2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

AugVLA-3D: Depth-Driven Feature Augmentation for Vision-Language-Action Models

Zhifeng Rao[1,2*], Wenlong Chen[1*], Lei Xie[1], Xia Hua[3], Dongfu Yin[1], Zhen Tian[1†], F. Richard Yu[4]

***Abstract*— Vision-Language-Action (VLA) models have recently achieved remarkable progress in robotic perception and control, yet most existing approaches primarily rely on VLM trained using 2D images, which limits their spatial understanding and action grounding in complex 3D environments. To address this limitation, we propose a novel framework that integrates depth estimation into VLA models to enrich 3D feature representations. Specifically, we employ a depth estimation baseline called VGGT to extract geometry-aware 3D cues from standard RGB inputs, enabling efficient utilization of existing large-scale 2D datasets while implicitly recovering 3D structural information. To further enhance the reliability of these depth-derived features, we introduce a new module called action assistant, which constrains the learned 3D representations with action priors and ensures their consistency with downstream control tasks. By fusing the enhanced 3D features with conventional 2D visual tokens, our approach significantly improves the generalization ability and robustness of VLA models. Experimental results demonstrate that the proposed method not only strengthens perception in geometrically ambiguous scenarios but also leads to superior action prediction accuracy. This work highlights the potential of depth-driven data augmentation and auxiliary expert supervision for bridging the gap between 2D observations and 3D-aware decision-making in robotic systems.**

I. Introduction

A central ambition in robotics is to develop agents with broad versatility: the capability to execute diverse tasks across varying environments while following natural language instructions, adapting to situational constraints, and remaining resilient to unexpected disturbances. Yet, attaining such generality is still highly challenging due to the inherent complexity of real-world perception–action loops and the wide variability of manipulation scenarios. Recent advances in imitation learning [1]–[4] together with the rise of Vision-Language-Action (VLA) models [5]–[10] provide a promising route toward building robots with adaptable and transferable skills. By combining large-scale vision-language models (VLMs) with action-generation modules, VLAs have created new opportunities for scalable and generalizable policy learning.

*These authors contributed equally to this work.
† Corresponding author: Zhen Tian, tianzhen@gml.ac.cn.
This work was supported by the Guangdong Natural Science Foundation (2025A1515012083) and the National Natural Science Foundation of China (62271324, 62231020).
[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China.
[2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China.
[3]School of Communication and Information Engineering, Shanghai University, Shanghai, China.
[4]School of Information Technology, Carleton University, Ottawa, Canada.

Recent years have witnessed remarkable progress in this line of research. OpenVLA [13] demonstrates that scaling vision-language pretraining to billions of parameters enables robots to generalize across a wide range of manipulation tasks. Building on this foundation, $\pi 0$ [14] leverages pretrained VLMs to align visual observations with textual task descriptions, thereby enabling flexible policy learning from human demonstrations. Its successor, $\pi 0.5$ [15], further refines multimodal alignment strategies and incorporates broader datasets, leading to improved task generalization and grounding. Despite these advances, a key limitation remains: these models fundamentally rely on 2D representations inherited from VLMs. While 2D features capture rich semantic context, they lack explicit 3D structural awareness that is indispensable for reasoning about geometry, depth, and spatial interactions in the physical world. Consequently, such models often falter in scenarios requiring fine-grained spatial reasoning, such as collision avoidance, object stacking, or reachability analysis.

To overcome the shortcomings of purely 2D approaches, recent works have attempted to introduce explicit 3D information. 3D-VLA [5] proposes a generative world model that integrates 3D perception, reasoning, and action, offering a new framework for embodied intelligence. However, this approach requires large-scale 3D embodied datasets that are expensive to collect and thus difficult to scale. SpatialVLA [16] injects Ego3D position encoding and adaptive action grids to improve spatial reasoning and trajectory generalization, but its dependence on large quantities of real-world robot trajectories makes it resource-intensive and less adaptable to new domains. PointVLA [12], on the other hand, augments pretrained VLAs with point cloud features through lightweight injection blocks, effectively improving geometric awareness. Yet, it relies on specialized 3D sensors such as LiDAR, which restricts training to datasets with 3D annotations and prevents large-scale pretraining with the abundant 2D corpora that have driven the success of VLAs. Collectively, these methods highlight the importance of 3D awareness but also reveal significant barriers: reliance on expensive data collection, limited scalability, or difficulty in fully leveraging existing 2D datasets.

In this paper, we address these limitations by introducing AugVLA-3D, a novel framework that enhances VLA models with sensor-free 3D structural features while preserving their compatibility with large-scale 2D training data. Specifically, our approach leverages a state-of-the-art depth estimation model (VGGT) to transform 2D RGB inputs into dense 3D point clouds, which are subsequently encoded by a PointNet-

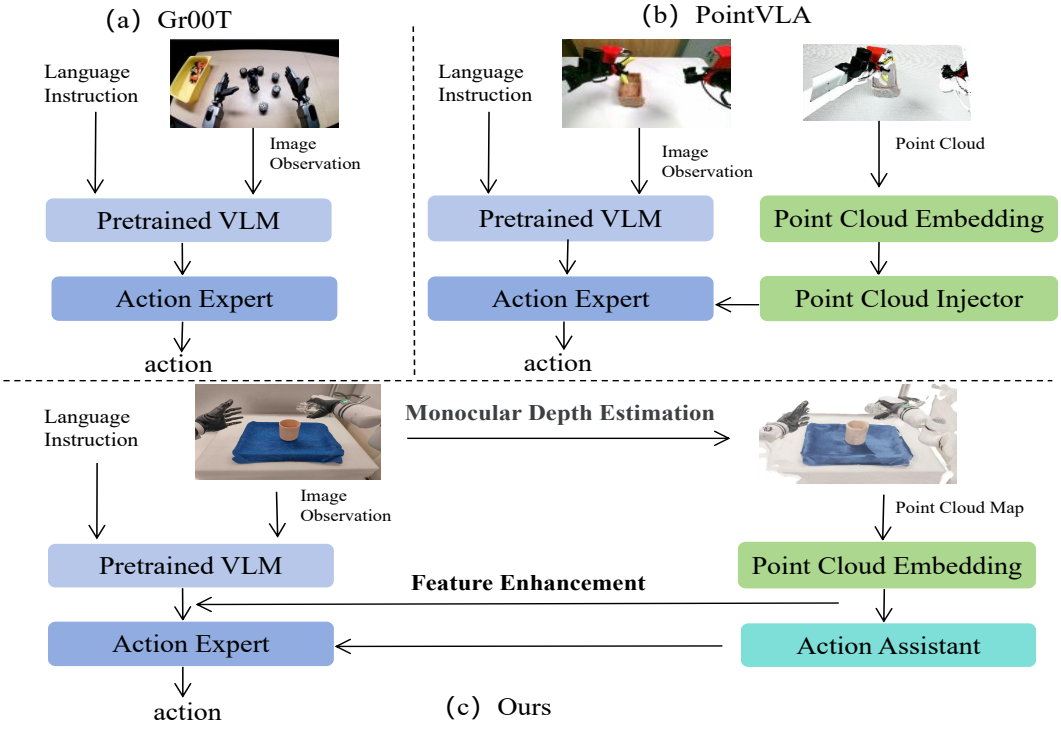


Fig. 1: The architecture comparison with different methods. (a) Gr00t [11]: Only 2D visual features are used without explicit 3D reasoning. (b) PointVLA [12]: LiDAR-based point clouds are introduced but rely on specialized 3D sensors. (c) AugVLA-3D: Our AugVLA-3D leverages depth estimation to inject 3D structural features in a sensor-free manner, enabling scalable training and stronger 3D generalization.

based extractor to yield compact geometric descriptors. This design bypasses the need for specialized 3D hardware and enables maximal reuse of existing 2D VLA datasets at scale. To ensure that the extracted 3D features are task-aligned and do not interfere with pretrained representations, we introduce an Action Assistant module that mirrors the main action head but with fewer parameters. Acting as a lightweight regularizer, it constrains the 3D features through structural alignment and layer-wise feature injection, thereby improving stability and downstream performance. By combining sensor-free 3D feature extraction with auxiliary regularization, AugVLA-3D achieves tight integration of semantic and geometric reasoning, leading to substantially improved generalization in real-world 3D environments. The comparison of our model with other models is shown in Fig. 1. Our paper presents several notable contributions, which are as follows:

1) We propose a sensor-free 3D feature extraction method that leverages a depth estimation model (VGGT) to convert 2D RGB images into point clouds, from which compact 3D features are derived to enhance the original VLA model.
2) We design a new module called Action Assistant to regularize the extracted 3D features, which constrains the learned 3D representations with action priors and ensures their consistency with downstream control tasks.

II. Related Works

2D Vision-Language-Action (VLA) Models. In recent years, researchers have increasingly explored training general-purpose robotic control policies from large-scale robot learning datasets [17]–[21]. Vision-Language-Action (VLA) models [6]–[8], [22]–[24], extending visual-language models (VLMs), have emerged as a promising approach by leveraging internet-scale data and fine-tuning on robot manipulation datasets, demonstrating strong generalization and adaptability. OpenVLA [13], trained on 4,000 hours of open-source data, established an early baseline for generalist VLAs. More advanced methods such as $\pi 0$ [14] and $\pi 0.5$ [15] scale up to 10,000 hours of internal data to further improve grounding and generalization, but remain computationally heavy and reliant on large-scale pre-training. To improve efficiency, TinyVLA [25] eliminates pre-training while achieving faster inference and comparable performance, and SmolVLA [26] accelerates inference by skipping selected VLM layers. Despite these advances, current approaches are fundamentally built on 2D VLMs, lacking explicit 3D structural reasoning and thus struggling to ensure robust generalization in real-world environments. This limitation has motivated a shift toward integrating richer spatial awareness into VLA frameworks.

3D-Enhanced VLA Models. Developing resilient visuomotor policies in 3D settings [27]–[29] has emerged as a core challenge in robotic learning. Recent studies increas-

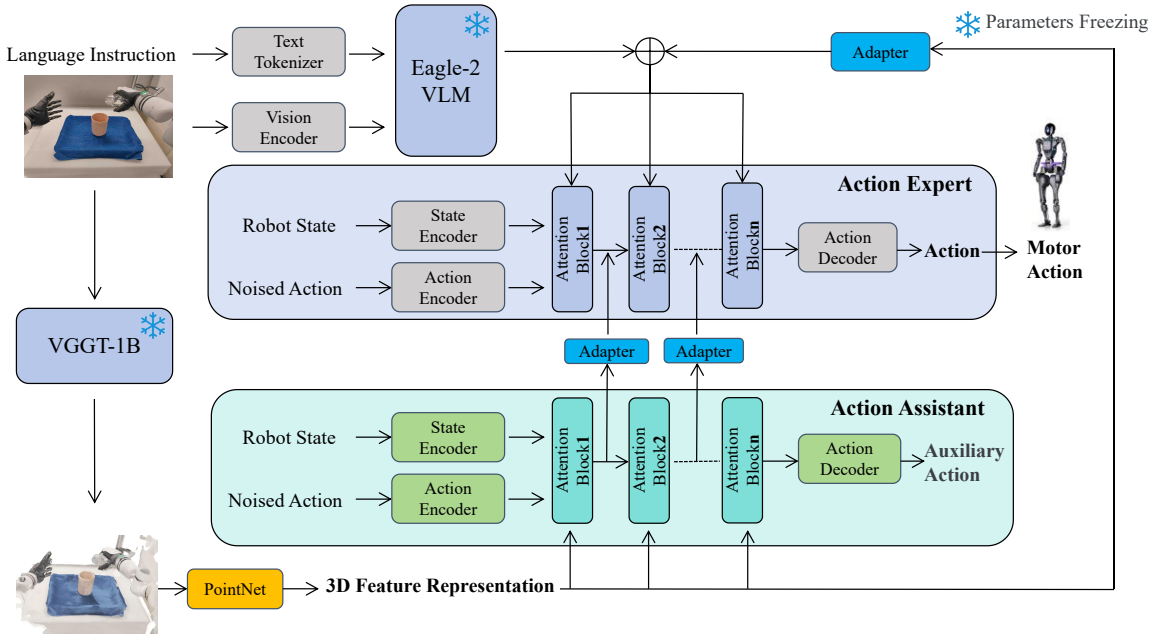


Fig. 2: Architecture of our proposed AugVLA-3D framework. The overall model design largely follows the GR00t backbone, while we introduce a dedicated 3D feature injection module to enhance the Action Expert with depth-derived geometric information. To ensure that the injected 3D features are aligned with task objectives without introducing excessive computational overhead, we further design an Action Assistant. This module is structurally consistent with the Action Expert but adopts a lightweight parameterization, effectively constraining the 3D features while keeping the additional cost minimal.

ingly strive to integrate perception, reasoning, and action into unified 3D-aware models. A representative example is 3DVLA [5], which builds a holistic vision-language-action framework capable of tackling generalization, visual question answering, scene interpretation, and robot control in a single system. Building on this trend, iDP3 [30] reinforces the 3D vision backbone and demonstrates strong adaptability on humanoid robots across both egocentric and third-person viewpoints. SpatialVLA [16] further incorporates Ego3D positional encodings and adaptive action grids to promote spatial reasoning, though its dependence on large collections of real-world trajectories limits scalability. PointVLA [12] employs lightweight adapters for multi-modal feature fusion [31] to enhance geometric awareness. However, its reliance on specialized 3D sensors restricts scalability and prevents utilizing abundant 2D corpora. Overall, while 3D-enhanced VLAs move closer to bridging the gap between perception and action in realistic environments, their reliance on specialized sensors and heavy data collection hinders broader applicability.

VLA Models for Dexterous Hands. Dexterous manipulation [11], [32]–[34] is a key challenge in robotics, requiring precise control and generalization across objects and tasks. GR00T-N1 [11] is a Vision-Language-Action foundation model for humanoid robots that jointly interprets instructions and generates real-time motor actions, trained on a mix of real-robot, human video, and synthetic data. Being-H0 [35] treats human hands as the ultimate "foundation manipulators," learning hand motions from large-scale human videos and transferring these skills to robots. DexVLG [36] constructs a massive dataset of 170 million grasp poses mapped to 174,000 simulated objects with detailed part-level descriptions, enabling instruction-aligned dexterous grasping. These works collectively demonstrate the potential of data-driven and foundation-model approaches for advancing general-purpose dexterous manipulation. More broadly, they also highlight a recurring theme across the VLA literature: the need to balance scalability, efficiency, and embodiment-specific adaptability, which continues to drive the design of next-generation robotic foundation models.

III. Methods

A key limitation of existing Vision-Language-Action (VLA) models lies in their reliance on 2D visual features, which lack explicit 3D structural reasoning and thus hinder generalization in real-world environments. To address this, we propose a novel 3D Feature Injection framework that augments VLA models with depth-based geometric cues while maintaining compatibility with large-scale 2D training corpora. Specifically, we leverage a state-of-the-art depth estimation model (VGGT) to transform standard 2D RGB observations into dense 3D point clouds, enabling sensor-free acquisition of structural information. A PointNet encoder is then employed to extract compact 3D features, which serve as geometric priors for action prediction. To further align these 3D features with task objectives, we introduce a new module called Action Assistant that mirrors the Action Expert head but with fewer parameters. Acting as a lightweight

regularizer, this module constrains the learned 3D features and outputs intermediate activations that are injected, together with PointNet features, into the corresponding layers of the original VLA backbone. This design achieves a tight integration of 2D semantics and 3D geometry, reinforcing multimodal fusion while avoiding dependence on 3D sensors. The architecture of our model is shown in Fig. 2.

A. 3D Feature Extraction based on Depth estimation

Our framework is motivated by a critical limitation of current vision-language-action (VLA) models: their reliance on 2D features inherited from large-scale vision-language models (VLMs). While such features provide strong semantic grounding, they lack explicit geometric reasoning, which is essential for accurate interaction in 3D environments. As a result, current VLA models often struggle in tasks requiring precise depth understanding, collision avoidance, or spatial planning. For instance, ambiguous 2D cues can lead to failures when objects overlap in the image plane or when perspective distortions alter apparent distances, highlighting the need for explicit structural priors.

Recent attempts have begun addressing this gap by incorporating 3D information into VLA architectures. Among them, PointVLA augments pretrained VLAs with LiDAR-derived point clouds through lightweight injection blocks. Although this approach improves geometric awareness, it introduces substantial practical limitations. Specifically, reliance on LiDAR sensors restricts applicability to domains where high-quality 3D data is available, and the resulting datasets are typically limited in scale. Consequently, such methods can only be applied during fine-tuning, preventing the full reuse of the massive 2D datasets that have driven recent progress in VLMs and VLAs.

To overcome these shortcomings, we propose a sensor-free approach to infuse 3D reasoning into VLA models. Our framework leverages VGGT, a state-of-the-art monocular depth estimator, to predict dense depth maps from standard RGB images. The estimated depth maps are back-projected into camera-centered point clouds using known intrinsics, followed by light preprocessing steps such as outlier filtering and normalization. These point clouds are then encoded by a PointNet backbone, producing compact descriptors that capture both local geometric structures and global spatial layouts. By converting ubiquitous 2D data into 3D representations, our method enables large-scale reuse of existing VLA training corpora while providing explicit structural cues absent in purely 2D models. The extracted 3D feature f_{3D} is formulated as follows:

$$\begin{aligned} P &= f\big((I_i)_{i=1}^{N}\big), & I_i &\in \mathbb{R}^{3\times H\times W} \\ \tilde{P} &= \mathcal{S}(P), & \tilde{P} &\in \mathbb{R}^{M'\times 3},\ M' < H\times W \\ f_{3D} &= \text{PointNet}(\tilde{P}), & F_{3D} &\in \mathbb{R}^{M'\times C} \end{aligned} \tag{1}$$

where $f(\cdot)$ denotes the VGGT depth estimation model, $(I_i)_{i=1}^{N}$ represents a sequence of N RGB images and N represents the number of observation perspectives of the robot; if there is only one observation perspective, N equals 1. P is the raw point cloud map obtained from VGGT, $\mathcal{S}(\cdot)$ is a point cloud sampling operator used to reduce the number of points to M' and C is the dimension encoded by PointNet.

Our framework achieves a tight integration of 2D semantics and 3D structure without requiring dedicated 3D sensors. By transforming RGB data into actionable geometric priors, the model gains robust spatial reasoning while retaining the scalability of 2D training pipelines. Compared to LiDAR-dependent approaches such as PointVLA, our method is more broadly applicable: it leverages monocular depth estimation to scale 3D feature injection to the full extent of existing 2D corpora. This sensor-free design maximizes scalability, reduces deployment costs, and substantially improves generalization in real-world 3D environments.

B. Action Assistant

While depth-derived features introduce valuable geometric priors, injecting them directly into a pretrained VLA backbone often leads to unstable optimization and performance degradation, as raw geometric signals are not naturally aligned with the task-specific action representation space. To mitigate this, we design a new module called the Action Assistant that mirrors the Action Expert head but with a lightweight parameterization. This auxiliary expert acts as a task-aligned regularizer: it leverages PointNet-extracted 3D features to generate motion constraints, and both the original PointNet features and the intermediate activations from each layer of the motion assistant are injected into the corresponding layers of the primary motion expert. This design constrains the learning of 3D features, ensuring that they align with manipulation objectives and enhance the pretrained 2D pathway rather than disrupting it.

The action assistant mirrors the structure of the primary action head but uses significantly fewer parameters. Its role is twofold: (1) to act as a *task-guided projector* that transforms PointNet-extracted 3D features into action-relevant embeddings, and (2) to serve as an *intermediate regularizer* that injects stable guidance into the VLA backbone at multiple depths. Concretely, for the l-th layer, we compute

$$\tilde{h}^{(l)} = h_{\text{orig}}^{(l)} + \alpha^{(l)} \cdot \mathcal{T}\big(h_{\text{aux}}^{(l)}, f_{\text{3D}}\big), \tag{2}$$

where $h_{\text{orig}}^{(l)}$ is the original hidden state of the VLA, $h_{\text{aux}}^{(l)}$ is the corresponding activation from the auxiliary expert, f_{3D} denotes the PointNet features, and $\alpha^{(l)}$ is a learnable scalar gate. The transformation $\mathcal{T}(\cdot)$ is implemented as a lightweight projection or cross-attention module, ensuring that injected features are smoothly aligned with the pretrained representation.

In order to minimize the additional parameter and computational cost, the action assistant adopts a compact *transformer-diffusion* architecture. Specifically, we reduce the hidden dimensions, share weights across denoising steps, and limit the diffusion horizon to a few iterations. This structure not only maintains efficiency but also leverages the generative prior of diffusion modeling to provide stronger regularization of 3D-to-action mappings.

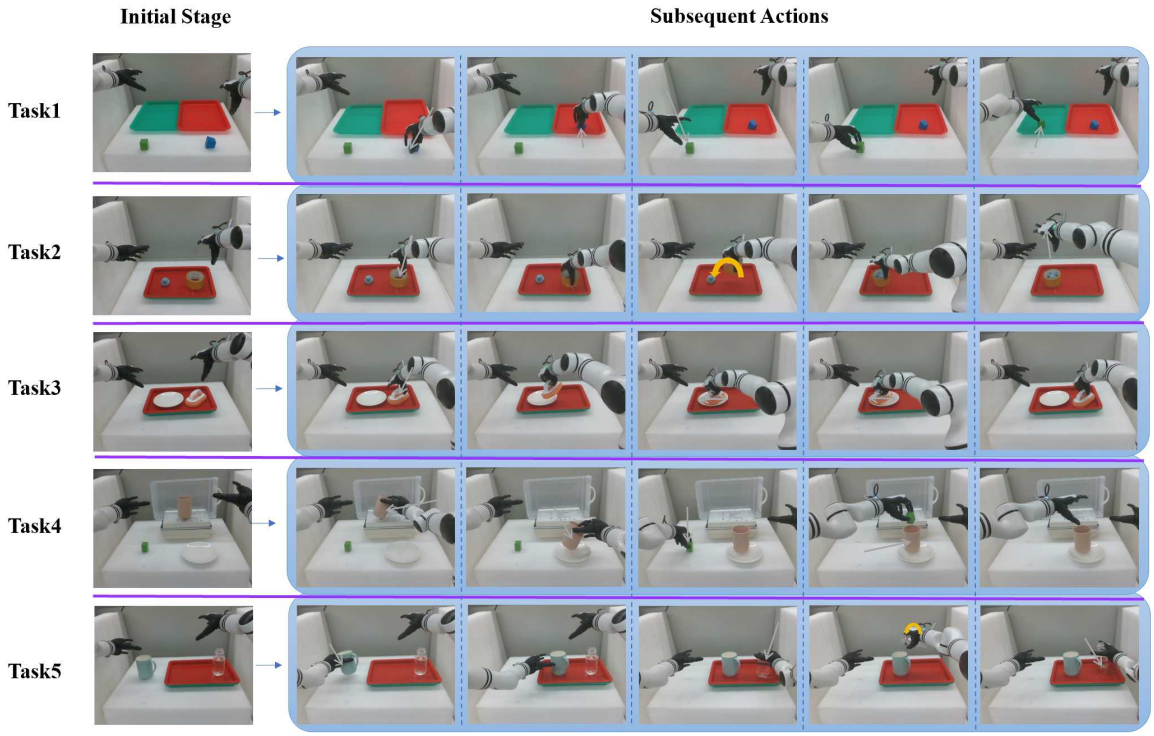


Fig. 3: Illustrations of the five experimental tasks: Task 1: Place the wooden blocks into the corresponding plates; Task 2: Cover the duck toy with a tape; Task 3: Wipe the plates with a dishcloth; Task 4: Take out the cup and put in the block; Task 5: Place the cup and pour water into it.

Critically, the auxiliary actions generated by the action assistant are used solely for computing auxiliary losses to constrain 3D feature learning, and these actions are never directly executed or used to update the motor commands of the robot. This design ensures that the primary action expert retains full control over policy execution while still benefiting from the regularization provided by the assistant. The separation preserves the stability of the pretrained VLA backbone, as the outputs from the Assistant influence feature representations without bypassing the decision-making process of the main policy.

C. Intropy Analysis of AugVLA-3D

Under the Intropy framework [37], [38], AugVLA-3D increases intelligence gain $\mathrm{d}L = \delta S/R$ by enriching geometric discrepancy signals while reducing spatial ambiguity. Depth-derived 3D features inject dense, task-relevant information into the action pathway, increasing δS beyond 2D vision-language alignment alone. Simultaneously, geometry-aware feature routing and action-assisted regularization reduce effective resistance R by constraining optimization with physically meaningful structure. The result is higher Intropy, yielding more robust and generalizable 3D manipulation.

IV. Experiments

This section evaluates the effectiveness of our approach through a series of experiments, including real-world physical trials across diverse scenarios and multi-task validation in the RoboCasa simulation environment, following the experimental protocol of Gr00T [11].

A. Implementation Details

Our model is built upon the GR00t backbone, with the proposed 3D feature injection module integrated into its action prediction pipeline. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Due to limited computational resources, we train our framework using only the `PhysicalAI-Robotics-GR00T-X-Embodiment-Sim` dataset, and specifically restrict training to 10% of its data for only 1 epoch. While this constraint prevents us from scaling training to larger and more diverse datasets, or from conducting extensive evaluations across multiple benchmarks, our results nevertheless highlight the effectiveness of the proposed design.

B. Experimental Results on Real-Life Scenarios

We evaluate our approach on the dexterous robotic hand ROH-A001 across five distinct tasks: Place the wooden blocks into the corresponding plates, Cover the duck toy with a tape, Wipe the plates with a dishcloth, Take out the cup and put in the block and Place the cup and pour water into it, as illustrated in Fig. 3.

TABLE I: Simulation results comparing our method (AugVLA-3D) with Diffusion Policy and Gr00T across various manipulation tasks. AugVLA-3D, enhanced with 3D features, outperforms Gr00T in most tasks, demonstrating the effectiveness of our approach.

Methods	Diffusion Policy		Gr00T		AugVLA-3D	
	30 demos	100 demos	30 demos	100 demos	30 demos	100 demos
Cutting Board to Pot	23%	37%	59%	58%	57%	60%
Cutting Board to Basket	20%	42%	43%	62%	46%	65%
Cutting Board to Tiered Basket	14%	14%	14%	24%	18%	28%
Cutting Board to Pan	28%	48%	68%	66%	70%	71%
Cutting Board to Cardboard Box	12%	16%	31%	30%	36%	33%
Placemat to Bowl	15%	19%	31%	39%	34%	39%
Placemat to Plate	16%	24%	33%	37%	38%	41%
Placemat to Basket	16%	24%	50%	46%	49%	52%
Placemat to Tiered Shelf	7%	6%	12%	22%	18%	25%
Plate to Pan	14%	18%	35%	48%	38%	47%
Plate to Cardboard Box	13%	14%	34%	38%	39%	39%
Plate to Bowl	16%	19%	41%	42%	40%	43%
Plate to Plate	26%	39%	73%	85%	72%	87%
Tray to Tiered Shelf	2%	7%	18%	28%	22%	31%
Tray to Tiered Basket	13%	34%	33%	49%	36%	52%
Tray to Plate	27%	41%	54%	69%	60%	67%
Tray to Cardboard Box	22%	37%	51%	56%	56%	60%
Tray to Pot	22%	48%	52%	60%	55%	65%
Wine to Cabinet	43%	56%	58%	54%	60%	62%
Place Bottle to Cabinet	40%	63%	61%	81%	60%	83%
Place Milk to Microwave	37%	41%	42%	59%	45%	63%
Potato to Microwave	18%	30%	30%	27%	35%	35%
Cup to Drawer	25%	32%	36%	44%	34%	46%
Can to Drawer	48%	75%	78%	77%	75%	81%
Average	21%	33%	43%	50%	46%	54%

ROH-A001. The robotic hand is equipped with 11 joints and 6 active degrees of freedom, enabling fine-grained motion control and grasping capabilities comparable to those of a human hand.

Place the wooden blocks into the corresponding plates. The robot picks up green and red blocks and places them into their corresponding color-marked areas. The challenge lies in identifying the right locations and executing precise placement in a 3D space, which requires dynamic coordination between visual recognition and hand positioning.

Cover the duck toy with a tape. In this task, the robot grasps a roll of tape and places it over a toy duck, securing it in place. The main difficulty lies in positioning both the tape and the duck in 3D space. To align the tape correctly and ensure it covers the duck efficiently, 3D spatial awareness is essential for precise manipulation and placement.

Wipe the plates with a dishcloth. The robot grabs a dishcloth and uses it to clean a plate. The challenge here is to account for the dishcloth's deformation as it interacts with the plate's surface. Using 3D features helps the robot adjust its actions based on the cloth's shape and movement, leading to a more accurate and effective cleaning process.

Take out the cup and put in the block. In this task, the robot retrieves a cup from the cabinet and places a blue block inside it. The challenge is handling the precise placement of the block inside the cup, considering the cup's 3D orientation and position, and managing any potential block collisions or tipping.

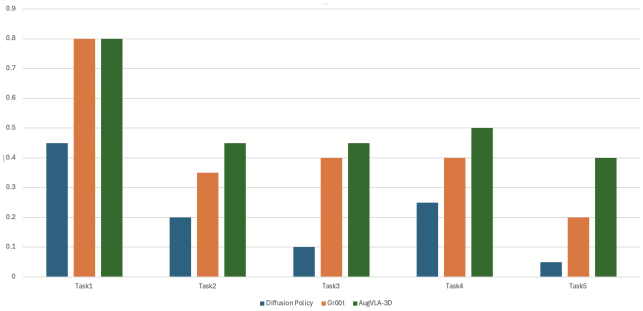
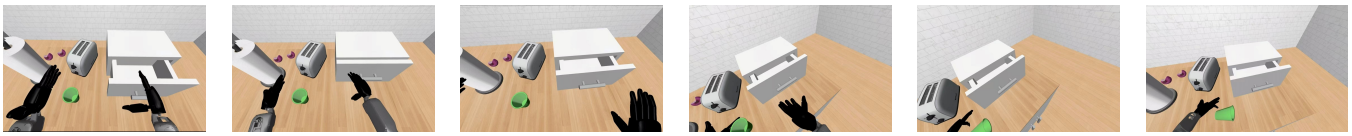


Fig. 4: Experimental results on real-life scenarios

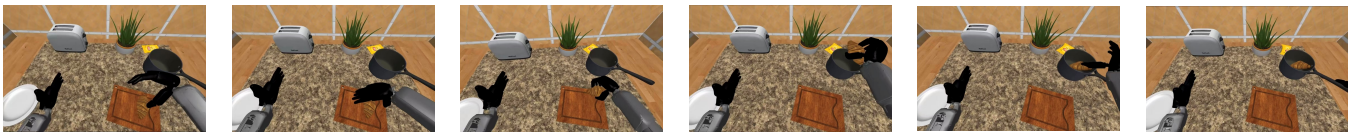
Place the cup and pour water into it. The robot places a cup and pours water into it. The task requires careful control of the pouring angle and speed, while maintaining the correct position of the cup. 3D feedback helps the robot make real-time adjustments to avoid spillage and achieve a controlled pour.

Experimental Results. We evaluate our model against two state-of-the-art dexterous hand baselines, Diffusion Policy [39] and Gr00T [11]. As shown in Fig. 4, AugVLA-3D consistently outperforms both methods across multiple real-world scenarios. This performance gain stems from the explicit incorporation of 3D features, which equip the policy with a stronger awareness of geometric structure and spatial

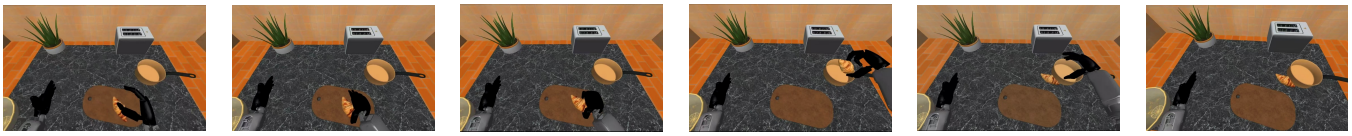
AugVLA-3D: pick up the cup and place it into the open drawer, then close it.



Gr00T: pick up the cup and place it into the drawer, then close it



AugVLA-3D: pick up the bread and put it in the pot



Gr00T: pick up the bread and put it in the pot

Fig. 5: Comparative experimental results between the AugVLA-3D and Gr00T models in complex manipulation scenarios with dexterous hands. The experiment involved two typical manipulation tasks: "pick up the cup and put it in a drawer, then close it" (rows 1-2) and "pick up a loaf of bread and put it in a pot" (rows 3-4). Each task consisted of six key action steps. To ensure fairness, the object layout, lighting conditions, and task instructions were kept consistent across all experimental scenarios. The results show that the AugVLA-3D model, which incorporates 3D spatial features, generally outperforms the Gr00T model in object positioning accuracy, motion trajectory smoothness, and task completion efficiency, validating the effectiveness of 3D features in improving robot manipulation intelligence.

relations capabilities that are often underrepresented in conventional 2D-based VLAs. As a result, our model achieves more reliable grasping, placement, and coordination even under distractors or shifting object poses. Importantly, these advantages emerge despite training on highly constrained computational resources, suggesting that the integration of depth-driven reasoning provides an intrinsic robustness that cannot be easily compensated for by scale alone. We thus believe that with sufficient pretraining, the performance gap over existing baselines would widen further.

C. Experimental Results on Simulation Benchmark

We further benchmark our approach on `robocasa-gr1-tabletop-tasks`, which is built on the RoboCasa simulation framework and includes 24 official tabletop tasks designed for NVIDIA's GR-1 humanoid foundation models. These tasks comprehensively assess the capacity of dexterous hands to operate across diverse manipulation contexts.

Experimental Results. Following the GR00T setup, we train on 30 and 100 demonstrations per task, without any large-scale pretraining due to resource limitations. As summarized in Table 1, AugVLA-3D achieves consistently higher success rates than both Diffusion Policy and Gr00T. This highlights the effectiveness of introducing 3D structural cues into VLA models, as they directly enhance spatial reasoning and object interaction fidelity: two bottlenecks for 2D-based approaches.

Qualitative results are shown in Fig. 5. In the first task, which requires coordinated bimanual manipulation ("pick up the cup and place it into a drawer, then close it"), AugVLA-3D demonstrates superior performance, leveraging 3D feature injection to maintain consistent spatial alignment between the two hands and the manipulated objects. By contrast, Gr00T exhibits frequent miscoordination, reflecting its limited ability to model inter-hand dependencies in 3D space. In the second task ("pick up the bread and put it in the pot"), our model precisely estimates the relative geometry between the bread and the container, enabling successful placement. Gr00T, relying only on 2D features, often misjudges depth and distance, leading to failure. These findings underscore a broader insight: the introduction of explicit 3D features not only improves local grasp accuracy but also enhances global spatial reasoning, which is essential for complex, multi-object manipulations.

V. Conclusion

In this paper, we presented AugVLA-3D, a novel Vision-Language-Action framework that enhances conventional models with sensor-free 3D geometric features derived from 2D RGB inputs via monocular depth estimation, and further stabilized through an action-guided regularization module. This design substantially improves spatial reasoning, action prediction accuracy, and robustness in complex manipulation scenarios, outperforming state-of-the-art baselines such as Gr00T and Diffusion Policy in both real-world and simulation experiments. Importantly, our model achieves these gains despite being trained on limited resources, highlighting its inherent efficiency and scalability. Looking forward, large-scale pretraining, deployment to broader robotic platforms, and self-supervised depth refinement hold promise for further strengthening geometric fidelity and advancing the generalization of embodied AI systems.

References

[1] F. Lin, Y. He, and F. R. Yu, "PP-TIL: Personalized planning for autonomous driving with instance-based transfer imitation learning," in *Proc IEEE IROS*, (Abu Dhabi, UAE), Oct. 2024.

[2] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," in *Proc. IEEE ICRA*, pp. 5637–5643, 2025.

[3] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," *arXiv preprint arXiv:2501.06994*, 2025.

[4] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu, "Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation," in *Proc IEEE ICRA*, pp. 1105–1112, 2025.

[5] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3D-VLA: A 3D vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.

[6] Q. Zhao, Y. Lu, *et al.*, "Cot-VLA: Visual chain-of-thought reasoning for vision-language-action models," in *Proc. IEEE CVPR*, pp. 1702–1713, 2025.

[7] H.-T. L. Chiang, Z. Xu, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah, *et al.*, "Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs," *arXiv preprint arXiv:2407.07775*, 2024.

[8] Y. Yue, Y. Wang, B. Kang, Y. Han, S. Wang, S. Song, J. Feng, and G. Huang, "Deer-VLA: Dynamic inference of multimodal large language models for efficient robot execution," vol. 37, pp. 56619–56643, 2024.

[9] P. Ding, H. Zhao, W. Zhang, W. Song, M. Zhang, S. Huang, N. Yang, and D. Wang, "Quar-VLA: Vision-language-action model for quadruped robots," in *Proc. ECCV*, pp. 352–367, Springer, 2024.

[10] A. W. Yu and A. Nayak, "The Internet of humanoids: A survey of technologies, applications, and challenges," *IEEE Internet of Things Journal*, 2026. Online early access.

[11] J. Bjorck, F. Castañeda, *et al.*, "Gr00t-N1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[12] C. Li, J. Wen, Y. Peng, Y. Peng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," *IEEE Robotics and Automation Letters*, vol. 11, no. 3, pp. 2506–2513, 2026.

[13] M. J. Kim, K. Pertsch, *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[14] K. Black, N. Brown, D. Driess, *et al.*, "π0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv: 2410.24164*, 2025.

[15] K. Black, N. Brown, *et al.*, "π0. 5: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[16] D. Qu, H. Song, *et al.*, "SpatialVLA: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025.

[17] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "RoboNet: Large-scale multi-robot learning," in *Proc. Machine Learning Research*, 2019.

[18] A. Brohan, N. Brown, *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[19] B. Zitkovich, T. Yu, *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. Conference on Robot Learning*, pp. 2165–2183, PMLR, 2023.

[20] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "Bridgedata V2: A dataset for robot learning at scale," in *Conference on Robot Learning*, pp. 1723–1736, PMLR, 2023.

[21] G. Zhou, V. Dean, *et al.*, "Train offline, test online: A real robot learning benchmark," *arXiv preprint arXiv:2306.00942*, 2023.

[22] G. Lu, W. Guo, C. Zhang, Y. Zhou, H. Jiang, Z. Gao, Y. Tang, and Z. Wang, "VLA-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning," *arXiv preprint arXiv:2505.18719*, 2025.

[23] O. Sautenkov, Y. Yaqoot, *et al.*, "UAV-VLA: Vision-language-action system for large scale aerial mission generation," in *Proc. ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1588–1592, 2025.

[24] C. Fan, X. Jia, *et al.*, "Interleave-VLA: Enhancing robot manipulation with interleaved image-text instructions," *arXiv preprint arXiv:2505.02152*, 2025.

[25] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, *et al.*, "TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, pp. 3988 – 3995, 2025.

[26] M. Shukor, D. Aubakirova, *et al.*, "SmolVLA: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[27] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An embodied generalist agent in 3D world," *arXiv preprint arXiv:2311.12871*, 2023.

[28] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3D object manipulation," in *Proc. Conference on Robot Learning*, pp. 694–710, PMLR, 2023.

[29] S. Chen, R. Garcia, I. Laptev, and C. Schmid, "Sugar: Pre-training 3D visual representations for robotics," in *Proc. IEEE CVPR*, pp. 18049–18060, 2024.

[30] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with improved 3D diffusion policies," *arXiv e-prints*, pp. arXiv–2410, 2024.

[31] H. Xue, H. Zhu, Z. Ran, X. Tang, G. Qi, Z. Zhu, S.-C. Kuok, and H. Leung, "Feature fusion and enhancement for lightweight visible-thermal infrared tracking via multiple adapters," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[32] F. Yang, W. Chen, H. Lin, S. Wu, X. Li, Z. Li, and Y. Wang, "Task-oriented tool manipulation with robotic dexterous hands: A knowledge graph approach from fingers to functionality," *IEEE Trans. Cybernetics*, pp. 395 – 408, 2024.

[33] L. Huang, H. Zhang, Z. Wu, S. Christen, and J. Song, "Fungrasp: functional grasping for diverse dexterous hands," *IEEE Robotics and Automation Letters*, pp. 6175 – 6182, 2025.

[34] Y. Zhang, T. Liang, Z. Chen, Y. Ze, and H. Xu, "Catch it! learning to catch in flight with mobile dexterous hands," in *Proc. IEEE ICRA*, pp. 14385–14391, 2025.

[35] H. Luo, Y. Feng, *et al.*, "Being-H0: vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[36] J. He, D. Li, X. Yu, Z. Qi, W. Zhang, J. Chen, Z. Zhang, Z. Zhang, L. Yi, and H. Wang, "DexVLG: Dexterous vision-language-grasp model at scale," *arXiv preprint arXiv:2507.02747*, 2025.

[37] F. R. Yu, *Intropy: A Framework for Modeling Intelligence*. Amazon Digital Services, 2026. Kindle edition.

[38] Y. Ren, H. Zhang, F. R. Yu, *et al.*, "Industrial internet of things with large language models (llms): An intelligence-based reinforcement learning approach," *IEEE Trans. Mobile Computing*, vol. 24, no. 5, pp. 4136–4152, 2025.

[39] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations," *arXiv preprint arXiv:2403.03954*, 2024.