Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

StereoVLA: Enhancing Vision-Language-Action Models with Stereo Vision

Shengliang Deng[1,3*] Mi Yan[1,2*] Yixin Zheng[1,4,5*] Jiayi Su[1,6] Wenhao Zhang[1,2]
Xiaoguang Zhao[4] Heming Cui[3] Zhizheng Zhang[1,5†] He Wang[1,2,5†]
[1]Galbot [2]CFCS, School of CS, Peking University [3]The University of Hong Kong [4]Institute of Automation, Chinese Academy of Sciences
[5]Beijing Academy of Artificial Intelligence [6]Xiamen University Malaysia

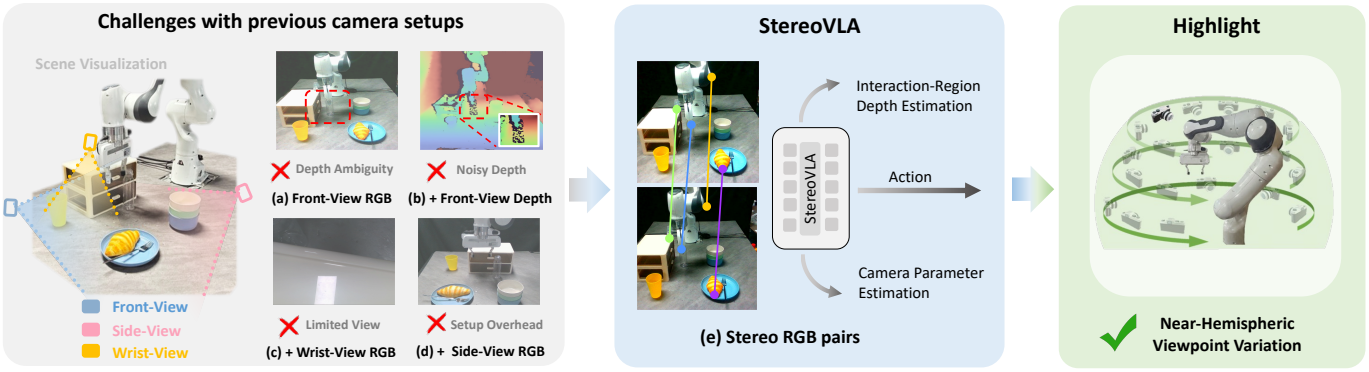


Fig. 1: **Left:** To enhance spatial perception, prior vision-language-action models typically supplement single-view RGB with sensor depth, wrist-view RGB, or side-view RGB. These setups face inherent challenges: (a) single-view RGB suffers from depth ambiguity; (b) depth sensors yield noisy estimates for transparent objects; (c) wrist-view RGB captures a limited view and the additional hardware increases collision risk; (d) multi-camera settings incur additional deployment overhead. **Middle:** (e) Stereo RGB pairs provide robust spatial cues for comprehensive scene understanding, mitigate view limit and collision risk, while enabling a simpler setup with a single stereo camera. Leveraging these distinctive strengths, we present StereoVLA, a novel VLA framework that takes in rich spatial and geometric cues from stereo RGB pairs, coupled with a dual-objective co-training strategy. **Right:** StereoVLA demonstrates robustness to near-hemispheric camera perspectives.

***Abstract*—While Vision-Language-Action (VLA) models excel in generalist manipulation, they often lack fine-grained spatial awareness and show limited viewpoint robustness. This limitation largely stems from the reliance on pretrained RGB encoders, which lack explicit geometric cues and prioritize semantic alignment over geometric representation. We argue that effective visual representations for VLA models must jointly encode both semantic and geometric information. In this paper, we introduce StereoVLA, the first VLA model to incorporate rich geometric cues from large-scale synthetic stereo data. StereoVLA employs a Geometric-and-Semantic (GeoSem) vision encoder that extracts geometric cues from subtle stereo-view disparities for precise spatial perception, while simultaneously capturing semantic features from pixel observations to support language-conditioned manipulation. Additionally, we introduce two synergistic co-training objectives: Interaction-Region Depth Estimation for precise spatial reasoning, and Camera Parameter Estimation to implicitly align perception and action coordinate systems. Compared with baselines that employ various input modalities, StereoVLA achieves a 33.4% absolute gain in success rate in real-world experiments and demonstrates robust generalization to near-hemispheric camera perspectives. Project page: shengliangd.github.io/StereoVLA-Webpage/.**

* Equal contribution. † denotes corresponding authors.
Correspondence to zhangzz@galbot.com, hewang@pku.edu.cn.

I. Introduction

Leveraging powerful pretrained Vision-Language Models (VLMs) [3, 2], VLA models can effectively acquire diverse skills from demonstrations while exhibiting robust semantic reasoning. Nonetheless, pioneering VLA models (e.g., OpenVLA [28]) are confined to single-view RGB inputs, limiting their fine-grained spatial awareness and viewpoint robustness. Consequently, providing VLA models with rich geometric cues remains a key challenge.

Previous works explored three types of additional sensors: wrist-mounted cameras [6, 4, 41], depth sensors [21, 83], and extra third-person cameras [13]. However, as shown in Figure 1, wrist cameras provide limited views that are easily occluded, and the wrist-mounted hardware increases collision risk with the environment. The reliability of depth sensors is often compromised by surface properties (e.g., transparent, specular, or light-absorbing materials) and external lighting conditions (e.g., sunlight saturation). Extra third-person cameras offer complementary information, but they complicate hardware for mobile robots and humanoids, and the increased view diversity hinders viewpoint robustness [67]. These settings do not explicitly provide stereoscopic perception for embodied manipulation, nor learning methods that are well

aligned with such perceptual cues.

Beyond the limitations imposed by camera configurations, we further observe that most existing VLA models rely on pretrained RGB vision encoders, which lack effective representations of geometric information. However, effective vision representations should jointly encode semantic and geometric information to ensure that robots can not only understand instructions, but also interact precisely with objects of diverse shapes, positions, and other geometric properties. Given these limitations, we explore enhancing VLA model with stereo vision to enable more precise robotic manipulation. While prior work incorporates only limited stereo data in a straightforward manner [27], and recent vision research has developed strong stereo-based depth foundation models [61], how to learn and integrate effective stereo representations into VLA models remains an important yet underexplored research problem.

In this paper, we introduce StereoVLA, the first VLA model to incorporate rich geometric cues from large-scale synthetic stereo data. We experimentally find that directly applying existing RGB encoders to stereo inputs based on current state-of-the-art VLA models, such as [6], is not sufficiently effective for learning embodied manipulation. This stems from two primary reasons. First, existing VLA models rely on pretrained RGB encoders, which lack explicit geometric cues and prioritize semantic alignment over geometric representation. Second, collecting large-scale robot data with stereo inputs is labor-intensive.

To address afore-discussed issues, we design a Geometric-and-Semantic (GeoSem) vision encoder to learn informative visual representations from a large-scale simulated data. The GeoSem Vision Encoder extracts geometric cues from subtle stereo-view disparities for precise spatial perception, while simultaneously capturing semantic features from pixel observations to support language-conditioned manipulation. Specifically, the GeoSem Vision Encoder employs the learned geometry-centric features from FoundationStereo [61], a state-of-the-art depth estimation foundation model pretrained on millions of stereo pairs. It further integrate and spatially align semantically rich tokens from PrismaticVLM [26] with these geometry-centric features to generate hybrid tokens that effectively combine geometry precision with semantics richness.

Besides, to tackle the challenge of limited stereo data, we draw inspiration from recent works [13, 10, 56, 75] to generate 5 million synthetic stereo trajectories. These works demonstrate that the synergy between modern high-fidelity simulators and powerful vision-language foundation models can effectively bridge the sim-to-real gap, enabling seamless zero-shot sim-to-real transfer. To mitigate visual and physical gaps, we utilize extensive domain randomization and adopt a quasi-static assumption [44]. We also observe that while modern rendering excels in RGB photorealism, it struggles to replicate sensor-specific depth artifacts, positioning stereo as a more robust modality for geometric sim-to-real transfer.

Furthermore, to improve viewpoint robustness, we rethink the perception-to-action mapping by decoupling it into camera-frame spatial perception and robot-frame action generation, enhanced respectively by two co-training tasks. First, **interaction-region depth estimation** guides the model to predict depth within interaction regions, rather than uninformative backgrounds, to learn task-relevant spatial details. Second, **camera parameter estimation** enables the model to infer camera parameters, learning the mapping between camera and robot frames implicitly.

In summary, our contributions are as follows: a) we present StereoVLA, a geometry-aware VLA system that exploits rich geometric cues from stereo vision by integrating stereo sensing, stereo-aware representation learning, auxiliary geometric supervision, and large-scale synthetic-data pre-training for robust and precise manipulation, b) we introduce GeoSem Vision Encoder to produce visual tokens that carry dense semantic and geometric features from stereo inputs, c) we propose co-training with interaction-region depth estimation and camera parameter estimation tasks to facilitate perception-to-action mapping with enhanced viewpoint robustness, d) we show that our approach outperforms existing VLAs by a large margin in diverse tasks under stereo settings, and is robust to near-hemispheric camera pose variations. Evaluation shows that our approach significantly outperforms baselines under stereo configuration, achieving a 33.4% higher success rate on general tasks and notable improvements in precision-demanding scenarios, and demonstrates robustness to near-hemispheric camera pose variations. Ablation studies validate the effectiveness of the proposed components.

II. Related Work

A. Vision-Language-Action Models

Recent advances in vision-language models [14, 26, 35, 58, 2, 1, 66] and large-scale robot datasets [47, 7, 27, 18, 64] have accelerated Vision-Language-Action (VLA) models [84, 28, 5, 49, 6, 62, 40, 33, 13, 11, 63] for generalizable multimodal reasoning and control. Early methods [84, 28] discretized actions for autoregressive prediction, whereas subsequent work [5, 49, 6, 62, 40, 33, 13, 11] increasingly adopts diffusion or flow-based policies for smooth trajectories.

VLA camera setups have also evolved from single RGB views [84, 28] to wrist-mounted monocular cameras [5, 49, 62, 40, 11], side-facing views [13, 33], and multi-view layouts such as the four-corner configuration in $\pi_{0.5}$ [6]. Despite this shift toward multi-view sensing, prior VLA work has not leveraged stereo cameras which can provide explicit geometric cues via binocular disparity.

Synthetic data is becoming a key ingredient for scalable VLA training, alleviating real-world data collection costs. With modern simulators and automated generation pipelines [19, 45, 10, 75, 56], recent efforts [13, 42] show that combining synthetic trajectories with strong vision-language backbones improves real-world manipulation. Complementary lines of work explore real-to-sim-to-real augmentation to increase data diversity and robustness [12, 82, 71, 43, 74, 76, 59, 60].

B. Robot Learning with Geometry Cues

To address the limits of monocular RGB perception, a growing line of work [79, 78, 23] incorporates explicit geometry cues such as depth, point clouds, and reconstructed 3D structure. These geometry-aware inputs provide stronger spatial priors and improve robustness and generalization. For example, DP3 [79] shows point-cloud observations can be more robust to lighting and object variations, and iDP3 [78] further improves robustness by using a robot-centered egocentric frame that better handles environmental changes and supports both egocentric and third-person viewpoints.

Motivated by these advances, recent VLA methods have started to combine geometry with language and action to mitigate the spatial reasoning limitations of purely 2D models [83, 51, 30, 32, 73, 24, 37, 50, 55, 77, 34, 72, 17, 65, 80]. Representative examples include PointVLA [30], which injects depth cues from RGB-D or monocular depth estimation, and BridgeVLA [32], which projects 3D observations into multi-view 2D feature maps for action prediction. However, existing methods still leave open how to best fuse explicit geometry with the rich 2D representations of pretrained vision foundation models.

C. Stereo Perception and Policies for Robot Learning

Deep stereo matching has progressed from accurate but memory-intensive cost-volume 3D CNNs [9, 52, 53, 69, 36, 81] to iterative refinement [39, 68, 31, 25, 20, 70] avoiding explicit 4D volumes; recent work such as FoundationStereo [61] further improves generalization via large-scale pretraining.

Stereo has also been used for manipulation policies, including stereo visual servoing [16], distillation-based visuomotor control (DextrAH-RGB) [54], and synthetic-data-driven perception for robust grasping (SimNet) [29].

Different from these typically small-scale stereo-policy frameworks, we scale stereo perception to large VLA models, combining model capacity with vision-language grounding for stronger spatial awareness and cross-task generalization.

III. Method

The overall architecture of StereoVLA is illustrated in Figure 2(a). The **GeoSem Vision Encoder** module leverages state-of-the-art vision foundation models to encode the stereo images into visual tokens that capture both semantic and geometric information. These visual tokens, together with language tokens, are passed to a pre-trained large language model, InternLM-1.8B [8], for vision-language joint processing. Leveraging the intermediate vision-language features (i.e., the key-value cache), a 300M-parameter action expert predicts action chunks via flow-matching [38, 5] with a delta end-effector pose representation. To improve viewpoint robustness, we introduce two synergistic tasks, **interaction-region depth estimation** and **camera parameter estimation**, co-trained with action prediction to enhance spatial reasoning while preserving computational efficiency.

A. GeoSem Vision Encoder

As shown in Figure 2(b), given a stereo image pair $I_l, I_r \in \mathbb{R}^{H\times W\times 3}$, we carefully extract geometric features from FoundationStereo [61] from both views and semantic-rich features from the monocular left view I_l. These features are then fused into a sequence of visual tokens that jointly encode both geometric and semantic information.

Geometric Feature Extraction. FoundationStereo is tailored for depth estimation with several specialized modules, making feature adaptation to our task non-trivial. We therefore first outline the key components of FoundationStereo and then explain our feature choices.

For stereo image pairs I_l, I_r, FoundationStereo first computes monocular features f_l, f_r for each image with a unary feature extractor. These features are then processed via two parallel paths. In the first path, f_l, f_r are concatenated to form a 4D cost volume $V_c \in \mathbb{R}^{C\times \frac{D}{4}\times \frac{H}{4}\times \frac{W}{4}}$, where H and W denote the image height and width, D is the maximum disparity range considered, and C is the feature dimension. To model long-range correlations, the cost volume undergoes an attention-based hybrid cost filtering module, yielding a filtered cost volume V_c' with the same dimensions. In the second path, the dot product of the left and right unary features f_l, f_r are computed to generate a correlation volume $V_{corr} \in \mathbb{R}^{\frac{W}{4}\times \frac{H}{4}\times \frac{W}{4}}$. Using the filtered cost volume V_c' and correlation volume V_{corr}, FoundationStereo iteratively refines disparity estimates, resulting in a final disparity map.

When selecting features, we consider two criteria: (1) they should be dense feature volumes and capture rich geometric information, and (2) additional computation for depth estimation should be minimized.

Guided by these criteria, we choose the filtered cost volume V_c' as our geometric feature source. Although the iterative refinement in FoundationStereo further improves disparity estimation, it introduces considerable computational overhead. Similarly, the correlation volume V_{corr} consists of matching scores rather than dense features, so we discard the second computation path. From the first path, we choose V_c' over the raw cost volume V_c because it incorporates long-range correlations via the hybrid cost filtering module, providing dense geometric features crucial for manipulation.

Semantic Feature Extraction. As FoundationStereo is pretrained for depth estimation, it lacks the semantic and appearance information for effective vision-language grounding. To overcome this, we extract these features using SigLIP and DINOv2 following PrismaticVLM [26]. SigLIP, trained with a vision-language contrastive objective, effectively captures high-level semantics. DINOv2, trained with self-supervised discriminative objectives that align features across different image crops, excels at capturing visual details [46, 26]. Due to the high redundancy of the semantic information between left and right views, we apply SigLIP and DINOv2 exclusively on the left view for computational efficiency.

Feature Fusion. The spatial resolution of FoundationStereo features differ with those of SigLIP and DINOv2. We therefore

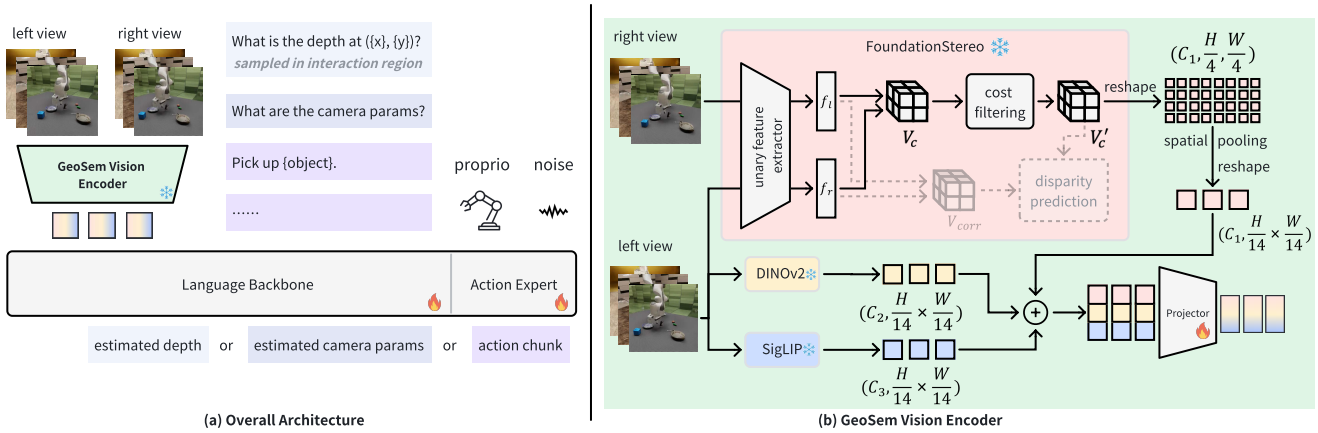


Fig. 2: (a) In StereoVLA, a stereo image pair is encoded by the GeoSem Vision Encoder module to generate visual tokens with geometric precision and semantic richness. Together with language tokens, they are processed by a large language model backbone (InternLM-1.8B). An action expert predicts delta end-effector poses, while auxiliary depth and camera parameter estimation tasks further enhance spatial reasoning during training. (b) The GeoSem Vision Encoder module extracts geometric features with FoundationStereo (bypassing disparity prediction components for efficiency) and semantic-rich features with SigLIP and DINOv2, then fuses them into a unified visual representation with an MLP projector.

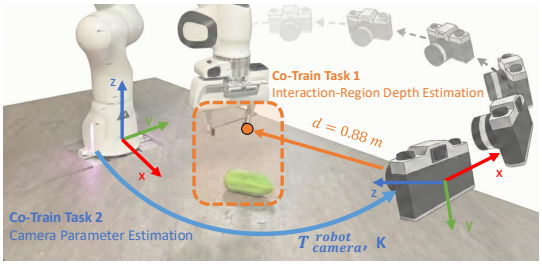


Fig. 3: **Synergistic co-training tasks for viewpoint robustness.** Task 1 performs Interaction-Region Depth Estimation (d) for precise spatial reasoning. Task 2 performs Camera Parameter Estimation (T^{robot}_{camera}, K) to align perception and action coordinate systems.

spatially pool the FoundationStereo features to match the 14-pixel stride of the other two. The final hybrid representation is obtained by concatenating the feature maps along the channel dimension. We avoid token sequence concatenation, as it increases the number of tokens and computational overhead during both training and inference.

B. Co-training Tasks for Viewpoint Robustness

While many existing VLA models operate as "black boxes", we decouple the perception-action mapping into camera-frame spatial perception and robot-frame action generation. To improve viewpoint robustness without increasing inference latency, we introduce two synergistic tasks co-trained alongside action chunk prediction, illustrated in Figure 3.

Interaction-Region Depth Estimation. To enhance fine-grained geometric awareness, we utilize an auxiliary task where the model predicts the metric depth d for sampled coordinates (x, y). Rather than uniform sampling, which often captures task-irrelevant background (e.g., walls), we employ task-aware sampling focused on the interaction region. This region, defined by the 2D bounding boxes of the gripper and target object, forces the model to prioritize local spatial relationships, improving precision and training convergence.

Formally, for each sampled (x, y), we query the model with *"What is the depth at x, y?"* and cast depth prediction as text generation [1, 22]. This avoids task-specific regression heads or discretized bins, preserving a unified architecture and leveraging pretrained knowledge. We supervise the generated depth string with cross-entropy loss $\mathcal{L}_{depth}$.

Camera Parameter Estimation. Our setup features a near-hemispheric orbital camera range, a scope largely underexplored in prior literature, requiring the model to reason about the spatial relationship between the camera and the robot. We supervise the prediction of camera extrinsics T^{robot}_{camera} and the intrinsic K alongside action prediction. This co-training strategy encourages camera-aware representation learning under camera pose variations. The model implicitly strengthens its spatial reasoning capabilities through this co-training task. Combined with the single-step visual input design, it achieves robustness to viewpoint changes during task execution.

Specifically, the model is queried with the prompt: *"What are the camera extrinsics and intrinsics?"*. It is then required to generate a string representation of the ground-truth parameters, comprising the translation (x, y, z), the Euler rotation $(roll, pitch, yaw)$ in the robot frame, and the intrinsic parameters (f_x, f_y, c_x, c_y). Consistent with our depth estimation approach, we cast the continuous value into plain text and supervise it with a standard cross-entropy loss $\mathcal{L}_{camera}$.

C. Dataset Curation and Sim-to-Real Transfer

While large-scale real-world robotic datasets, such as Open-X [47] and AgiBot-World [7], often lack stereo image pairs, recent studies (e.g., GraspVLA, InternData-A1, Genie Sim 3, RoboTwin 2.0) [13, 56, 75, 10] have highlighted the

promising scalability of using synthetic data as a primary source for training VLA models. By integrating high-fidelity modern simulation with vision-language foundation models, these works demonstrate robust zero-shot sim-to-real transfer. Inspired by these findings, we curate a large-scale synthetic dataset comprising 5 million stereo trajectories, covering both a table-top Franka and a humanoid robot to evaluate our method's versatility across different robot morphologies.

We mitigate sim-to-real gaps in both appearance and dynamics. For visuals, we adopted three key designs: (i) keep pretrained vision encoders frozen to retain robust foundation features, (ii) co-train on GRIT [48] with 2D box prediction to better align to real imagery, and (iii) use high-fidelity ray-traced rendering with extensive randomization of lighting, textures, and camera parameters. We also observe that while modern rendering achieves high RGB fidelity, it fails to accurately model sensor-specific depth noise. Consequently, stereo vision emerges as a more robust modality for sim-to-real geometric transfer. For dynamics, we use a quasi-static setting with simple positional control and randomize friction in simulation to improve robustness; dataset details are in the supplementary material.

D. Training Details

Following GraspVLA, we adopt progressive action generation, where the model first predicts the 2D bounding box of the target object and the next keyframe pose of the gripper as intermediate steps to guide action chunk prediction. Since the left and right stereo images contain largely redundant semantic information, we only require the model to predict the bounding box on the left image to reduce computational overhead. The overall training loss is defined as

$$\mathcal{L} = \mathcal{L}_{action} + \mathcal{L}_{depth} + \mathcal{L}_{camera} + \mathcal{L}_{bbox} + \mathcal{L}_{pose}, \quad (1)$$

where $\mathcal{L}_{action}$ is the flow-matching loss for action chunk prediction, and $\mathcal{L}_{depth}$, $\mathcal{L}_{camera}$, $\mathcal{L}_{bbox}$, and $\mathcal{L}_{pose}$ are cross-entropy losses for depth estimation, camera parameter estimation, bounding box prediction, and keyframe pose prediction, respectively. We balance samples from the five tasks with a ratio of 1:2:2:2:1. The final model is trained for 300k steps on 32 NVIDIA H800 GPUs with a batch size of 384 and a learning rate of 1.6e-4.

IV. Evaluation

We conduct comprehensive experiments to answer three key questions: 1) How does StereoVLA compare to existing VLAs on representative manipulation tasks? 2) How does the stereo camera setting compare to other camera settings under comparable training conditions with state-of-the-art VLAs? 3) What is the individual contribution of our proposed design components to the overall performance?

A. Real-world Experiments

1) Task Suite: We evaluate our model across six task categories: Common, Distractor, Environment, Transparent Object, Small Object, and Bar-shaped Object tasks. This suite systematically assesses manipulation capabilities, robustness to visual interference, and high-precision execution. We detail each task category below.

Common Tasks. These general tasks involve pick-and-place and stacking operations with daily items (e.g., bowls, cubes). They verify the model's ability to map language instructions to basic motor skills and spatial coordination.

Distractor Tasks. Target objects are placed in cluttered scenes alongside nine irrelevant items of similar appearance. This category evaluates the model's visual grounding and its ability to distinguish target objects from visual distractors.

Environment Tasks. We test generalizability by varying backgrounds and lighting conditions. These tasks evaluate the model's robustness to real-world variations.

Transparent Object Tasks. Focusing on materials like plastic, these tasks provide a detailed comparison with depth-based baselines regarding the handling of objects with special material properties.

Small Object Grasping Tasks. Manipulating small objects ($1 \sim 2$ cm) demands millimeter-level precision. These tasks evaluate the fine-grained integration of high-resolution perception and delicate motor control.

Bar-shaped Object Grasping Tasks. Bar-shaped objects like forks have a significantly larger length compared to their width and thickness. Their short axis requires precise estimation of the spatial relationship between the gripper and the object, while the long axis often leads to visual overlap with the gripper in camera views, misleading the model into grasping at improper positions. To systematically evaluate the performance, we test with objects oriented at $0°$, $45°$, and $90°$ relative to the camera.

2) Hardware Setup: The hardware setup is shown in Figure 5. Target objects and distractors are placed in a workspace sizing $0.5m \times 0.4m$. To systematically evaluate the impact of camera settings, we introduce three levels of pose randomization for both front and side cameras, as shown in the figure. Unless otherwise specified, we employ the small randomization range. Results for all randomization ranges are reported in Section IV-B.

3) Baselines: We select baselines to support two complementary comparisons: stereo RGB-based adapted baselines evaluate how existing VLA architectures perform with the same stereo observations, whereas cross-setting baselines compare system-level performance against representative VLAs under their respective camera configurations. For fairness, all baselines are trained or fine-tuned with matched dataset scale and task distribution.

Architectural comparison. Since no existing VLA models natively support stereo inputs, we adapt the aforementioned SOTAs, $\pi_{0.5}$ and GraspVLA, to process synchronized stereo RGB pairs, denoted as $\pi_{0.5}$**-S** and **GraspVLA-S**. GraspVLA-S is effectively StereoVLA without our stereo-centric designs, serving as a controlled baseline for testing whether a camera swap alone can explain the gains. To avoid bias toward GraspVLA's original front-side configuration, GraspVLA-S is trained from VLM initialization, while $\pi_{0.5}$-S is fine-tuned

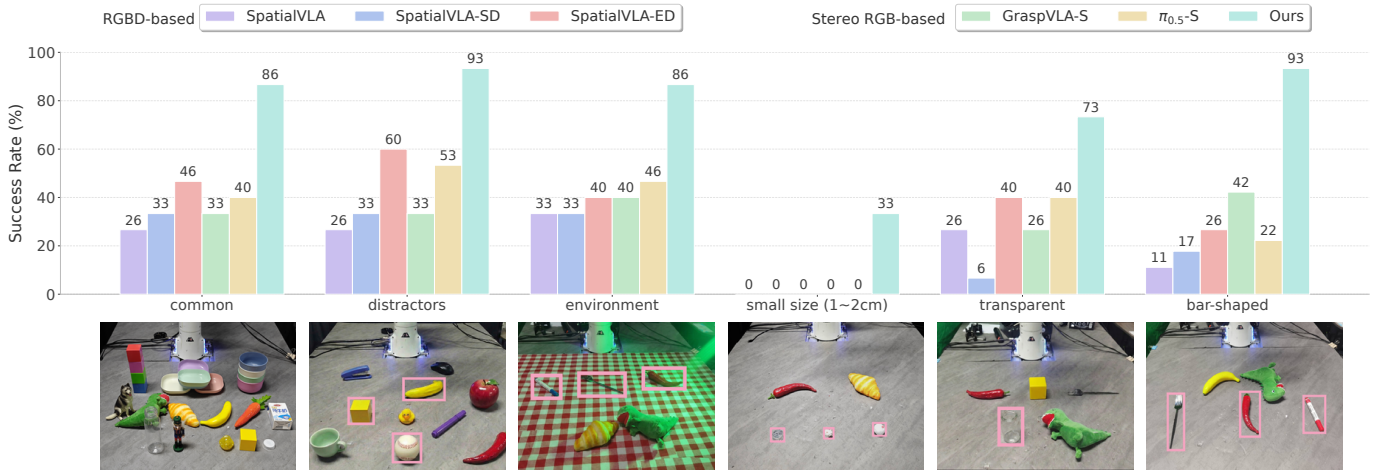


Fig. 4: **Real-world evaluation tasks and results.** We evaluate StereoVLA on a comprehensive suite of tasks requiring fine-grained perception, manipulation and generalizability. Objects enclosed by colored boxes denote the target objects used for evaluation, while other items in the scene serve as distractors. We include RGB-D-based baselines to compare system-level performance with representative depth-aware VLA systems, and Stereo RGB-based baselines to evaluate how different architectures exploit the same stereo observations. All models are trained on an identical synthetic dataset. StereoVLA consistently outperforms baselines across the evaluated tasks, demonstrating its effectiveness.

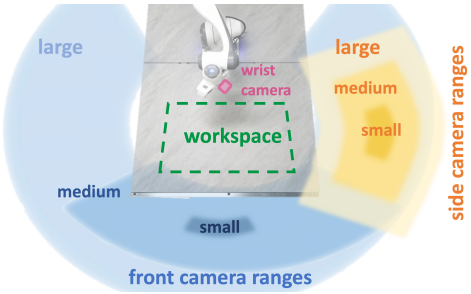


Fig. 5: **Layout of the robot arm, workspace, and cameras.** To evaluate viewpoint robustness, cameras are randomized within three spherical shell regions, ranging from localized sectors to a near-hemispheric volume. While these are 3D spaces, we provide only their top-down projections for visual clarity. See the Appendix for geometric parameters.

using its recommended hyperparameters.

Cross-setting comparison. We compare against representative SOTA models for different modalities. SpatialVLA [51] is selected for the RGB-D setup, as it explicitly uses depth as positional embedding to enhance spatial reasoning. For the front-wrist configuration, we employ $\pi_{0.5}$ [6], a generalist VLA recognized for its robustness and generalizability. Finally, we adopt GraspVLA [13] to represent the front-side setup.

Depth-based variants. To systematically assess the impact of depth quality, we address SpatialVLA's [51] reliance on potentially inaccurate monocular estimation by introducing two variants: **SpatialVLA-SD** (Sensor Depth), which utilizes raw depth from the Intel RealSense D435 sensor, and **SpatialVLA-ED** (Estimated Depth), which employs high-quality depth maps predicted from stereo RGB pairs via FoundationStereo.

4) Metrics: Task success rates are reported for each test set, with 15 trials per set, yielding a total of $15 \times 6 \times 5 = 450$ trials. To rigorously assess fine-grained spatial perception and manipulation, we incorporate three additional criteria. First, to prevent models from succeeding through repeated attempts, each trial allows only a single execution: one gripper-close before grasping and one gripper-open after grasping. Second, we disable the "gripper sticking" heuristic used in OpenVLA, SpatialVLA, and $\pi_{0.5}$, since it artificially delays gripper actions and can conceal inaccurate decisions. Third, a trial is deemed successful only when the task is fully completed, with no partial credit assigned. While these stricter criteria produce lower success rates than those reported in prior work, they provide a more accurate measure of model capability.

5) Results: Figure 4 shows the success rates across the evaluation tasks. Our model consistently outperforms all baselines across every task, demonstrating robustness to visual distractors and complex environments. In the rigorous small-size object ($1 \sim 2$ cm) grasping task where all baselines completely fail (0%), our model achieve a 33% success rate. Performance in this regime is currently limited by the 224×224 resolution, where small objects span only a few pixels; while higher-resolution backbones could enhance semantic grounding, they introduce a computational trade-off critical for future optimization.

Analysis of depth-based baselines. Our evaluation of depth-based models shows that performance scales from monocular estimation (lowest) to sensor depth, and finally to stereo estimation (highest). This is likely due to the sim-to-real divergence: simulated depth is noise-free, whereas sensor and monocular data contain significant artifacts. Furthermore, sensors fail on transparent surfaces due to their underlying physical principles, causing the model to struggle with these specific objects. While FoundationStereo [61] produces depth

modalities	viewpoint	model	small	medium	large	moving
RGBD	front	SpatialVLA-ED [51]	35.6±6.9%	6.7±3.7%	3.9±3.0%	2.8±2.6%
RGBs	front + wrist	$\pi_{0.5}$ [6]	65.6±6.9%	<u>51.7±7.2%</u>	<u>42.2±7.1%</u>	<u>39.4±7.1%</u>
RGBs	front + side	GraspVLA [13]	<u>71.7±6.5%</u>	23.3±6.1%	11.1±4.6%	7.2±3.8%
stereo RGBs	front	StereoVLA	**77.2±6.1%**	**62.2±7.0%**	**53.9±7.2%**	**52.2±7.2%**

TABLE I: **Comparison of camera settings.** StereoVLA demonstrates strong robustness to camera pose variations. To ensure fairness, all models are trained on the same amount of synthetic data for each camera randomization range. Section IV-B details the experimental setup. Best results are **bolded**, second-best results are <u>underlined</u>. Each reported success rate is estimated from three evaluation groups, each covering 6 tasks with 10 trials per task, for 180 real-world robot trials in total. The ± values denote 95% Wilson confidence intervals.

feature	w/o sem.	w/ sem.
V_{corr}	27.0%	54.0%
V_c	40.0%	69.0%
V_c'	51.0%	**77.0%**

TABLE II: **Comparison of stereo feature selections.** The adoption of the filtered cost volume V_c' and the semantic feature results in the highest success rate.

quality on par with simulation, their iterative optimization process results in a 70ms inference overhead. This latency makes them unsuitable for real-time action control.

6) Qualitative Analysis: $\pi_{0.5}$ and the original SpatialVLA model often approached correct target objects, but consistently closed the gripper too early, likely due to the lack of precise spatial perception. In contrast, StereoVLA achieves the highest success rate on all tasks, demonstrating its superior performance. Interestingly, SpatialVLA-ED struggled despite utilizing superior stereo-derived depth. This suggests that the downsampling of depth maps during spatial encoding may result in the loss of critical geometric details. Meanwhile, GraspVLA-S, which encodes left and right images independently, performed well on bar-shaped objects but failed significantly on more complex tasks. We hypothesize that while GraspVLA's additional pose supervision enhances general spatial awareness, it remains unable to capture fine-grained spatial relationships from the subtle disparities between stereo views.

Overall, these results suggest that the superior performance of StereoVLA stems from methodological innovations rather than from data scale or quality, emphasizing the critical role of model architectures and training strategies in the stereo setting.

B. Comparison of Camera Settings

This section provides a system-level comparison of representative VLA methods across common camera settings. Building on the baselines introduced in Sec. IV-A3, we follow each method's original model design and training protocol, with matched dataset scale and task distribution across methods. We evaluate four common configurations [51, 6, 28, 13, 47]: single-view RGBD (SpatialVLA-ED), stereo RGBs (StereoVLA), front + wrist RGBs (fine-tuned $\pi_{0.5}$), and front + side RGBs (GraspVLA).

To evaluate the robustness to camera pose variations, we generate equal-sized datasets across three randomization ranges (Figure 5) and train/fine-tune each model accordingly. To further account for potential motion blur induced by camera movement and vibrations, we include a 'moving' test track where the camera traverses from left to right relative to the robot. In all evaluations, we maintain the gripper and objects within the frame to ensure they are visible. The randomization ranges are defined as spherical shells centered on the workspace. For the front view:

- The **small range** corresponds to an elevation range of $20^\circ \sim 30^\circ$, an azimuth range of $-10^\circ \sim 10^\circ$, and a radius range of $1.2\,\text{m} \sim 1.3\,\text{m}$.
- The **large range** covers an elevation of $15^\circ \sim 45^\circ$ and a radius of $0.9\,\text{m} \sim 1.6\,\text{m}$. Instead of a full 2π azimuth, we restrict the range to $-150^\circ \sim 150^\circ$ to exclude the region directly behind the robot, thereby avoiding self-occlusion caused by the robotic arm.

Detailed geometric parameters are provided in the Appendix, and average task success rates are reported in Table I. While the front + side configuration with GraspVLA performs well under small randomization, it suffers a significant drop in larger randomization due to the difficulty of extracting coherent spatial cues from unaligned views. In this evaluation, StereoVLA obtains the highest success rates across the tested randomization ranges, indicating robustness to camera pose variations under the stereo setup.

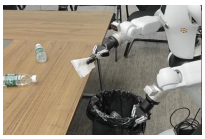
Regarding the "front + wrist" configuration ($\pi_{0.5}$), although it starts with a lower baseline under small randomization, it exhibits higher resilience to increasing pose variations. This is because the wrist camera provides egocentric views of gripper-object interactions that remain invariant to global camera shifts. However, this setup is prone to premature gripper closure, a failure mode likely caused by the wrist camera's limited precise spatial perception. Finally, RGBD-based models consistently underperform, particularly in large randomization, likely due to the lack of current patch-level depth encoding methods in capturing fine-grained spatial cues.

We believe that each camera setting has potential depending on the deployment scenario. Within this context, our results suggest that the stereo configuration is a promising alternative for balancing task performance, robustness to camera pose variation, and deployment simplicity.

C. Ablation of Key Design Components

Stereo	GeoSem	DE	CPE	small	medium	large	moving
✓				29.0	-	-	-
✓	✓			57.8	36.7	13.3	10.0
✓	✓	✓		74.9	51.8	30.2	27.1
✓	✓	✓	✓	**77.3**	**62.1**	**54.1**	**52.3**

TABLE IV: **Ablation of key design components.** DE refers to interaction-region Depth Estimation. CPE refers to Camera Parameter Estimation.




Model	Cleaning	Laundry
w/o pre-train	26.6%	13.3%
w/ pre-train	**73.3%**	**53.3%**

TABLE III: **Humanoid Fine-Tuning.** The model with pre-training demonstrates notably higher success rates in both scenarios.

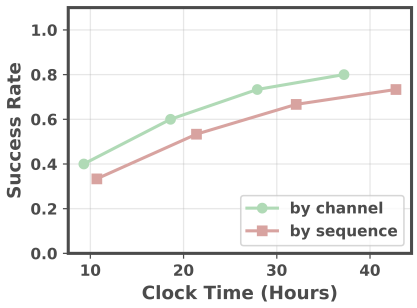


Fig. 6: **Comparison of feature fusion methods.** Sequence concatenation results in a lower success rate and requires a longer training time.

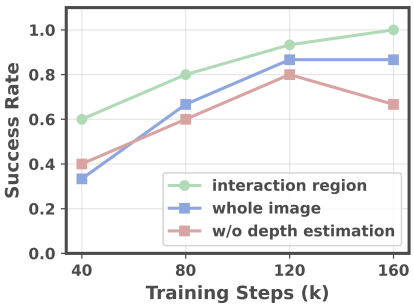


Fig. 7: **Comparison of depth estimation methods.** Depth estimation within the interaction region yields the highest success rate.

Table IV summarizes the component-wise ablation. We use GraspVLA-S, introduced in Sec. IV-A3, as the stereo-only baseline and progressively add GeoSem, DE, and CPE to isolate their contributions. The stereo-only baseline achieves moderate performance, indicating that a camera swap alone cannot explain the overall gains. GeoSem and DE provide substantial improvements by enhancing stereo geometry encoding and interaction-region spatial supervision, respectively. CPE further improves robustness under large and moving camera pose shifts, leading to the strongest overall performance.

D. Design Choices of GeoSem Vision Encoder

We conduct ablation experiments on the GeoSem Vision Encoder module to evaluate the effectiveness of our feature selection and fusion strategies.

For the selection of FoundationStereo features, we evaluate models in simulation after 100k training steps, averaging the performance of the three most recent checkpoints at 10k-step intervals. As shown in Table II, success rates progressively improve when using V_{corr}, V_c, and V_c', respectively. This supports our hypothesis that V_{corr} retains little geometric information, whereas the unfiltered cost volume V_c lacks the long-range spatial relationships present in the filtered version V_c'. The superior performance of V_c' validates our choice of this representation as the stereo feature.

In addition, the results show that incorporating the semantic feature consistently yields higher success rates than using the stereo feature alone. Manual inspection of the rollouts revealed that models without the semantic feature more frequently grasped incorrect objects, indicating that semantic features is essential for vision-language grounding. This justifies the necessity of fusing both features.

As is shown in Figure 6, for feature fusion strategies, concatenating features by sequence yields slightly lower success rates and increases computational overhead due to the doubled number of visual tokens. Channel-wise concatenation provides both efficiency and performance advantages.

In addition to FoundationStereo [61], other foundation models, such as VGGT [57], can be alternatives for stereo feature extraction. We compare the depth estimation precision of VGGT and FoundationStereo on our dataset to assess their performance in manipulation scenarios. VGGT obtains an AbsRel [15] error of 0.4121, whereas FoundationStereo achieves a substantially lower error of 0.0275. These results indicate that FoundationStereo provides more reliable depth information, likely due to its specialization in stereo inputs. In contrast, VGGT may excel in broader multi-view settings without stereo assumptions.

E. Effectiveness of Co-training Tasks

1) Effectiveness of Interaction-Region Depth Estimation: We compare two alternatives: predicting the depth of points uniformly sampled in the whole image, and no prediction of depth. As shown in Figure 7, interaction-region depth estimation facilitates model learning and achieves the highest performance. In contrast, predicting depth over the entire image negatively impacts training in the early stages, potentially due to the additional burden of learning irrelevant background depths that do not contribute to manipulation success. Nevertheless, both strategies stabilize training in later stages. This suggests that explicitly supervising the model to perceive critical spatial details is beneficial.

2) Effectiveness of Camera Parameter Estimation: Quantitative results in Table IV indicate that this task substantially improves viewpoint robustness, with the most significant gains observed in the large and moving categories. The model's stability under moving camera conditions is notable. We hypothesize this stems from the synergy between viewpoint randomization during training and the model's dependence solely on current-step observations.

F. Fine-Tuning on a Humanoid Platform

Motivated by the similarity between stereo vision and human visual perception, we extend our analysis to a humanoid platform and evaluate performance on two household tasks: Cleaning and Laundry. Using 500 trajectories for fine-tuning and 15 trials for evaluation, the results in Table III demonstrate that pre-training significantly improves downstream adaptation. However, Laundry achieves a lower success rate than Cleaning, which we attribute to the distributional shift from the table-top-only pre-training data. See the Appendix for details.

G. Inference Speed

We profiled the model to quantify inference time overhead of each component on a NVIDIA H800 GPU. A full FoundationStereo forward pass to predict depth takes 68 ms per iteration, while extracting the filtered cost volume V_c' requires only 14 ms, reducing computation by 54 ms. We therefore use V_c' as the stereo feature representation, yielding a vision-encoder inference time of 27 ms.

	vision	language	action	total
time (ms)	27	81	52	168

TABLE V: **Breakdown on inference time.**

V. Limitation and Future Work

Despite promising results, our work has two main limitations. First, the input resolution of 224×224 constrains performance on small objects; exploring higher resolutions while balancing computational cost is an important direction. Second, the model does not yet capture long-horizon temporal dependencies, which limits its ability to solve more complex, multi-stage tasks.

VI. Conclusion

We introduced StereoVLA, a vision-language-action model that exploits rich geometric cues from stereo vision to achieve robust and precise manipulation. Experiments demonstrate that our approach outperforms prior methods by a large margin in the stereo setting and is robust to camera pose variations. Furthermore, our comparison of camera setups provides practical guidance for balancing task performance and deployment overhead in future studies.

Acknowledgments

This work was partially supported by New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122905), and National Key R&D Program of China (2022ZD0160201). Additionally, we thank Yitao Zeng for assistance with the experiments.

References

[1] 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, 2025. URL https://arxiv.org/abs/2511.21631.

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

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[4] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[6] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. π0. 5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[7] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

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

[9] Liyan Chen, Weihan Wang, and Philippos Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17235–17244, 2023.

[10] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation, 2025. URL https://arxiv.org/abs/2506.18088.

[11] InternVLA-M1 Contributors. Internvla-m1: Latent spatial grounding for instruction-following robotic manipulation, 2025.

[12] Prithwish Dan, Kushal Kedia, Angela Chao, Edward Weiyi Duan, Maximus Adrian Pace, Wei-Chiu Ma,

and Sanjiban Choudhury. X-sim: Cross-embodiment learning via real-to-sim-to-real, 2025. URL https://arxiv.org/abs/2505.07096.

[13] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025.

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, 2014.

[16] Albert Enyedy, Ashay Aswale, Berk Calli, and Michael Gennert. Stereo image-based visual servoing towards feature-based grasping. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7325–7331. IEEE, 2024.

[17] Qingyu Fan, Zhaoxiang Li, Yi Lu, Wang Chen, Qiu Shen, Xiao xiao Long, Yinghao Cai, Tao Lu, Shuo Wang, and Xun Cao. Peafowl: Perception-enhanced multi-view vision-language-action for bimanual manipulation, 2026. URL https://arxiv.org/abs/2601.17885.

[18] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

[19] Ning Gao, Yilun Chen, Shuai Yang, Xinyi Chen, Yang Tian, Hao Li, Haifeng Huang, Hanqing Wang, Tai Wang, and Jiangmiao Pang. Genmanip: Llm-driven simulation for generalizable instruction-following manipulation, 2025. URL https://arxiv.org/abs/2506.10966.

[20] Rui Gong, Weide Liu, Zaiwang Gu, Xulei Yang, and Jun Cheng. Learning intra-view and cross-view geometric knowledge for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20752–20762, 2024.

[21] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.

[22] Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification, 2025. URL https://arxiv.org/abs/2510.13054.

[23] Siddhant Haldar, Lars Johannsmeier, Lerrel Pinto, Abhishek Gupta, Dieter Fox, Yashraj Narang, and Ajay Mandlekar. Point bridge: 3d representations for cross domain policy learning, 2026. URL https://arxiv.org/abs/2601.16212.

[24] Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilue Wang, Longzan Luo, Lily Lee, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, and Shanghang Zhang. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation, 2024. URL https://arxiv.org/abs/2411.18623.

[25] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal. Uncertainty guided adaptive warping for robust and efficient stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3318–3327, 2023.

[26] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

[27] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[29] Thomas Kollar, Michael Laskey, Kevin Stone, Brijen Thananjeyan, and Mark Tjersland. Simnet: Enabling robust unknown object manipulation from pure synthetic data via stereo. In *Conference on Robot Learning*, pages 938–948. PMLR, 2022.

[30] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511*, 2025.

[31] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022.

[32] Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. *arXiv preprint arXiv:2506.07961*, 2025.

[33] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

[34] Yixuan Li, Yuhui Chen, Mingcai Zhou, Haoran Li, Zhengtao Zhang, and Dongbin Zhao. Qdepth-vla: Quantized depth prediction as auxiliary supervision for vision-

language-action models, 2025. URL https://arxiv.org/abs/2510.14836.

[35] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-2: Faster inference of language models with dynamic draft trees. *arXiv preprint arXiv:2406.16858*, 2024.

[36] Bifa Liang, Yichao Wang, Zhicong Huang, Ziyang Hu, Haifeng Hu, Jianming Xu, and Dihu Chen. Arunet: Advancing real-time stereo matching for robotic perception on edge devices. *IEEE Robotics and Automation Letters*, 10(9):8970–8977, 2025. doi: 10.1109/LRA.2025.3589145.

[37] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025.

[38] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[39] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[40] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[41] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[42] Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, Scott Reed, Ken Goldberg, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation, 2025. URL https://arxiv.org/abs/2503.24361.

[43] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations, 2023. URL https://arxiv.org/abs/2310.17596.

[44] Matthew T Mason. *Mechanics of robotic manipulation*. MIT press, 2001.

[45] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots, 2024. URL https://arxiv.org/abs/2406.02523.

[46] 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. *arXiv preprint arXiv:2304.07193*, 2023.

[47] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[48] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.14824, 2023.

[49] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[50] Quanhao Qian, Guoyang Zhao, Gongjie Zhang, Jiuniu Wang, Ran Xu, Junlong Gao, and Deli Zhao. Gp3: A 3d geometry-aware policy with multi-view images for robotic manipulation. *arXiv preprint arXiv:2509.15733*, 2025. doi: 10.48550/arXiv.2509.15733.

[51] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[52] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2021.

[53] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European conference on computer vision*, pages 280–297. Springer, 2022.

[54] Ritvik Singh, Arthur Allshire, Ankur Handa, Nathan Ratliff, and Karl Van Wyk. Dextrah-rgb: Visuomotor policies to grasp anything with dexterous hands. *arXiv preprint arXiv:2412.01791*, 2024.

[55] Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models, 2025. URL https://arxiv.org/abs/2508.09071.

[56] Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy, 2025. URL https://arxiv.org/abs/2511.16651.

[57] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

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

[59] Wenhao Wang, Kehe Ye, Xinyu Zhou, Tianxing Chen, Cao Min, Qiaoming Zhu, Xiaokang Yang, Ping Luo, Yongjian Shen, Yang Yang, Maoqing Yao, and Yao Mu. Fieldgen: From teleoperated pre-manipulation trajectories to field-guided data generation, 2025. URL https://arxiv.org/abs/2510.20774.

[60] Yiming Wang, Ruogu Zhang, Minyang Li, Hao Shi, Junbo Wang, Deyi Li, Jieji Ren, Wenhai Liu, Weiming Wang, and Hao-Shu Fang. Exogs: A 4d real-to-sim-to-real framework for scalable manipulation data collection, 2026. URL https://arxiv.org/abs/2601.18629.

[61] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025.

[62] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025.

[63] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Tinyvla: Toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics Autom. Lett.*, 10(4):3988–3995, 2025. doi: 10.1109/LRA.2025.3544909. URL https://doi.org/10.1109/LRA.2025.3544909.

[64] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

[65] Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Shuai Zhou, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Zechen Wang, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A pragmatic vla foundation model, 2026. URL https://arxiv.org/abs/2601.18692.

[66] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[67] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation, 2023. URL https://arxiv.org/abs/2307.03659.

[68] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21919–21928, 2023.

[69] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1959–1968, 2020.

[70] Yuanfan Xu, Shuaiwen Chen, Xinting Yang, Yunfei Xiang, Jincheng Yu, Wenbo Ding, Jian Wang, and Yu Wang. Efficient and hardware-friendly online adaptation for deep stereo depth estimation on embedded robots. *IEEE Robotics and Automation Letters*, 10(5): 4308–4315, 2025. doi: 10.1109/LRA.2025.3548504.

[71] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning, 2025. URL https://arxiv.org/abs/2502.16932.

[72] Ganlin Yang, Tianyi Zhang, Haoran Hao, Weiyun Wang, Yibin Liu, Dehui Wang, Guanzhou Chen, Zijian Cai, Junting Chen, Weijie Su, Wengang Zhou, Yu Qiao, Jifeng Dai, Jiangmiao Pang, Gen Luo, Wenhai Wang, Yao Mu, and Zhi Hou. Vlaser: Vision-language-action model with synergistic embodied reasoning, 2026. URL https://arxiv.org/abs/2510.11027.

[73] Rujia Yang, Geng Chen, Chuan Wen, and Yang Gao. Fp3: A 3d foundation policy for robotic manipulation. *arXiv preprint arXiv:2503.08950*, 2025.

[74] Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation, 2025. URL https://arxiv.org/abs/2504.13175.

[75] Chenghao Yin, Da Huang, Di Yang, Jichao Wang, Nanshu Zhao, Chen Xu, Wenjun Sun, Linjie Hou, Zhijun Li, Junhui Wu, Zhaobo Liu, Zhen Xiao, Sheng Zhang, Lei Bao, Rui Feng, Zhenquan Pang, Jiayu Li, Qian Wang, and Maoqing Yao. Genie sim 3.0 : A high-fidelity comprehensive simulation platform for humanoid robot, 2026. URL https://arxiv.org/abs/2601.02078.

[76] Justin Yu, Letian Fu, Huang Huang, Karim El-Refai, Rares Andrei Ambrus, Richard Cheng, Muhammad Zubair Irshad, and Ken Goldberg. Real2render2real: Scaling robot data without dynamics simulation or robot hardware, 2025. URL https://arxiv.org/abs/2505.09601.

[77] Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Zhuoguang Chen, Tao Jiang, and Hang Zhao. Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning, 2025. URL https://arxiv.org/abs/2510.13375.

[78] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024.

[79] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

[80] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration, 2025. URL https://arxiv.org/abs/2506.22242.

[81] Yakai Zhang and Jinhui Zhang. Gfanet: Group fusion aggregation network for real time stereo matching. *IEEE Robotics and Automation Letters*, 8(7):4251–4258, 2023. doi: 10.1109/LRA.2023.3280818.

[82] Yujie Zhao, Hongwei Fan, Di Chen, Shengcong Chen, Liliang Chen, Xiaoqi Li, Guanghui Ren, and Hao Dong. Real2edit2real: Generating robotic demonstrations via a 3d control interface, 2025. URL https://arxiv.org/abs/2512.19402.

[83] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[84] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.