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

Universal Pose Pretraining for Generalizable Vision-Language-Action Policies

Haitao Lin[1,2*,†] Hanyang Yu[3*,‡] Jingshun Huang[4,5*,‡] He Zhang[1,2] Yonggen Ling[1,2]
Ping Tan[3†] Xiangyang Xue[4] Yanwei Fu[4,5†]

[1] Tencent Robotics X [2] Futian Laboratory
[3] The Hong Kong University of Science and Technology
[4] Fudan University [5] Shanghai Innovation Institute

Project Page: https://hetolin.github.io/PoseVLA

***Abstract*—Existing Vision-Language-Action (VLA) models often suffer from feature collapse and low training efficiency because they entangle high-level perception with sparse, embodiment-specific action supervision. Since these models typically rely on VLM backbones optimized for Visual Question Answering (VQA), they excel at semantic identification but often overlook subtle 3D state variations that dictate distinct action patterns. To resolve these misalignments, we propose Pose-VLA, a decoupled paradigm that separates VLA training into a pre-training phase for extracting universal 3D spatial priors in a unified camera-centric space, and a post-training phase for efficient embodiment alignment within robot-specific action space. By introducing discrete pose tokens as a universal representation, Pose-VLA seamlessly integrates spatial grounding from diverse 3D datasets with geometry-level trajectories from robotic demonstrations. Our framework follows a two-stage pre-training pipeline, establishing fundamental spatial grounding via poses followed by motion alignment through trajectory supervision. Extensive evaluations demonstrate that Pose-VLA achieves state-of-the-art results on RoboTwin 2.0 with a 79.5% average success rate and competitive performance on LIBERO at 96.0%. Real-world experiments further showcase robust generalization across diverse objects using only 100 demonstrations per task, validating the efficiency of our pre-training paradigm.**

I. Introduction

The pursuit of general-purpose embodied intelligence has been significantly accelerated by the emergence of Vision-Language-Action (VLA) models. State-of-the-art frameworks such as π [5, 6] and the GR00T series [4] have demonstrated remarkable potential in end-to-end robotic manipulation through the adoption of dual-system architectures. In this paradigm, a Vision-Language Model (VLM) acts as a high-level semantic interpreter, while a specialized action expert that employs generative techniques like flow matching [22] serves as the low-level controller for precise action denoising.

Despite these advances, scaling vision-language representations into physically grounded, generalizable action policies remains a fundamental challenge. While existing VLAs leverage off-the-shelf VLMs [3, 2, 33] as perception backbones, their general semantic features often fail to translate into robust downstream policy performance. This discrepancy is reinforced by recent findings [46] showing that fine-tuning on auxiliary tasks like depth estimation or VQA does not inherently guarantee improved control. This persistent gap underscores a critical open question: *how can we efficiently adapt VLMs to acquire transferable embodied priors that directly facilitate downstream policy learning?*

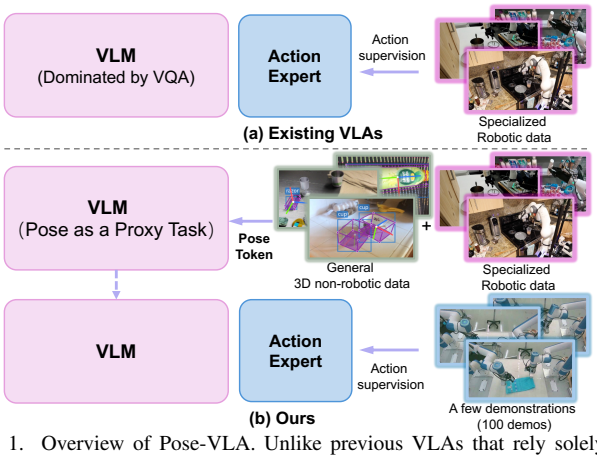

Fig. 1. Overview of Pose-VLA. Unlike previous VLAs that rely solely on sparse action supervision, our approach decouples policy learning into two stages by *using unified pose prediction as a proxy task*: (1) Pre-training, extracting universal 3D spatial priors in a unified camera-centric space; and (2) Alignment, adapting these priors to specific embodiments. This decoupling allows the model to leverage diverse 3D datasets, enabling efficient transfer as backbone when adapting to robotic control with only few-shot fine-tuning.

We observe that this failure is not incidental, but structural. It stems from two persistent misalignments between how VLMs are pretrained and how robotic actions are defined. (1) **Granularity Mismatch**: VLM pretraining is dominated by tasks such as image–text alignment and visual question answering, which emphasize categorical recognition and high-level semantics. Robotic manipulation, in contrast, depends critically on fine-grained 3D state variations, such as subtle changes in pose, contact geometry, or relative motion that demand qualitatively different actions. As a result, a VLM may correctly recognize what an object is while remaining insensitive to how its physical state evolves. (2) **Data Heterogeneity Gap**: Internet-scale visual corpora lack physical grounding, while robotic demonstration datasets are scarce,

* Equal contribution.
† Corresponding authors.
‡ Work done during an internship at Tencent Robotics X.

narrowly distributed, and expensive to collect. Existing VLAs struggle to reconcile these two extremes, preventing them from absorbing diverse spatial experiences while remaining relevant to real-world control.

To address these challenges, we propose a decoupled learning paradigm for Vision–Language–Action modeling. Rather than entangling perception with embodiment-specific actions as in prior VLAs [5, 17], we separate learning into two stages as in Fig. 1: (1) large-scale pretraining of universal spatial priors in a unified, camera-centric observation space, and (2) lightweight post-training for embodiment alignment. This decoupling fundamentally reshapes the objective of large-scale pre-training. By incorporating 3D spatial supervision, the model learns to distinguish subtle 3D variations that dictate distinct action patterns. Once these robust spatial priors are established, the VLM serves as a powerful initialization that makes adapting to a new robotic policy significantly more efficient than learning from scratch.

Building on this principle, we introduce Pose-VLA, a framework that re-centers VLA learning around a universal pose representation. In this architecture, pose acts as a *structural bridge* that maps high-level visual perception to low-level physical execution. Specifically, (1) to bridge the *granularity mismatch*, we represent object states and actions as 3D poses and temporal sequences, explicitly constraining the model to reason over fine-grained spatial supervision. (2) To bridge the *data heterogeneity gap*, pose tokens provide a common language across diverse 3D sources, enabling Pose-VLA to ingest large-scale non-robotic 3D datasets alongside limited robot demonstrations. By unifying these elements, Pose-VLA transforms the VLM into a geometrically-aware backbone optimized for downstream manipulation.

Technically, Pose-VLA builds upon the PaliGemma [3] architecture to unify semantic understanding with spatial grounding. We integrate RGB images with depth maps and camera intrinsics encoded as raymaps to instill intrinsic 3D awareness. A modality masking strategy is employed during training to ensure robustness to varying sensor availability at inference time. We further extend the language model vocabulary with discrete pose tokens that represent 3D transformations in the camera frame, enabling a unified tokenization of spatial information across heterogeneous datasets.

Our training follows a two-stage pipeline. First, spatial foundation pre-training on large-scale 3D datasets establishes geometric grounding. Second, pose alignment pre-training leverages dense multi-view supervision to project robotic trajectories into the same camera-centric observation space. Together, these stages transform the VLM from a semantic describer into a foundation for embodied control. During post-training, we follow [5] by appending a lightweight action expert that maps pretrained representations into robot-specific commands. This strategy ensures that the VLM acquires rich, 3D-aware features that drastically reduce the amount of demonstration data required for downstream policy adaptation.

Our contributions are summarized as follows: (1) we propose a unified VLM framework that integrates RGB images, depth maps, and camera intrinsics to instill intrinsic 3D awareness, facilitating the effective transfer of vision-language knowledge to robotic control; (2) we introduce discrete Pose Tokens as a universal interface (within consistent camera-centric observation space) for aligning and ingesting spatial priors from heterogeneous non-robotic 3D data and specialized robotic demonstrations; and (3) we curate a comprehensive pre-training corpus comprising 1.4M images with 6.5M 3D annotations for spatial grounding, complemented by approximately 1.55M diverse robotic trajectories for motion alignment. (4) we demonstrate that Pose-VLA achieves state-of-the-art performance on RoboTwin 2.0 (avg. 79.5% success) and competitive results on LIBERO (avg. 96.0%); notably, a single multi-task model exhibits robust real-world generalization across rigid, articulated, and deformable objects, requiring as few as 100 demonstrations per task.

II. Related Work

Vision-Language Models for 3D Grounding Recently, an increasing number of approaches have sought to extend general-purpose VLMs for 3D understanding [14, 33, 37, 42, 11, 41]. Several methods focus specifically on bridging the gap between vision-language reasoning and 3D geometric tasks. For instance, SpatialLM [29] performs 3D grounding from point clouds to generate 3D bounding boxes, although it is limited to a small number of object categories. Spatial-Reasoner [27] estimates object positions and orientations as intermediate outputs to enhance spatial awareness, while the concurrent N3D-VLA [40] utilizes a massive depth-lifted dataset to enable 3D localization via bounding boxes, yet lacks explicit orientation modeling. Unlike previous methods, Pose-VLA utilizes pose tokens to unify object poses and motion trajectories within a shared representation. This enables joint pre-training on heterogeneous 3D datasets and robotic demonstrations, while incorporating camera rays and depth as spatial priors to provide the additional geometric awareness for 3D grounding and robotic manipulation.

Vision-Language-Action Models Vision-Language-Action (VLA) models extend the reasoning capabilities of Vision-Language Models (VLMs) to the direct prediction of robotic actions. Current VLA methods can be generally categorized into discrete and continuous action outputs. Discrete approaches [51, 17, 30, 44, 39, 20, 24, 13] treat action prediction as a next-token prediction task in order to align with the native training paradigm of Large Language Models (LLMs). Alternatively, continuous approaches [38, 5, 4, 6, 18, 26, 34, 24, 43, 31] add generative heads like diffusion or flow-matching to achieve higher control fidelity. Despite these advances, how to effectively adapt a VLM's semantic and reasoning abilities to the action modality remains an open question. Recent studies indicate that while VLM pre-training is fundamental, a model's general VQA competence does not always predict its downstream control performance, highlighting a persistent domain gap. To bridge this gap, strategies like co-training with vision-text data [6, 50, 21, 31] or incorporating Embodied Chain-of-Thought (ECoT) [44, 18, 10, 36]

have been proposed to preserve reasoning traces. Furthermore, some VLAs [19, 16, 48] introduce regularization within the VLM's latent space, such as aligning mid-layer features with a vision teacher or utilizing hidden space world modeling. While VLM4VLA [46] indicates that fine-tuning VLMs on auxiliary tasks rarely translates to better policy performance, we show that 3D pose pre-training establishes a vital geometric foundation for downstream control.

III. Method

A. Preliminaries

Given a multi-modal observation $\mathbf{O}$ and a language instruction $\mathbf{L}$, Pose-VLA aims to model the joint probability of an output token sequence $\mathbf{S}$. Following the auto-regressive paradigm, the generation process is formulated as:

$$P(\mathbf{S}|\mathbf{O},\mathbf{L}) = \prod_{k=1}^{K} P(\mathbf{s}_k|\mathbf{O},\mathbf{L},\mathbf{s}_{<k}) \tag{1}$$

where $\mathbf{s}_k$ denotes the k-th token in a sequence of length K, and $\mathbf{L}$ represents the natural language instruction. To bridge high-level semantics with fine-grained 3D control, we define the output as a structured sequence of tuples $\mathcal{S} = (\tau_1, \tau_2, \ldots, \tau_T)$, where each tuple τ_t serves as the fundamental unit for spatial representation:

$$\tau_t = \{\mathbf{c}_t, \mathbf{b}_t, \mathbf{p}_t\} \tag{2}$$

In this formulation, $\mathbf{c}_t$ represents the object category, $\mathbf{b}_t \in \mathbb{R}^2$ denotes the box center in image coordinates, and $\mathbf{p}_t \in SE(3)$ signifies the pose in the camera-centric frame. This design enables τ_t to serve as a universal geometric primitive: in static contexts, it localizes the object for spatial grounding; in temporal sequences, it characterizes the trajectory waypoints for motion estimation. By unifying 2D centroids with 3D poses, τ_t provides a versatile representation that describes both discrete physical entities and continuous motion paths within a unified generative framework.

B. VLM architecture

Our approach adopts PaliGemma [3] as the backbone. PaliGemma employs SigLip [45] as the visual encoder, which extracts rich semantic features from images. However, PaliGemma is trained mainly on the VQA task and detection task. We empirically find that such supervision on text level or 2D location-level may focus on high-level semantics, while losing the fine-grained details for control-oriented features. Thus, we use the pose as the universal representation to bridge the non-robotics data in computer vision task and robotics data in manipulation task as in Fig. 2. As pose is described in the 3D world, it can force the model to understand the spatial relationship in the 3D world. To maintain the inherent 2D visual grounding capabilities, we reuse the text and `<loc>` vocabularies from the original PaliGemma model for representing $\mathbf{c}_t$ and $\mathbf{b}_t$.

To incorporate 3D priors, we leverage auxiliary inputs, including depth maps and camera intrinsics, which are commonly available in 3D-annotated, non-robotic datasets. The inclusion of such auxiliary information has been shown to markedly improve 3D prediction performance at test time.

For 3D-annotated non-robotics data, poses are defined within the camera coordinate frame. To maintain alignment with these diverse geometric datasets, our approach requires the model to predict actions directly within the camera frame of the image view, rather than estimating robot-centric actions defined in the robot's base frame. This design ensures a shared camera-centric representation across heterogeneous data sources, enabling the model to fully leverage the 3D awareness acquired during VLM pre-training for precise robotic control.

C. Unified Pose Representation

We adopt 3D pose as a universal representation to unify the feature spaces of non-robotic 3D datasets and robotic manipulation data. This choice is motivated by the fact that both objects and robotic grippers can be effectively parameterized within this shared geometric space. By imposing this representation, we constrain the model to estimate spatial coordinates directly from visual observations within the camera frame, which naturally fosters the development of robust 3D spatial grounding capabilities. To implement this, we introduce a dedicated **pose token** defined in the camera coordinate system. This design choice explicitly addresses the challenge of coordinate misalignment encountered during pre-training, facilitating a seamless integration of diverse data sources into a consistent framework.

We represent object states and robotic motion trajectories using a parameterization of translation and rotation. In the following sections, we detail the implementation of these pose tokens and demonstrate their versatility across both general 3D spatial grounding and embodied robotic manipulation tasks.

Object-Level Representation. Each object is characterized by its rigid-body transformation, comprising translation and rotation. We discretize these continuous parameters to integrate them into the VLM as an extended vocabulary of specialized tokens. For rotation, we perform uniform quantization of the Euler angles and assign a dedicated `<rot>` token to each principal axis.

Regarding translation, an empirical analysis of our large-scale pre-training statistics reveals a distinct distributional divergence between the lateral (x, y) and longitudinal (z) axes as shown in our Appendix. While the x and y distributions exhibit comparable means and variances, the z-axis (depth) distribution is significantly shifted, reflecting the inherent perspective projection and depth range of the camera. To accommodate this discrepancy and enforce depth-aware geometric grounding, we employ a shared `<trans_xy>` token for the x and y dimensions and a distinct `<trans_z>` token for the depth dimension.

To further evaluate 3D grounding performance in general 3D benchmark [7], we incorporate object scale into the rep-

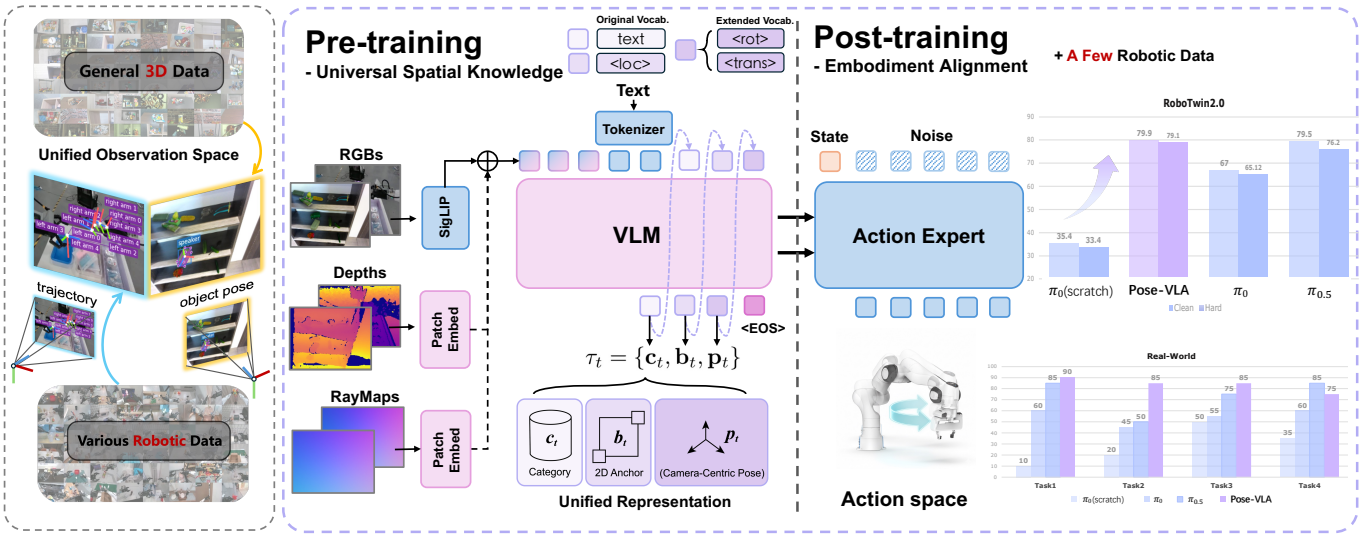

Fig. 2. Pipeline of Pose-VLA. Pose-VLA decouples VLA training into: (1) **Pre-training** for extracting universal 3D spatial priors in a camera-centric space, and (2) **Post-training** for embodiment alignment. The VLM predicts a structured sequence $\mathcal{S} = (\tau_1, \ldots, \tau_T)$ via next-token prediction, where each tuple $\tau_t = \{\mathbf{c}_t, \mathbf{b}_t, \mathbf{p}_t\}$ consists of a category $\mathbf{c}_t$, 2D box center $\mathbf{b}_t$, and camera-centric pose $\mathbf{p}_t$. To enhance spatial reasoning, auxiliary 3D geometry priors are integrated via additive fusion with RGB embeddings, analogous to positional encodings. This unified format enables seamless knowledge transfer from diverse 3D datasets to robotic domains, achieving robust alignment with minimal demonstrations.

resentation. Since the physical dimensions of objects across all three axes exhibit similar statistical characteristics, we represent object scale using a unified `<size>` token for spatial dimensions.

By mapping image observations into structured pose tokens, Pose-VLA establishes a direct correspondence between image pixels and physical entities within a metric-aligned coordinate system. Unlike previous methods [2, 28] that rely on decomposing numerical coordinates into individual digit tokens, our tokenization scheme significantly enhances efficiency by reducing the required token sequence length per coordinate. Crucially, this approach ensures that all task data are optimized within a shared real-world metric space, enabling a unified physical vocabulary across both perception and action. By supervising the model to estimate target object poses, we drive the network to implicitly learn 3D spatial features, leading to a level of fine-grained geometric reasoning that far surpasses the capabilities of conventional VLMs dominated by VQA supervision.

Trajectory-Level Representation. Robot actions can be parameterized by the 6-DoF pose of the gripper. By adopting this convention, the gripper pose can directly reuse the discretized tokens established for object-level representations. For dynamic manipulation, we represent action trajectories as temporally ordered sequences of these unified pose tokens.

Unlike contemporary VLA frameworks [17] that define actions as joint angles or base-frame poses, our approach addresses the fundamental misalignment between the observation space (camera-centric) and the action space (robot-centric). Such spatial discrepancies often hinder cross-embodiment generalization, as the model must implicitly bridge the gap between localized visual observations and a global robot-base coordinate system. To resolve this conflict, we project all action trajectories into the respective camera coordinate frame of each view. This alignment ensures that both non-robotic 3D grounding data and robotic demonstration data share a consistent geometric reference frame.

By supervising the model to estimate future trajectories directly within the image observation space, we provide a significantly denser and more grounded supervision signal compared to mapping pixels to embodiment-specific tokens [17]. Furthermore, this camera-centric formulation enhances implicit 3D correspondence learning across diverse viewpoints, such as head-mounted and wrist-mounted cameras. This dense supervision encourages the model to learn representative, control-oriented features that are sensitive to subtle variations between consecutive observations. Ultimately, representing the gripper and objects within a shared token space unifies the optimization objectives of large-scale computer vision and robotics tasks, facilitating the scalable transfer of 3D geometric knowledge to downstream robotic manipulation.

D. Adding Prior Conditioning

To enhance the framework's geometric awareness, we incorporate auxiliary spatial priors. Inspired by Pow3R [15], these embeddings are integrated via additive fusion with the RGB token embeddings prior to the initial Transformer block, analogous to standard positional encodings.

Camera Ray Encoding. To establish a direct correspondence between pixels and their physical viewing directions, we construct camera rays derived from the intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$. The ray $\mathbf{r}$ for each pixel (u, v) is computed as $\mathbf{r} = \mathbf{K}^{-1}[u, v, 1]^\top$, which encodes the viewing direction relative to the camera's optical center. This ray encoding enables the model to effectively process non-centered crops

and facilitates high-resolution inference by providing absolute geometric anchors. Following the RGB processing pipeline, these dense ray maps are patchified and projected into the latent space before being fused with the visual features.

Depth Prior Integration. For the depth prior, given a depth map $\mathbf{D}$ and its corresponding sparsity mask $\mathbf{M}$, we deliberately avoid value normalization to maintain explicit alignment with the metric space. We stack the depth and mask into a joint representation $[\mathbf{D}, \mathbf{M}] \in \mathbb{R}^{H \times W \times 2}$, which is then patchified and embedded. This configuration allows the encoder to robustly handle varying levels of sparsity through integrated valid-pixel masking.

This multi-modal conditioning ensures the seamless integration of semantic and 3D geometric information, significantly bolstering the model's capacity for precise 3D reasoning in complex manipulation scenarios.

E. Training Strategy

Pre-training stage. In this stage, we employ a next-token prediction loss to instill fundamental spatial priors into the model. To preserve the extensive vision-language capabilities of the base VLM, we jointly supervise the generation of 2D localization tokens and our extended 3D pose tokens. Given the multi-modal input $\mathbf{O}$ and language instruction $\mathbf{L}$, we utilize the standard auto-regressive objective:

$$\mathcal{L}_{pre}(\theta) = -\sum_{k=1}^{K} \log p_\theta(\mathbf{s}_k | \phi(\mathbf{O}), \mathbf{L}, \mathbf{s}_{<k}) \tag{3}$$

where s_k denotes the k-th token in the structured sequence $\mathcal{S}$, and $\phi(\mathbf{O})$ represents the modality masking strategy. Specifically, during training, we randomly zero-pad raymaps or depth values to ensure the model maintains robust inference capabilities even with RGB-only inputs. Furthermore, for depth images, we sample sparse depth points to enhance robustness to diverse sensor noise patterns. This training phase equips the VLM with coarse yet semantically grounded spatial awareness, enabling it to transition from a semantic describer to a geometric reasoner.

Post-training stage. For embodiment alignment, we follow [5] by appending a lightweight action expert to map the pre-trained representations from the VLM backbone into robot-specific commands. This decoupling allows our primary focus to remain on large-scale spatial prior acquisition during pre-training. Specifically, we employ a flow matching objective [22, 25] to denoise the action tokens, where the VLM backbone interacts with the action expert through self-attention layers to provide semantic and geometric conditioning. Notably, the action expert is trained from scratch without inheriting initialization parameters from [5], ensuring that the learned alignment is natively grounded in our Pose-VLA representations.

Remark on Inference: We emphasize that 3D pose prediction serves strictly as an *auxiliary proxy task* during pre-training to learn spatial representations. At inference time, the action expert maps implicit features directly to continuous actions end-to-end, introducing zero overhead for explicit pose or camera trajectory estimation.

F. Pre-training Datasets

Our pre-training utilizes two types of datasets: **(1) Spatial Grounding and 3D Perception.** We curate 1.4M images with 6.5M 3D annotations across three pillars: (1) *General 3D Detection* via Omni3D [7] (Objectron, SUN RGB-D, ARKitScene); (2) *6D Pose Estimation* via Omni6DPose [47] (SOPE, ROPE); and (3) *Manipulation-centric Perception* from the BOP Challenge (YCB-V, LineMOD, HOT3D). This mixture of real and synthetic data ensures robust awareness of 3D geometry and lighting variations. **(2) Robotic Trajectory Pre-training.** This stage aligns spatial features with motion control using nearly 1.55M trajectories. All end-effector poses are transformed into a unified camera-centric frame via calibration parameters provided by datasets. (1) *AgibotWorld Beta (Real):* 1M trajectories from 100 isomorphic robots, covering 200 tasks and 87 atomic skills. (2) *InternData-A1 (Sim):* 550K trajectories across heterogeneous platforms, including ARX, Aloha, Franka. Spanning 4 embodiments and 227 scenes, it provides dense supervision for articulated objects and long-horizon tasks.

Implementation Details. The model is optimized via AdamW with a cosine schedule, reaching a peak learning rate of 5×10^{-5} after a 1,000-step warmup. This is followed by a cosine decay to a minimum of 2.5×10^{-6}. Training is performed on 16 NVIDIA H20 GPUs with bfloat16 precision and a per-GPU batch size of 8. Specifically, both 3D spatial grounding and trajectory estimation tasks are completed within 2 days each on 16 H20 GPUs. This approach significantly reduces GPU hours compared to previous VLA training [4, 5] while effectively instilling robust spatial priors for the VLM.

IV. Experiment

A. Evaluation in 3D Grounding Benchmarks

Baselines. For the 3D grounding task, we compare against recent open-source and closed-source vision-language models capable of end-to-end 3D spatial grounding directly from language prompts. Our open-source baselines include the Qwen3-VL series [33] and the Visual Spatial Tuning (VST) models [42]. For closed-source baselines, we evaluate against Seed1.5-VL [14], Gemini-2.0-Pro [12], and the state-of-the-art Gemini Robotics-ER [37], as their settings align most closely with our experimental protocol.

Evaluation Metrics. Following the protocols established in Qwen3-VL [33] and Seed1.5-VL [14], we adopt Average Precision (AP) as our primary evaluation metric. Each evaluation input consists of an image-text pair, where the textual prompt specifies the target object category. To ensure a fair and consistent comparison with existing VLMs, we set the Intersection over Union (IoU) threshold to 0.15 and report mAP@0.15 on the Omni3D test set [7]. We primarily conduct evaluations on the SUN-RGBD [35] and Objectron [1] subsets of Omni3D [7], which cover a wide range of indoor and tabletop environments. In line with previous benchmarks,

the detection confidence for all predictions is fixed at 1.0 to evaluate the spatial grounding accuracy independently of confidence scoring.

Results and Analysis Table I summarizes the 3D spatial grounding performance across various foundational models. Our proposed Pose-VLA achieves a substantial performance leap on the *Objectron* dataset, reaching an AP_{15} of 87.3. This represents a 16.1% absolute improvement over the strongest open-source baseline, Qwen3-VL [33] (235B, *Thinking*), and a significant margin over closed-source models such as Seed1.5-VL [14] and Gemini-2.0-Pro [37]. On the SUN RGB-D benchmark [35], Pose-VLA delivers a competitive score of 45.5, outperforming all open-source variants and the VST-3B-SFT [42] model, while remaining closely comparable to the state-of-the-art Gemini Robotics-ER [37].

The results on Objectron [1] are particularly significant as they highlight the model's robust capability in handling diverse daily objects essential for robotic manipulation, such as *bottles, cups, and cameras*. While SUN RGB-D evaluates general indoor scene understanding, Objectron demands precise geometric localization of individual objects. The fact that Pose-VLA maintains superior spatial awareness across both large-scale indoor environments and object-centric contexts validates our *pose-centric* pre-training paradigm.

Crucially, Fig. 3 demonstrates that Pose-VLA generalizes robustly across unseen scenarios, ranging from tabletop layouts to complex robotic workspaces. Unlike baselines that frequently suffer from orientation misalignment in novel environments, Pose-VLA provides precise localization and orientation estimation. This superior generalization highlights the effectiveness of utilizing auxiliary spatial priors and diverse 3D datasets, enabling the model to maintain high-fidelity spatial awareness even in out-of-distribution scenarios.

TABLE I
COMPARISON OF 3D SPATIAL GROUNDING PERFORMANCE USING THE AP_{15} METRIC ACROSS DIFFERENT VLM BACKBONES. BASELINE RESULTS ARE SOURCED FROM THE OFFICIAL QWEN3-VL REPORT [33].

Category	Model	SUN RGB-D [35]	Objectron [1]
Open Source	Qwen3-VL [33] (235B, Thinking)	34.9	71.2
	Qwen3-VL [33] (235B, Instruct)	39.4	-
	Qwen3-VL [33] (4B, Instruct)	34.7	-
	Qwen3-VL [33] (2B, Instruct)	33.8	-
	VST-SFT(3B) [42]	37.3	-
	VST-RL(3B) [42]	40.1	-
Closed Source	Seed1.5-VL [14]	33.5	8.1
	Gemini-2.0-Pro [12]	32.5	5.5
	Gemini Robotics-ER [37]	**48.3**	-
	Ours-VLM (3B)	45.5	**87.3**

B. Evaluation in Simulation Benchmarks

To evaluate the cross-embodiment generalization and task-level robustness of Pose-VLA, we conduct extensive experiments on two large-scale simulation benchmarks: RoboTwin-2.0 [9] and LIBERO [23].

RoboTwin-2.0. To evaluate the generalization capabilities of our method, we perform *multi-task training* on the RoboTwin-2.0 benchmark [9]. Specifically, Pose-VLA and all baseline models [5, 6, 3] are trained on a comprehensive dataset comprising 2,500 demonstrations collected in clean scenes (50 per task), augmented with an additional 25,000 demonstrations gathered in heavily randomized scenes (500 per task). To ensure a strictly fair comparison across this dataset, all models are initialized from their respective pre-trained checkpoints and fine-tuned using an identical training budget of 80K optimization steps with a total batch size of 32.

To rigorously assess performance under distribution shifts, we conduct inference across both "Easy" and "Hard" scenarios. While the "Easy" setting serves as a controlled baseline, the "Hard" scenes introduce significant visual and geometric noise to evaluate the model's robustness. We report the average success rate over 100 evaluation trials per task.

LIBERO. Following the protocol of OpenVLA [17], we evaluate our model on four distinct task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. Each suite comprises 10 diverse tasks with 500 demonstrations in total. These benchmarks allow us to evaluate Pose-VLA across multiple dimensions, ranging from spatial reasoning to long-horizon planning. We compare our approach against a broad range of state-of-the-art generalist policies, including OpenVLA [17], SpatialVLA [32], π_0 [5], $\pi_{0.5}$ [6], π_0-FAST [30], CoT-VLA [49], WorldVLA [8], and GR00T-N1 [4].

Results and Analysis. Table II and Table III present the comprehensive evaluation results across simulation benchmarks. To ensure a fair comparison with RGB-only baselines, we evaluate Pose-VLA using only RGB input by masking out depth and raymap modalities. On RoboTwin 2.0, Pose-VLA establishes a new state-of-the-art, achieving an average success rate of 79.1% in the challenging *Hard* setting. This represents a substantial margin of 14.0% over the strong baseline π_0 and a dramatic improvement of over 45% compared to the vanilla PaliGemma baseline, confirming that our spatial pre-training effectively prevents feature collapse where standard VQA-oriented backbones fail. Extending this evaluation to LIBERO, Pose-VLA demonstrates superior cross-task generalization with an overall average of 96.0%, surpassing π_0 and ranking second only to $\pi_{0.5}$. Notably, in the LIBERO-Long suite which demands multi-stage reasoning, our model attains 92.4%, tying with $\pi_{0.5}$ for the top position. Thus, these results validate that integrating 3D spatial priors not only enhances robustness against visual perturbations but also provides the structural consistency required for long-horizon planning.

C. Evaluation in Real-world Tasks.

Real-world Experimental Setup. We evaluate our model using a Dual-arm Xtrainer robotic platform, where each arm is equipped with a 1-DoF parallel gripper. For visual perception, we employ a multi-camera configuration: a RealSense D455 depth camera serves as the head-mounted *Eye-on-Base* sensor for global scene understanding, while a RealSense D405 depth

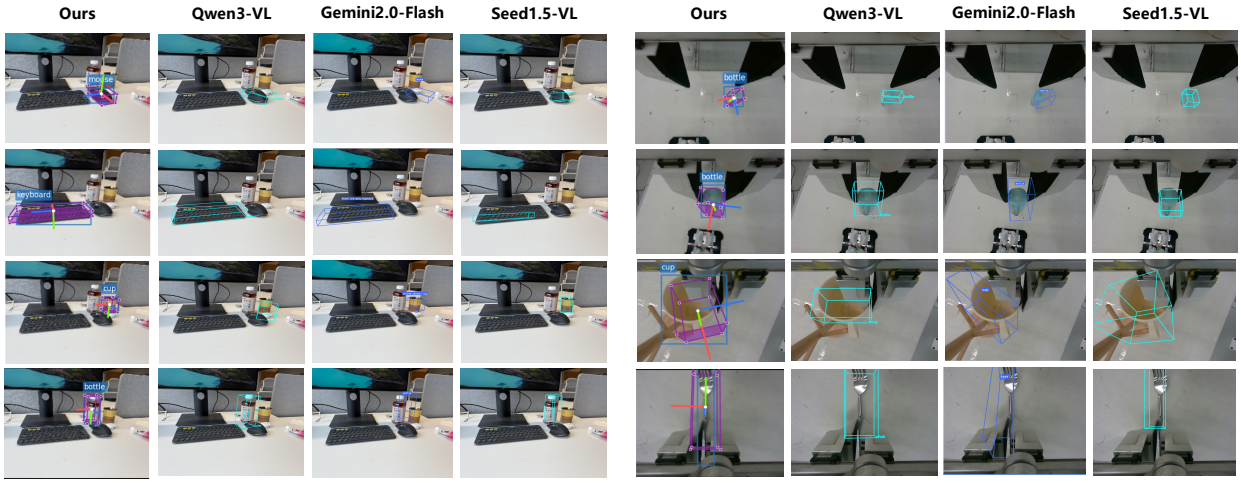

Fig. 3. Generalization of 3D spatial grounding across unseen scenarios. Pose-VLA exhibits robust generalization across various unseen settings, ranging from indoor tabletop layouts to complex robotic manipulation workspaces, providing more precise geometric localization than baseline methods.

TABLE II
COMPARISON ON ROBOTWIN 2.0 SIMULATION. ALL POST-TRAINED FOR 80K STEPS (BATCH SIZE=32).

Manipulation Task	π_0 [5]		$\pi_{0.5}$ [6]		PaliGemma_expert		Pose-VLA(w/o depth)	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Adjust Bottle	70%	57%	62%	69%	67%	47%	97%	77%
Beat Block Hammer	80%	73%	78%	93%	40%	37%	100%	87%
Click Alarmclock	83%	73%	92%	89%	80%	57%	83%	93%
Dump Bin Bigbin	100%	90%	89%	97%	73%	77%	97%	97%
Grab Roller	93%	97%	100%	100%	53%	57%	97%	100%
Handover Block	33%	40%	84%	57%	7%	17%	73%	80%
Lift Pot	63%	73%	100%	85%	37%	30%	100%	97%
Move Pillbottle Pad	73%	67%	85%	61%	10%	20%	90%	87%
...(50 tasks)	...	...	...	...	...	...	...	...
Open Laptop	73%	70%	88%	96%	67%	70%	93%	93%
Pick Dual Bottles	47%	50%	21%	63%	10%	0%	87%	87%
Place A2b Left	60%	43%	84%	82%	27%	10%	90%	80%
Place Cans Plasticbox	83%	73%	90%	84%	23%	17%	97%	87%
Place Container Plate	93%	100%	89%	95%	90%	80%	97%	100%
Place Dual Shoes	63%	50%	93%	75%	3%	0%	87%	77%
Place Object Scale	70%	43%	82%	80%	10%	10%	67%	83%
Place Phone Stand	70%	57%	83%	81%	30%	20%	87%	87%
Place Shoe	87%	80%	96%	93%	23%	30%	100%	87%
Average (%)	67.00	65.12	79.48	76.16	35.40	33.36	**79.91**	**79.10**

TABLE III
SUCCESS RATES (%) ON LIBERO BENCHMARK. BOLD AND UNDERLINE DENOTE THE BEST AND SECOND BEST RESULTS ACROSS FOUR SUITES AND THEIR AVERAGE.

Method	Spatial	Object	Goal	Long	Avg
OpenVLA [17]	84.7	88.4	79.2	53.7	76.5
SpatialVLA [32]	88.2	89.9	78.6	55.5	78.1
CoT-VLA [49]	87.5	91.6	87.6	69.0	83.9
WorldVLA [8]	87.6	96.2	83.4	60.0	79.1
GR00T-N1 [4]	94.4	97.6	93.0	90.6	93.9
π_0 [5]	96.8	98.8	95.8	85.2	94.1
$\pi_{0.5}$ [6]	**98.8**	**98.2**	**98.0**	**92.4**	**96.8**
π_0-FAST [30]	96.4	96.8	88.6	60.2	85.5
Pose-VLA(w/o depth)	<u>96.5</u>	<u>98.0</u>	<u>97.1</u>	**92.4**	<u>96.0</u>

camera is integrated as an *Eye-on-Hand* sensor to provide localized, high-resolution visual feedback.

Tasks and Evaluation. To verify the generalizability of Pose-VLA, we define four representative and challenging manipulation tasks as in Fig. 4: (1) *Stacking*: stacking three nested bowls; (2) *Hanging*: hanging a mug onto a designated peg on a wooden stand; (3) *Long-horizon Interaction*: a complex multi-stage task involving opening a drawer, picking and placing a pen inside, and closing the drawer; (4) *Deformable Object Manipulation*: folding a fabric towel. We collect an average of 100 demonstrations per task to facilitate embodiment alignment. Each model is evaluated over 20 trials per random seed, totaling 60 trials per task with diverse object placements. The success rate across these 60 trials is reported as the primary performance metric.

Results and Analysis. As shown in Fig 5, Pose-VLA demonstrates superior manipulation capability with an average success rate of 81.25%, significantly outperforming the vanilla PaliGemma (29.6%) and the strong baseline $\pi_{0.5}$ (71.65%). Our method consistently outperforms or matches $\pi_{0.5}$ across

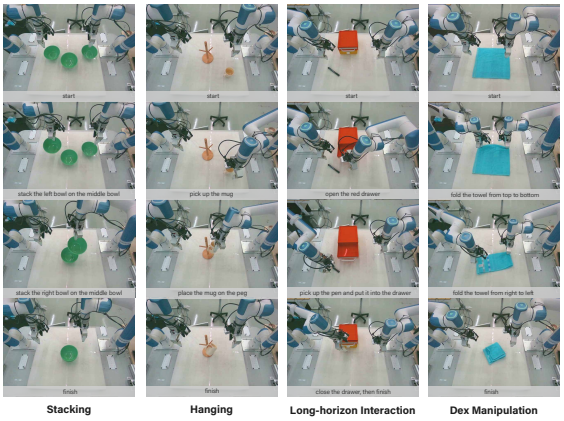

Fig. 4. Real-world setup of four representative tasks. Our platform uses a dual-arm Xtrainer with head and wrist cameras. The benchmark includes: (1) *Tableware Arrangement*, (2) *Hanging* a mug, (3) *Long-horizon* drawer interaction, and (4) *Deformable* towel folding. Success rates are evaluated over 20 trials per task.

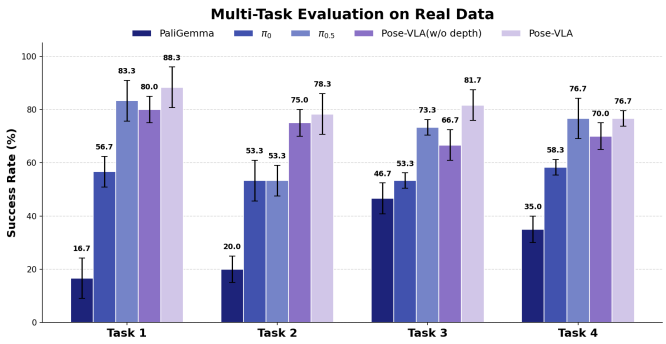

Fig. 5. Success rate comparison of Pose-VLA and baseline models across four real-world manipulation tasks. Each model is evaluated over 60 trials per task, with success rates reported as percentages. Under the same demonstration scale, Pose-VLA consistently outperforms current vision-language-action baselines, especially in long-horizon and deformable object tasks.

all tasks, notably achieving significant improvements in Task 2 (78.3% vs. 53.3%) and Task 3 (81.7% vs. 73.3%). Furthermore, the ablation study highlights the critical role of the depth modality; removing depth information leads to a marked performance decline, particularly in Task 3 where the success rate drops by 15.0 percentage points (from 81.7% to 66.7%). This substantial gap, alongside the 8.3% drop in Task 1 (from 88.3% to 80.0%), confirms that explicit 3D geometry is indispensable for precise spatial alignment and minimizing execution ambiguity.

D. Ablation Study

To verify how our 3D spatial pre-training and key design choices facilitate downstream performance, we conduct a series of ablation studies on both the 3D perception and RoboTwin 2.0 benchmarks. We systematically evaluate the impact of input modalities, pre-training objectives, architectural backbones, and action representations.

Ablation of Geometric Modalities. We first investigate the contribution of each geometric modality on 3D grounding tasks, as shown in Table IV. On the SUN RGB-D benchmark, which contains complex indoor layouts, the integration of depth information provides a pivotal boost, increasing the AP_{15} from 39.0 (*RGB Only*) to 45.1 (*RGB+Depth*). In contrast, the performance gain on the Objectron dataset is less pronounced. We attribute this to the nature of the Objectron data, which primarily provides sparse and localized depth points centered around specific objects, offering limited additional geometric context compared to the dense depth maps in SUN RGB-D. Notably, the combination of Raymap and Depth (*Full Model*) yields the best overall performance, confirming the synergy between explicit ray-based priors and metric depth.

TABLE IV
ABLATION STUDY OF GEOMETRIC MODALITIES ON 3D GROUNDING.

Configuration	SUN RGB-D [35]	Objectron [1]
RGB Only	39.0	86.4
RGB + Raymap	38.5	86.7
RGB + Depth	45.1	85.4
Full Model	**45.5**	**87.3**

To thoroughly evaluate our design choices for robotic tasks, we perform two sets of ablation studies on the RoboTwin 2.0 benchmark. First, in Table V, we present an *incremental analysis* to trace how each component, ranging from data scaling to action space optimization, contributes to the final performance of Pose-VLA. Second, in Table VI, we conduct an *orthogonal comparison* to validate specific architectural choices, such as the VLM backbone and tokenization strategy, against prominent alternatives. Crucially, to eliminate confounding variables and ensure a strictly fair comparison, every evaluated variant is fine-tuned under a unified regime of 80K optimization steps with a batch size of 32, starting directly from its original pre-trained checkpoint.

TABLE V
INCREMENTAL ABLATION STUDY ON THE ROBOTWIN 2.0 BENCHMARK. GREEN AND RED DENOTE RELATIVE PERFORMANCE CHANGES FROM THE PRECEDING ROW.

Type	**Improvements**	**Success Rate (%)**	
		Easy	**Hard**
Baseline	PaliGemma + Action-Expert	35.4	33.4
Data	+ 3D Data	70.2 (+34.6)	69.1 (+35.7)
	+ 3D Data & Robotic Data	72.9 (+2.7)	72.0 (+2.9)
Action	Post-train: Joint → EE pose	**79.9** (+7.0)	**79.1** (+7.1)
	Pre-train: Cam → Base	80.4 (+0.5)	76.7 (-2.4)

Ablation of Pose Pre-training Data. We evaluate our data scaling strategy and mixture compositions across both Table V and Table VI. As shown in the incremental analysis (Table V), incorporating diverse non-robotic 3D grounding data to the vanilla *PaliGemma* baseline yields a substantial +35.7% boost in success rate. This confirms that pre-training on 3D spatial tasks establishes a more robust geometric foundation than conventional VQA-dominated pre-training. To further isolate the impact of these data sources, Table VI compares orthogonal data mixtures. Crucially, pre-training on robot data only severely degrades results, proving that generic 3D vision

data is the key driver of our framework's spatial capabilities. Conversely, relying solely on non-robotic 3D data (*3D data only*) maintains highly competitive performance. Together, these results demonstrate that a strong grasp of 3D geometry constitutes the most significant contribution to success in our setting, effectively enhancing data efficiency by alleviating the reliance on costly, large-scale robotic demonstrations.

Ablation of Action Representation. As shown in Table V, shifting the action space from joint angles to end-effector (EE) pose yields an average performance gain of over $+7.0\%$. This improvement stems from the fact that our VLM is pre-trained to reason over 3D spatial poses; thus, adopting an EE pose interface allows the model to more directly leverage its learned geometric features. In contrast, joint-space control introduces a transformation overhead that degrades knowledge transfer. Conversely, we observe that predicting actions in the base frame leads to a performance drop (-2.4%) in *RoboTwin-Hard* scenarios. This indicates that camera-centric representations provide more robustness when facing significant visual variations in downstream policy transfer.

TABLE VI
ORTHOGONAL ABLATION STUDY ON THE ROBOTWIN 2.0 BENCHMARK. SR DENOTES SUCCESS RATE.

Dimension	Variant	Easy SR (%)	Hard SR (%)
Backbone	Qwen3PI	57.7	55.3
Tokenization	Pose-VLA (OpenVLA binning)	68.5	65.9
Data Mixture	Pose-VLA (Robotic data only)	71.3	68.3
	Pose-VLA (3D data only)	77.2	76.2
Full Model	**Pose-VLA**	**79.9**	**79.1**

Ablation of VLM Backbone. To verify whether the performance gain of Pose-VLA stems from our specific pre-training rather than simply the scale of the VLM backbone, we evaluate a baseline using Qwen3-VL [33] (denoted as *Qwen3PI*). Although Qwen3-VL possesses state-of-the-art 3D grounding capabilities, Table VI shows that Qwen3PI, which undergoes only post-training, trails Pose-VLA by more than 20%, despite outperforming the scratch baseline. This gap confirms that even powerful VLMs lack the robust 3D representations necessary for fine-grained manipulation without explicit pose-centric pre-training. Furthermore, our approach demonstrates superior data efficiency under *tighter demonstration budgets*. Remarkably, this advantage persists even when the pre-training stage utilizes *solely non-robotic 3D data*, validating the effectiveness of pose tokens as a "universal interface" for transferring geometric knowledge to robotic policies.

Ablation of Tokenization. We evaluate the structural advantages of our pose token representation by comparing it against standard methods, as reported in Table VI. Specifically, we contrast our approach with the uniform binning of delta actions used in OpenVLA [17]. OpenVLA's tokenization is fundamentally limited in two ways: **1)** It relies on *local delta* representations tied to specific robot kinematics, which inherently blocks the absorption of non-robotic data. In contrast, our *global spatial* representation decouples actions from specific embodiments, serving as a "universal interface" that seamlessly integrates both robotic and non-robotic 3D trajectories. **2)** OpenVLA employs *uniform binning*, which inefficiently utilizes token capacity. Our *non-uniform* approach, detailed in the Appendix, allocates higher resolution at close proximities, yielding significantly higher precision for delicate, near-contact manipulation. Consequently, reverting to OpenVLA's standard tokenization results in strictly inferior performance, validating that our unique design is essential for both data scaling and precise 3D manipulation.

V. LIMITATIONS AND FUTURE WORK

While *Pose-VLA* provides a robust spatial foundation for robotic manipulation, its current scope presents promising avenues for future exploration. First, due to computational constraints, the full potential of large-scale 3D data scaling and high-efficiency 3D annotation for robotic videos remains partially unexplored. Furthermore, while our framework inherently supports deformable object manipulation during downstream adaptation, such as towel folding, the pre-training stage is currently driven by rigid-object datasets. Explicitly formulating deformable states, such as 3D keypoints or dense grasp annotations, into the pose token vocabulary during the large-scale pre-training phase represents a critical next step. Alongside this, extending these spatial priors from parallel-jaw grippers to multi-fingered dexterous hands will further unlock complex manipulation capabilities. Additionally, scaling our real-world experiments to a broader diversity of behaviors in unstructured environments remains a key objective. In particular, transitioning from static tabletop setups to mobile and egocentric manipulation presents an exciting frontier for future work. Finally, as our core focus in this work is maximizing spatial priors, we observe a natural trade-off regarding the backbone's original semantic reasoning when facing complex semantic distractors. Exploring semantic-preserving training paradigms to maintain general reasoning capabilities without compromising 3D awareness offers a compelling direction for future research.

VI. CONCLUSION

We present Pose-VLA, a framework utilizing a unified pose token as a universal interface to bridge spatial perception and robotic control. By integrating RGB-D data with camera intrinsics, our architecture instills an intrinsic 3D awareness that enables the learning of robust spatial and motion priors through large-scale pre-training. Pose-VLA achieves state-of-the-art results in 3D grounding and provides a generalizable physical foundation for downstream robotic tasks, as validated across LIBERO, RoboTwin, and real-world experiments. Our work demonstrates a viable avenue for scaling VLA models by leveraging heterogeneous non-robotic 3D data alongside specialized robotic demonstrations. We advocate for a shift from VQA-based foundations toward pre-training embodied-aware VLMs, fostering the development of backbones that are inherently grounded in the physical world.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021.

[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, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: 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 Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023.

[8] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[9] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[10] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025.

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

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[13] Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*, 2025.

[14] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[15] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1071–1081, 2025.

[16] Nikita Kachaev, Mikhail Kolosov, Daniil Zelezetsky, Alexey K Kovalev, and Aleksandr I Panov. Don't blind your vla: Aligning visual representations for ood generalization. *arXiv preprint arXiv:2510.25616*, 2025.

[17] 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.

[18] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

[19] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025.

[20] Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Liuao Pei, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025.

[21] Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2505.11917*, 2025.

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maxim-

ilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[23] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] 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.

[25] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[26] 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.

[27] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025.

[28] Yunze Man, Shihao Wang, Guowen Zhang, Johan Bjorck, Zhiqi Li, Liang-Yan Gui, Jim Fan, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Locateanything3d: Vision-language 3d detection with chain-of-sight. *arXiv preprint arXiv:2511.20648*, 2025.

[29] Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. Spatiallm: Training large language models for structured indoor modeling. *arXiv preprint arXiv:2506.07491*, 2025.

[30] 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.

[31] Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, et al. Eo-1: Interleaved vision-text-action pretraining for general robot control. *arXiv preprint arXiv:2508.21112*, 2025.

[32] 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.

[33] Qwen Team. Qwen3-vl: A frontier multimodal large language model. https://github.com/QwenLM/Qwen3-VL, 2025. Accessed: 2026-01-22.

[34] Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025.

[35] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[36] Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, and Soujanya Poria. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14199–14214, 2025.

[37] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[38] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[39] Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. *arXiv preprint arXiv:2507.01016*, 2025.

[40] Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025.

[41] Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024.

[42] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

[43] Shuai Yang, Hao Li, Yilun Chen, Bin Wang, Yang Tian, Tai Wang, Hanqing Wang, Feng Zhao, Yiyi Liao, and Jiangmiao Pang. Instructvla: Vision-language-action instruction tuning from understanding to manipulation. *arXiv preprint arXiv:2507.17520*, 2025.

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

[45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[46] Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action mod-

els. *arXiv preprint arXiv:2601.03309*, 2026.

[47] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024.

[48] Zhilong Zhang, Xiong-Hui Chen, Yidi Wang, Yihao Sun, Wenyu Luo, Haoxiang Ren, Haoxin Lin, and Yang Yu. On the representation degradation in vision-language-action models.

[49] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025.

[50] Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5377–5395, 2025.

[51] 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.