3

[28] Ziyu Guo*, Renrui Zhang*#, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*, 2024. 3

[29] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 3

[30] Xiankang He, Guangkai Xu, Bo Zhang, Hao Chen, Ying Cui, and Dongyan Guo. Diffcalib: Reformulating monocular camera calibration as diffusion-based dense incident map generation. *arXiv preprint arXiv: 2405.15619*, 2024. 5

[31] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 2

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

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 3, 6

[34] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *ECCVW*, 2018. 6

[35] Maksim Kolodiazhnyi, Anna Vorontsova, Matvey Skripkin, Danila Rukhovich, and Anton Konushin. Unidet3d: Multi-dataset indoor 3d object detection. *arXiv preprint arXiv:2409.04234*, 2024. 2

[36] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *CVPR*, 2019. 2

[37] 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. 3

[38] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, 2024. 2

[39] 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. *IEEE TPAMI*, 2024. 3

[40] Zhuoling Li, Xiaogang Xu, SerNam Lim, and Hengshuang Zhao. Unimode: Unified monocular 3d object detection. In *CVPR*, 2024. 2, 3, 5

[41] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *NeurIPS*, 2022. 3

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

[43] Luyang Liu, Hongyu Li, and Marco Gruteser. Edge assisted real-time object detection for mobile augmented reality. In *MobiCom*, 2019. 2

[44] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2, 3, 5, 6

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

[46] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[48] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3d object detection from images for autonomous driving: a survey. *IEEE TPAMI*, 2023. 2

[49] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *IJCV*, 2023. 2

[50] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *CVPR*, 2020. 2

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 2, 3, 6

[52] Youngmin Park, Vincent Lepetit, and Woontack Woo. Multiple 3d object tracking for augmented reality. In *ISMAR*, 2008. 2

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[54] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 2, 3, 4, 5, 6

[55] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025. 3

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[57] 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. 6

[58] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*, 2022. 7

[59] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 6

[60] 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. In *ECCV*, 2024. 3

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

[62] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 6

[63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[64] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 5

[65] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *ICCV*, 2019. 6

[66] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, 2021. 3, 5

[67] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 2

[68] Zhenyu Wang, Ya-Li Li, Xi Chen, Hengshuang Zhao, and Shengjin Wang. Uni3detr: Unified 3d detection transformer. In *NeurIPS*, 2023. 2, 7

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

[70] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets*, 2023. 6

[71] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 2019. 6

[72] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *CVPR*, 2024. 3

[73] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. In *NeurIPS*, 2025. 2

[74] Jin Yao, Hao Gu, Xuweiyi Chen, Jiayun Wang, and Zezhou Cheng. Open vocabulary monocular 3d object detection. *arXiv preprint arXiv:2411.16833*, 2024. 2, 3, 5, 6, 7

[75] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *arXiv preprint arXiv:2502.12894*, 2025. 2

[76] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. 2

[77] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 6

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

[79] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *ICLR*, 2023. 3

[80] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, 2023. 3, 5

[81] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR*, 2024. 3

[82] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024. 3

[83] Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Hengshuang Zhao, Chunhua Shen, Yu

Qiao, Tong He, et al. Ponderv2: Pave the way for 3d foundation model with a universal pre-training paradigm. *arXiv preprint arXiv:2310.08586*, 2023. 3

[84] Menglong Zhu, Konstantinos G Derpanis, Yinfei Yang, Samarth Brahmbhatt, Mabel Zhang, Cody Phillips, Matthieu Lecce, and Kostas Daniilidis. Single image 3d object detection and pose estimation for grasping. In *ICRA*, 2014. 2

[85] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *ECCV*, 2024. 3

[86] Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? *arXiv preprint arXiv:2410.10799*, 2024. 2