ISF-VLA: An implicit 3D Vision-Language-Action Model Capitalizing on 3D-aware Vision Foundation Models Through Implicit Spatial Fusion

Sheng Zang[1,2], Songlin Wei[4], Haoran Geng[5], Yikai Tang[5],
Bo An[1†], Senthilnath Jayavelu[2,3†], Pieter Abbeel[5], Jitendra Malik[5]
[1]Nanyang Technological University [2]Institute for Infocomm Research, A*STAR
[3]National University of Singapore [4]University of Southern California [5]University of California, Berkeley
† Corresponding Author

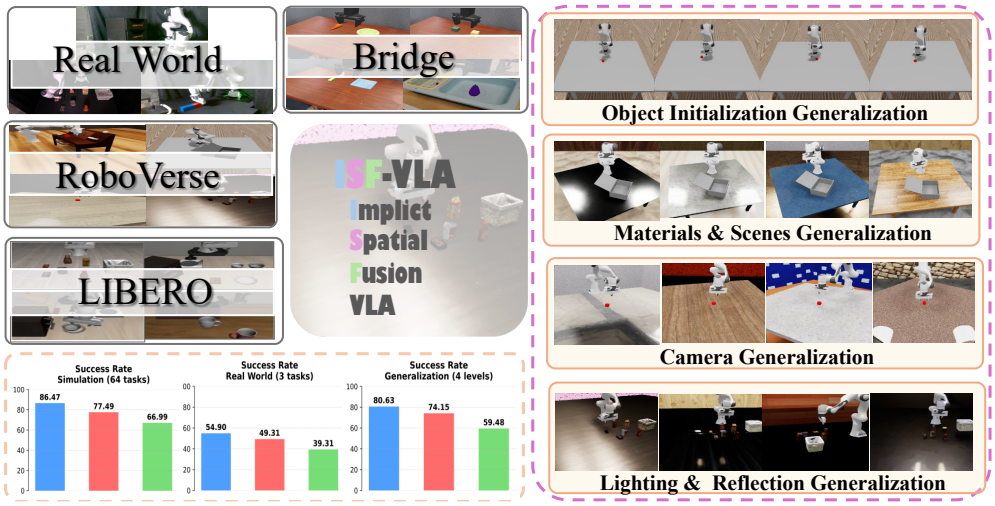


Fig. 1: We present **ISF-VLA**, an implicit 3D vision-language-action (VLA) model leveraging 3D-aware vision foundation models through implicit spatial fusion. Evaluated on 64 simulation tasks where four progressively more challenging levels of randomization were applied, as well as three real-world tasks, **ISF-VLA** ( ) consistently achieves superior zero-shot generalization to unseen configurations and manipulation performance, outperforming the state-of-the-art 2D VLA models ( ) and the explicit 3D VLA model ( ).

***Abstract*— Robotic manipulation requires agents to reason about both 2D visual semantics and 3D spatial geometries, yet existing vision-language-action (VLA) models either rely solely on 2D observations or use explicit 3D representations that are data-inefficient or sensitive to sensor characteristics. To address these challenges, we propose ISF-VLA, an implicit 3D vision-language-action model that capitalizes on 3D-aware vision foundation models through implicit spatial fusion. Our model fuses 2D visual features with implicit 3D information, leveraging pretrained 3D-aware vision encoders with manipulation-aware finetuning and 3D-grounding alignment, thereby endowing the model with 3D priors and facilitating generalizable object manipulation. To promote 2D–3D feature interactions and improve viewpoint generalization, ISF-VLA integrates 2D and 3D tokens via cross-attention, incorporates an additional camera token, and applies visual token selection to reduce the computational overhead of additional 3D tokens. The resulting fused representations exhibit cross-view consistency and interpretable attention patterns. Experiments on real-world robotic tasks and simulation benchmarks demonstrate that ISF-VLA achieves superior generalization and manipulation performance compared to both 2D-only and explicit 3D baselines.**

I. Introduction

Recent progress in vision–language–action (VLA) models [1], [2], [3] has advanced robotic manipulation by combining visual perception, language understanding, and action generation into unified architectures. Nevertheless, many state-of-the-art VLAs [4], [5], [6], [7] remain dominated by 2D visual representations and struggle with spatially demanding manipulation tasks that require reasoning about depth, viewpoint, and object geometry. While explicit 3D pipelines (e.g., point clouds, voxel grids, depth-based methods) provide concrete geometric signals, they are costly, sensitive to noisy sensors, and do not naturally integrate with the visual token semantics expected by large language models. This gap limits robust generalization under camera pose changes in spatially demanding manipulation tasks.

Prior work incorporating explicit 3D information, including SpatialVLA [8], demonstrates one possible direction but also reveals two key limitations. First, explicit 3D inputs require heavy supervision and are sensitive to changes in depth quality or incomplete observations. Second, existing approaches rarely study how to effectively integrate 2D image features with 3D representations in the VLM, where naive fusion can lead to misaligned semantics and degraded decision making. Several concurrent methods [9], [10] ex-

plore implicit spatial signals, but their evaluations do not fully examine 3D spatial reasoning or the design of 2D–3D feature fusion under viewpoint and scene variations.

To overcome these limitations, we propose Implicit Spatial Fusion Vision-Language-Action (ISF-VLA), which introduces implicit 3D reasoning into VLA architectures through a carefully designed fusion of 2D semantics and 3D geometries. ISF-VLA employs Visual geometry grounded transformer (VGGT) [11] as an implicit 3D encoder and fine-tunes it on robotic interaction data to capture manipulation-relevant spatial priors. To align the language model with these 3D-aware visual tokens, we further train the system with a 3D grounding objective that bridges visual geometry and language-driven reasoning.

With these implicit 3D representations, ISF-VLA performs cross-attention–based 2D–3D fusion, where 2D queries attend to 3D keys and values, effectively integrating spatial and semantic cues. Then a dedicated camera token is introduced to model viewpoint variations and improve viewpoint generalization. Finally, we perform cross-view visual feature analysis to study the properties of the fused representations and introduce visual token selection to reduce the computational overhead of additional 3D tokens. Together, these components enable ISF-VLA to achieve robust spatial understanding while remaining plug-and-play compatible with existing VLA frameworks.

We evaluate ISF-VLA on simulation and real-world robotic manipulation benchmarks, including three real-world tasks—LiberoPickButter, FetchBowl, and LiftPegUpright—and additional experiments on RoboVerse [12], LIBERO [13], and Bridge [14]. Across all settings, ISF-VLA consistently outperforms both 2D-only and explicit-3D baselines, demonstrating that implicit spatial fusion improves spatial generalization while preserving semantic alignment. Our main contributions are summarized as follows:

1) **ISF-VLA: Implicit 3D Fusion VLA for Robotic Manipulation.** We integrate an off-the-shelf implicit 3D encoder into a vision-language-action model, enabling spatially grounded visual representations for manipulation without explicit 3D reconstruction.
2) **Manipulation-oriented 3D-grounding Training.** We fine-tune the 3D encoder with depth supervision and align 3D tokens with task-relevant 2D visual tokens via a 3D-grounding objective, encouraging geometry-aware representations for manipulation.
3) **Comprehensive Evaluation and Feature Analysis.** We evaluate ISF-VLA on 67 tasks across simulation and real-world settings, demonstrating strong zero-shot generalization to unseen configurations across four increasingly challenging randomization levels and achieving an average 8.82% improvement over the state-of-the-art baseline, with feature analysis confirming effective 2D–3D fusion and improved cross-view consistency.

All code, training scripts, and the newly rendered RoboVerse dataset will be publicly released to facilitate reproducible research.

II. Related Work

Vision-Language-Action Models. Recent advances in embodied foundation models have led to the emergence of vision-language-action (VLA) models that unify perception, reasoning, and control across multiple embodiments. A growing line of work [1], [2], [15] has explored pretraining VLA models on large-scale cross-embodiment datasets [16], enabling effective fine-tuning for downstream robot tasks. These VLAs extend pretrained vision-language models (VLMs) with action generation, allowing robots to ground visual and linguistic understanding in physical interactions. Representative examples include RT-2 [1], OpenVLA [3], RoboVLM [17], and TraceVLA [18], which employ either discrete or continuous action representations to support both zero-shot and fine-tuned manipulation. More recent systems such as $\pi 0$ [4], $\pi 0.5$ [5], GR00T-N1 [6], and Gemini Robotics 1.5 [7] adopt a dual-system design, where a dedicated action head is trained alongside a general vision-language backbone. Building upon these developments, our work introduces an Implicit 3D VLA that integrates 3D scene understanding into the VLA framework, enhancing spatial perception and generalization in robotic manipulation.

3D Vision-Language-Action Models. Recent efforts have extended VLA models from 2D perception to 3D understanding to advance embodied intelligence. Early works such as LEO [19] and 3D-VLA [20] unify 3D perception, language, and action for generalist agents. More recent 3D VLA models—including SpatialVLA [8], BridgeVLA [21], PointVLA [22] and GeoVLA [23]—explicitly leverage depth, multi-view, or point-cloud representations to boost manipulation performance, but they often suffer from low data efficiency and high sensitivity to sensor characteristics. In contrast, geometry-aware approaches using implicit 3D priors (e.g., VGGT [11]-based encoders) enable 3D spatial perception without explicit reconstruction, providing more robust and efficient representations for downstream manipulation tasks. While several concurrent works integrate 3D encoders into VLAs, such as Evo-0 [9], they do not optimize architectures for manipulation or analyze how 2D and 3D features interact, and their evaluations provide limited analysis of viewpoint and environment generalization. Similarly, [10] aligns intermediate VLA embeddings with pretrained 3D foundation models, which can sacrifice semantic knowledge learned during OXE [16] pretraining. In contrast, ISF-VLA adopts a plug-and-play design that preserves pretrained semantics while leveraging implicit 3D priors, improving both data efficiency and spatial generalization.

III. Methodology

We present ISF-VLA, a vision-language-action model for robotic manipulation that integrates implicit 3D representations with 2D visual features. The model aims to improve spatial perception and cross-view generalization for manipulation tasks. We next describe the key components and design

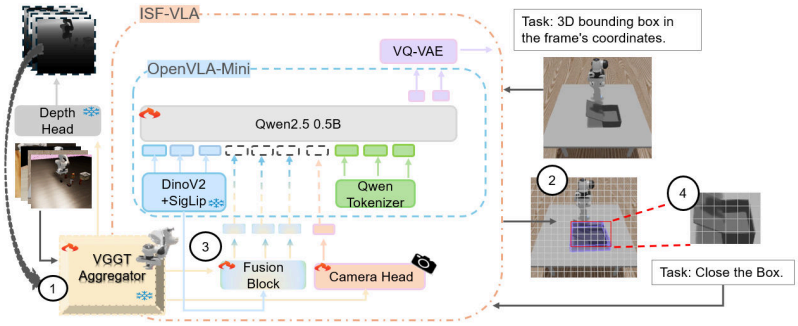


Fig. 2: **Overview of ISF-VLA.** (1) Manipulation-aware 3D encoder fine-tuning with depth supervision from the RoboVerse dataset to capture task-relevant spatial priors. (2) 3D-grounding alignment of the VLM with VGGT-generated 3D and camera tokens using OVMono3D-labeled bounding boxes. (3) 2D–3D fusion via cross-attention with camera token integration for improved viewpoint generalization. (4) Visual token selection based on projected 3D boxes to reduce computational overhead. Downstream VLA training and evaluation rely solely on RGB observations and language instructions.

choices of ISF-VLA. A schematic overview of the model is illustrated in Fig. 2.

A. Manipulation-Aware 3D Encoder and VLM Alignment

ISF-VLA builds upon a 3D encoder that provides spatial representations for manipulation and a VLM capable of interpreting the resulting visual tokens. This section describes the encoder selection, manipulation-aware fine-tuning, and 3D-grounding alignment. Unlike explicit 3D pipelines that rely on depth or point clouds during policy learning, ISF-VLA uses depth supervision only to adapt the implicit 3D encoder, while downstream training and inference operate purely on RGB observations and language. This design reduces sensitivity to depth noise or 3D grounding errors while preserving geometry-aware representations for manipulation.

Off-the-shelf 3D Encoder Selection. To endow ISF-VLA with spatial reasoning capability, we adopt the implicit 3D encoder VGGT [11] as the vision backbone. Our framework is not architecturally tied to a specific 3D encoder; VGGT is chosen primarily for its flexibility in handling single- or multi-view inputs. In contrast, other implicit 3D encoders, such as DUSt3R [24] and CUT3R [25], require paired images or multiple consecutive frames, which may misalign with the problem setting we target. Empirical comparisons with different 3D encoders are reported in Table VI.

Manipulation-Aware 3D Encoder Fine-tuning. To adapt VGGT for manipulation perception, we fine-tune it using depth images from RoboVerse dataset. This encourages the encoder to capture task-relevant spatial priors. Depth supervision provides geometric cues that help the encoder represent object distances and spatial relationships. As a result, the fine-tuned encoder produces 3D tokens with robust spatial information for downstream vision-language-action fusion.

We adopt the depth loss $\mathcal{L}_{\text{depth}}$ from VGGT [11], which penalizes the discrepancy between predicted depth $\hat{D}_i$ and ground-truth D_i using the predicted uncertainty map $\hat{\Sigma}_i^D$, and also includes a gradient-based term: $\sum_{i=1}^{N} \|\Sigma_i^D \odot (\hat{D}_i - D_i)\| + \|\Sigma_i^D \odot (\nabla\hat{D}_i - \nabla D_i)\| - \alpha \log \Sigma_i^D$, where $\odot$ denotes channel-broadcast element-wise product.

3D-Grounding Alignment of the Base VLM. Our base VLM follows the base model of OpenVLA-mini [3], which replaces the original LLM with Qwen2.5-0.5B [26] and uses a VQ-VAE as the action tokenizer. The VLM is initially trained following the LLaVA 1.5 data mixture [27], which includes approximately 1M image-text and text-only samples from open-source datasets. However, during this initial VLM training, the VGGT encoder is not involved, meaning that the LLM has never seen VGGT-generated visual tokens.

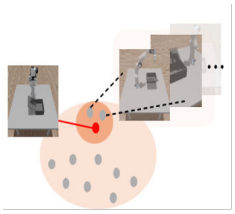
To better align the LLM with VGGT outputs, we introduce a 3D grounding finetuning task. Specifically, we use OVMono3D [28] to annotate the RoboVerse dataset, where each 3D bounding box is represented as a 9-dimensional vector consisting of position $t_i = [x_i, y_i, z_i]^\top$, dimensions $d_i = [w_i, h_i, l_i]^\top$, and orientation r_i . Similar to actions, the 3D bounding box vectors are normalized and predicted autoregressively. This stage teaches the model to interpret VGGT visual tokens with spatial context, aligning implicit 3D representations with the VLM and enabling downstream 2D–3D fusion for action generation.

B. 2D–3D Fusion and Camera Token Integration

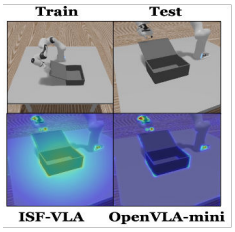
To integrate the 3D-aware representations from VGGT with the 2D visual features of the base VLM, we adopt a cross-attention fusion mechanism. One motivation is the mismatch in image resolution: VGGT is trained on 518×518 images, while the base VLM of VLA uses 224×224. Directly resizing VGGT inputs degrades its 3D reconstruction quality and leads to a mismatch in the number of visual tokens produced by VGGT and the DinoV2 [29] and SigLIP [30] encoders. Cross-attention naturally accommodates this discrepancy by allowing 2D tokens X_{2D} to attend to 3D tokens X_{3D}:

$$X'_{3D} = \text{softmax}\left(\frac{(X_{2D}W_Q)(X_{3D}W_K)^\top}{\sqrt{d_k}}\right)(X_{3D}W_V), \tag{1}$$

where W_Q, W_K, and W_V are learnable projection matrices and d_k denotes the dimensionality of the key vectors.



(a) KNN overlap ratio.



(b) Attention heatmaps.

Fig. 3: **Cross-view visual feature analysis.** Left: KNN overlap ratio (cross-view consistency). Right: attention heatmaps under viewpoint shifts.

The resulting fused tokens X'_{3D} are concatenated with the original X_{2D} tokens for downstream VLA processing.

We also evaluated alternative fusion strategies, including linear concatenation and element-wise addition after normalization, but both underperform compared to cross-attention (Table VI). Cross-attention naturally handles token number mismatches while enabling effective interaction between 2D and 3D features.

To improve viewpoint generalization under camera randomization, we further introduce a camera token C derived from the VGGT camera head. The token is appended to the input sequence, encouraging the model to be aware of camera pose under viewpoint randomization. Ablation results show that finetuning the camera head yields better performance than deriving the token directly from the aggregator (VGGT encoders) (Table VI). The final VLA input sequence is (X_{2D}, X'_{3D}, C).

C. Fused Visual Feature Analysis and Visual Token Selection

Although ISF-VLA demonstrates strong manipulation performance through the previous two components, the effectiveness of 2D–3D feature fusion remains less understood. To investigate this, we analyze fused features across viewpoints and observe improved cross-view consistency and interpretable attention patterns after fusion, which further motivates visual token selection to reduce the overhead of additional 3D tokens.

Cross-view Visual Feature Analysis. We evaluate fused visual features using the K-nearest neighbors (KNN) overlap ratio, which measures how often embeddings from different viewpoints of the same scene appear among each other's nearest neighbors in the latent space (Fig. 3a). A higher overlap indicates stronger cross-view consistency. After 2D–3D fusion, the KNN overlap ratio increases substantially ($0.435 \rightarrow 0.718$), suggesting that implicit 3D information improves viewpoint-invariant visual grounding. Additionally, attention heatmaps (Fig. 3b) show that ISF-VLA maintains task-relevant attention on manipulated objects under viewpoint shifts, whereas 2D-only baselines exhibit noticeable attention drift.

Visual Token Filtering. To reduce the computational overhead introduced by additional 3D tokens, we retain visual tokens corresponding to manipulated objects by projecting 3D bounding boxes onto the 2D image plane (see Fig. 2). As shown in Table I, this filtering keeps training time close to the 2D baseline while avoiding the large overhead observed without filtering, and the ablation results in Table VI show that performance remains stable.

TABLE I: **Training time comparison on RoboVerse Close-Box.** (50 epochs, 400 demonstrations, 8×A100 GPUs)

Model	Training Time
OpenVLA-mini	12h 36m
ISF-VLA (w/o token selection)	1d 1h 30m
ISF-VLA (w/ token selection)	14h 42m

TABLE II: **Randomization factors at different generalization levels.**

Level	Object Initialization	Camera	Materials & Scenes	Lighting & Reflection
Level 0	✓	✗	✗	✗
Level 0.5	✓	✓	✗	✗
Level 1	✓	✗	✓	✗
Level 2	✓	✓	✓	✗
Level 3	✓	✓	✓	✓

IV. Experiments

We evaluate **ISF-VLA** on both simulation and real-world robotic manipulation tasks along three aspects: (i) **Effectiveness:** whether ISF-VLA's implicit 3D representations improve success rates over strong 2D and 3D baselines; (ii) **Generalization:** performance under unseen object positions, camera viewpoints, and environmental variations; and (iii) **Ablation:** the contribution of key components, including implicit 3D features, camera tokens, the fusion module, and other supporting elements.

A. Simulation Experiments

1) Experimental Setup: We evaluate ISF-VLA in simulation using three established robotic manipulation benchmarks: LIBERO, Bridge, and RoboVerse.

- **LIBERO** [13]: LIBERO provides language-conditioned, procedurally generated manipulation tasks organized into four suites for transfer evaluation. Each suite contains 10 tasks with 50 demonstrations, covering spatial reasoning (LIBERO-Spatial), object recognition (LIBERO-Object), task execution (LIBERO-Goal), and long-horizon generalization (LIBERO-Long). We follow the OpenVLA-OFT [31] LoRA fine-tuning setup (including PD&AC and Cont-L1 losses). As LIBERO does not clearly reflect the benefits of implicit 3D modeling, our primary benchmark and ablations focus on RoboVerse, where generalization can be evaluated more systematically.
- **Bridge** [14]: This benchmark consists of constrained pick-and-place and kitchen-style manipulation tasks drawn from BridgeData V2, a large-scale real-world dataset containing thousands of human demonstrations for evaluating cross-environment and cross-object generalization. The dataset was collected using the WidowX

robot, and we evaluate its performance in simulation using the widely adopted SimplerEnv [32].

- **RoboVerse** [12]: RoboVerse is a unified simulation platform that integrates tasks from multiple benchmarks and provides a standardized multi-level generalization protocol. In particular, Level-1 (Table II) corresponds to scene-level generalization, where the task remains unchanged but the environment (materials and backgrounds) is unseen during training. We evaluate four representative tasks—*MoveSliderLeft* (CALVIN [33]), *CloseBox* (RLBench [34]), *PickCube* (ManiSkill [35]), and *PickButter* (LIBERO [13])—across four generalization levels: (i) task-space, (ii) environment, (iii) camera, and (iv) lighting/reflection. To highlight the effect of camera tokens in ISF-VLA, we introduce Level 0.5, a camera-only generalization based on Level 0.

Baselines. We compare ISF-VLA to a diverse set of baselines, including classical planners including DP [36] and ACT [37], 2D-only VLA models including OpenVLA-mini [3] (trained from scratch), OpenVLA [3] (fine-tuned), OpenVLA-OFT [31], Octo [2], and $\pi 0$ [4] / $\pi 0.5$ [5], 3D-aware VLA models including SpatialVLA [8], and other recent strong methods including TraceVLA [18], RoboVLM [17].

2) Benchmark Results:

- **LIBERO**: Table III compares ISF-VLA with strong baselines. ISF-VLA achieves the highest success on most suites, including LIBERO-Spatial (99.0%), LIBERO-Object (99.0%), and LIBERO-Long (95.8%), outperforming $\pi 0$ and OpenVLA-OFT. It slightly trails OpenVLA-OFT on LIBERO-Goal (97.4% vs. 97.9%). Overall, its implicit 3D representations capture scene geometry and relations, supporting robust generalization across tasks.
- **Bridge**: Table IV shows ISF-VLA achieves 45.3% average success, surpassing SpatialVLA (42.7%) and 2D-only models like OpenVLA. Leveraging implicit 3D representations, it generalizes well to unseen object positions and complex pick-and-place scenarios, highlighting its strength in real-world-like manipulation tasks.
- **RoboVerse**: As shown in Table V, ISF-VLA maintains 80–90% success at higher generalization levels (Level 2–3) on tasks such as *PickButter* and *PickCube*, while SpatialVLA drops to 40–50%. Its implicit 3D features preserve spatial consistency under unseen viewpoints, outperforming both explicit 3D and 2D-only baselines, demonstrating robust multi-level generalization for complex manipulation.

3) Ablation Studies: In this section, we conduct ablation studies to investigate the components of ISF-VLA on the *CloseBox* task in RoboVerse across multiple generalization levels (see Table VI).

- **Manipulation Fine-tuning**: We first evaluate the effect of manipulation-aware fine-tuning of the VGGT aggregator and 3D-grounding alignment, and find that removing either component consistently reduces success rates across all levels, highlighting their crucial role.
- **Fusion Module**: Next, we study the fusion mechanism, where removing cross-attention or using only the fused 2D–3D tokens without the original DinoSigLiP tokens reduces performance by 10–20% at higher levels, showing that effective multi-modal integration is crucial. This also helps explain the performance gap with **Evo-0** [9], which mainly introduces cross-attention fusion for VLA, while ISF-VLA incorporates other key components for improved spatial perception, particularly under more challenging generalization such as camera viewpoint changes.
- **Camera Tokens**: We also demonstrate that camera tokens play a significant role, as removing them or using tokens directly obtained from the aggregator without fine-tuning the camera head reduces success rates by 15–30% at Levels 0.5, 2, and 3, indicating that explicit camera-aware features are crucial for viewpoint generalization.
- **Visual Token Selection**: Visual token selection maintains stable performance across levels without noticeable degradation, while significantly reducing training time.
- **Data Scaling and Multi-view Input**: We examine the impact of data scaling and multi-view input. Doubling demonstration numbers for high-level tasks improves success rates by 5–10%, highlighting the importance of sufficient coverage for rare scenarios, while multi-view inputs provide limited gains, suggesting that ISF-VLA's implicit 3D representation already captures spatial structure from single-view observations.
- **Depth Noise**: To further examine robustness to depth supervision noise, we introduce Gaussian perturbations to depth signals during the manipulation-aware fine-tuning stage. Results show only minor performance degradation, confirming that ISF-VLA relies primarily on implicit 3D priors rather than explicit depth signals.
- **Different 3D encoder Choices**: We further study the sensitivity of ISF-VLA to the implicit 3D encoder. Replacing VGGT with alternatives (DUSt3R [24] or CUT3R [25]) leads to moderate performance drops, indicating that VGGT provides the most suitable spatial representation for our setting.

We also report ablation results on LIBERO Spatial and Object in Table VII. Due to the relatively high baseline performance, the impact of individual components appears smaller, but removing key modules still leads to consistent degradation, confirming their contribution to spatial reasoning.

4) Online Reinforcement Learning Finetuning: We further evaluate ISF-VLA as a policy prior in reinforcement learning (Fig. 4) by integrating it into the SimpleVLA-RL framework [39] to improve out-of-distribution generalization. Leveraging the pre-trained ISF-VLA representation, the agent achieves substantial gains after only 35 RL training steps: *PickCube* Level 3 success increases from 65% to 92%, and

TABLE III: **Simulation results on the LIBERO benchmark.** Success rates (SR) and rankings are provided for each method across four task suites, averaged over three random seeds with 500 trials.

	LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long		Average	
	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)
Diffusion Policy (scratch) [36]	78.3%	9	92.5%	5	68.3%	9	50.5%	9	72.4%	9
Octo (fine-tuned) [2]	78.9%	8	85.7%	8	84.6%	5	51.1%	8	75.1%	7
OpenVLA (fine-tuned) [3]	84.7%	6	88.4%	7	79.2%	6	53.7%	7	76.5%	6
TraceVLA (fine-tuned) [18]	84.6%	7	85.2%	9	75.1%	8	54.1%	6	74.8%	8
SpatialVLA (fine-tuned) [8]	88.2%	5	89.9%	6	78.6%	7	55.5%	5	78.1%	5
$\pi0$ + FAST (fine-tuned) [38]	96.8%	3	98.8%	2	95.8%	3	85.2%	3	94.2%	3
$\pi0$ (fine-tuned) [4]	96.4%	4	96.8%	4	88.6%	4	60.2%	4	85.5%	4
OpenVLA-OFT (fine-tuned) [31]	97.6%	2	98.4%	3	**97.9%**	1	94.5%	2	97.1%	2
ISF-VLA (fine-tuned)	**99.0%**	1	**99.0%**	1	97.4%	2	**95.8%**	1	**97.8%**	1

TABLE IV: **Evaluation on Bridge tasks using the WidowX robot in SimplerEnv.** ISF-VLA achieves the highest average performance.

Task	Octo-Base [2]	Octo-Small [2]	OpenVLA [3]	RoboVLM [17]	SpatialVLA [8]	$\pi0$ [4]	$\pi0$+FAST [38]	ISF-VLA
Put Spoon on Towel	12.5%	47.2%	0.0%	29.2%	16.7%	29.1%	29.1%	**43.0%**
Put Carrot on Plate	8.3%	9.7%	0.0%	25.0%	25.0%	0.0%	21.9%	**25.0%**
Stack Green on Yellow	0.0%	4.2%	0.0%	12.5%	29.2%	16.6%	10.8%	**38.0%**
Put Eggplant in Basket	43.1%	56.9%	4.1%	58.3%	**100.0%**	62.5%	66.6%	75.3%
Average	16.0%	30.0%	1.0%	31.3%	42.7%	27.1%	32.1%	**45.3%**

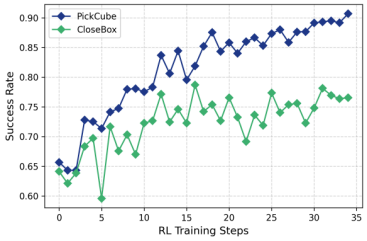


Fig. 4: **Reinforcement learning fine-tuning results.** Performance of ISF-VLA as a policy prior in online reinforcement learning.

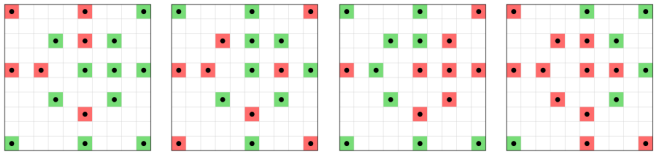
CloseBox Level 3 from 65% to 77%. The slower improvement on *CloseBox* likely results from its larger exploration space and the difficulty of random exploration for non-prehensible objects. These results highlight the effectiveness of ISF-VLA as a strong prior for rapid RL adaptation.

B. Real-World Experiments

1) Experimental Setup: We evaluate **ISF-VLA** on three real-world manipulation tasks to test spatial perception and generalization under realistic noise. Policies are trained with 100 demonstrations collected via human teleoperation using Gello [40] on a Franka Panda and evaluated over 20 trials following prior protocols. Due to the time cost of physical rollouts, we benchmark three most representative baselines: $\pi0.5$ [5], OpenVLA-mini [3], SpatialVLA [8], and compare them to our model, ISF-VLA.

As shown in Fig. 7, we present a brief description of our real-world tasks as follows:

- **LiberoPickButter:** Seven household objects of varying shapes and textures are placed on the table. The robot must identify and pick the butter and place it into a designated basket. The small target and visually complex background demand both precise spatial control and fine-grained semantic discrimination.
- **FetchBowl:** Several bowls with different colors and sizes are randomly arranged. The robot must locate and pick the specific target bowl and place it on top of the drawer. The task requires robust perception under occlusion and accurate grasp planning for similarly shaped objects.
- **LiftPegUpright:** The robot must lift the peg and then place it to the right on the table. Demonstrations are collected from a 10×10 grid of starting positions, with each grid cell dimension being 3 cm, to evaluate unseen spatial configurations. This task demands high precision and tests the model's ability to generalize across fine-grained position and orientation changes.



(a) ISF-VLA (b) SpatialVLA (c) $\pi0.5$ (d) OpenVLA-mini

Fig. 5: **Evaluation under spatial randomization.** Comparison of success rates for ISF-VLA, SpatialVLA, $\pi0.5$ and OpenVLA-mini on LiftPegUpright with varied initial peg positions.

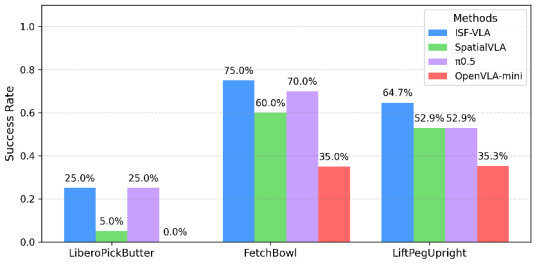


Fig. 6: **Quantitative results on real-world tasks.** ISF-VLA achieves the highest success rates across tasks.

2) Benchmark Results: As shown in Fig. 6, ISF-VLA consistently outperforms both 2D models ($\pi0.5$, OpenVLA-mini) and the explicit 3D model (SpatialVLA) across all three real-world tasks. In **LiftPegUpright**, which requires

TABLE V: **Simulation results on the RoboVerse benchmark.** We report performances on representative tasks from the benchmark across multiple generalization levels.

Task and Generalization Level	MoveSliderLeft (CALVIN)				CloseBox (RLBench)				PickCube (ManiSkill)				PickButter (LIBERO)			
	Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3	Level 0	Level 1	Level 2	Level 3
Diffusion Policy [36]	76.5	81.3	**72.0**	60.0	51.5	42.8	20.0	10.4	52.7	11.1	0.0	0.0	40.0	20.0	0.0	0.0
ACT [37]	**85.0**	**83.3**	43.3	16.6	68.3	73.3	0.0	20.0	31.7	30.0	6.7	3.3	35.0	25.0	15.0	10.0
OpenVLA [3]	45.0	40.0	35.0	30.0	0.0	0.0	0.0	0.0	40.0	15.0	0.0	0.0	25.0	10.0	0.0	0.0
OpenVLA-mini [3]	65.0	70.0	55.0	70.0	80.0	80.0	60.0	50.0	79.0	69.0	55.0	53.0	80.0	85.0	80.0	85.0
π0 [4]	70.0	72.5	55.0	68.0	82.5	85.0	70.0	60.0	88.0	**85.5**	76.8	55.0	77.5	80.5	85.0	75.0
SpatialVLA [8]	67.5	65.0	45.0	55.5	70.0	65.0	67.5	55.0	86.0	73.0	48.2	45.0	65.5	59.0	44.5	40.0
ISF-VLA	75.0	80.0	60.0	**75.0**	**95.0**	**90.0**	**75.0**	**70.0**	**95.0**	80.0	**80.0**	**65.0**	**90.0**	**85.0**	**90.0**	**85.0**

TABLE VI: **Ablation study of ISF-VLA components on RoboVerse CloseBox across generalization levels.**

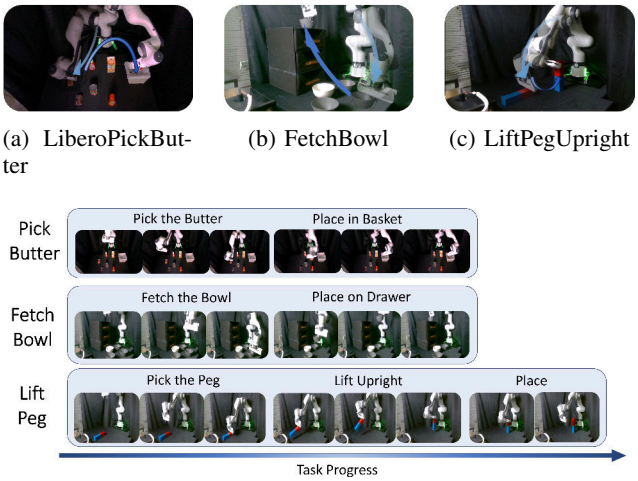
Task and Generalization Level	Settings	CloseBox (RLBench)				
		L0	L0.5	L1	L2	L3
#All	ISF-VLA	**95.0**	77.5	**90.0**	**75.0**	**70.0**
Manipulation Fine-tuning	w/o manipulation-aware fine-tuning	90.0	70.0	85.0	75.0	65.0
	w/o 3D-grounding alignment	85.0	72.5	87.5	70.0	65.0
Fusion Module	w/o cross-attention	80.0	60.0	80.0	65.0	50.0
	w/o DinoSigLiP token concat	85.0	65.0	80.0	70.0	60.0
Camera Tokens	w/o camera tokens	85.0	55.0	80.0	55.0	40.0
	from aggregator	85.0	50.0	85.0	45.0	60.0
Visual Token Selection	w/o visual token selection	95.0	**80.0**	85.0	70.0	70.0
Data Scaling	w/o demo scaling	90.0	72.5	80.0	65.0	70.0
Multi-view	multi-view input (three views)	87.5	65.0	85.0	65.0	40.0
Depth Noise	w/ depth noise	95.0	75.0	85.0	75.0	65.0
Different 3D encoder Choices	w/ DUSt3R [24]	85.0	75.0	80.0	70.0	70.0
	w/ CUT3R [25]	95.0	75.0	87.5	77.5	67.5

TABLE VII: **Ablation results on LIBERO Spatial and Object.**

Settings	Spatial	Object
ISF-VLA	99.0	**99.0**
w/o manipulation fine-tuning	98.4	98.6
w/o cross-attention fusion	97.9	98.2
w/o camera tokens	98.6	98.8
w/o visual token selection	**99.2**	99.0

high precision and generalization across fine-grained spatial variations, ISF-VLA achieves **64.7%** success, surpassing SpatialVLA and π0.5 (52.9%) and OpenVLA-mini (35.3%). In the highly challenging **LiberoPickButter** task, where the small, soft, and nearly textureless target is placed among cluttered objects, ISF-VLA attains **25%** success, outperforming other baselines (5% and 0%) and matching the performance of π0.5. Depth-based 3D models struggle with noisy measurements on reflective or deformable surfaces, while 2D models lack spatial perception; ISF-VLA mitigates these issues via implicit geometric understanding. In **FetchBowl**, which involves varying bowl positions, occlusions, and visually similar objects, ISF-VLA achieves **75%** success, demonstrating its ability to integrate global geometric context with semantic cues for reliable manipulation. Overall, these results confirm that implicit 3D fusion enhances robustness in cluttered, occluded, and fine-grained real-world scenarios.

3) Spatial Generalization in the Real World: To evaluate the model's ability of spatial generalization in the real world, we adopt the **LiftPegUpright** task. The task trains on a 10×10 grid of initial peg positions and tests on 17 unseen placements (black dots in Fig. 5). The robot must lift the peg and place it to the right with high precision, making it highly sensitive to spatial misalignment and control errors.



(a) LiberoPickButter (b) FetchBowl (c) LiftPegUpright

(d) Example stages illustrating our real-world tasks.

Fig. 7: **Our real robot benchmark** consists of three challenging tasks, LiberoPickButter, FetchBowl and LiftPegUpright.

As shown in Fig. 5, ISF-VLA achieves **11/17** successful trials, outperforming SpatialVLA (9/17), OpenVLA-mini (6/17), and even π0.5 (9/17). Its successes are widely distributed across the workspace, demonstrating better extrapolation to unseen poses rather than mere interpolation. This reflects ISF-VLA's ability to capture implicit 3D spatial relations within the feature space, enabling reasoning over fine geometric variations without explicit depth or voxelization. In contrast, SpatialVLA is prone to local spatial bias and depth noise, while π0.5 and OpenVLA-mini struggle with 2D geometric ambiguity. These results underscore that learning implicit, geometry-aware representations supports more robust spatial generalization in real-world manipulation.

V. Conclusion

In this work, we introduced ISF-VLA, a vision–language–action model that integrates implicit 3D fusion for robotic manipulation. By combining 2D semantics with implicit 3D geometry, ISF-VLA learns multi-view–consistent and manipulation-aware representations without requiring explicit 3D reconstruction. Experiments on RoboVerse, LIBERO, and Bridge, including real-world tasks such as LiberoPickButter, FetchBowl, and LiftPegUpright, show that ISF-VLA consistently outperforms both 2D-only and explicit-3D baselines in task success and camera-pose generalization. These results demonstrate that implicit

spatial fusion is an effective approach for improving spatial reasoning in robotic manipulation.

Limitations. The current model relies on an autoregressive action generation architecture. Future work will explore integrating ISF-VLA into a dual-system framework with a flow-matching action head. In addition, multi-layer 2D–3D feature fusion and more efficient training strategies may further improve performance while preserving knowledge from large-scale robotics pretraining.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[2] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[4] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[5] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker *et al.*, "π0.5: a vision-language-action model with open-world generalization," in *9th Annual Conference on Robot Learning*, 2025.

[6] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[7] A. Abdolmaleki, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, A. Balakrishna, N. Batchelor, A. Bewley, J. Bingham, M. Bloesch *et al.*, "Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer," *arXiv preprint arXiv:2510.03342*, 2025.

[8] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025.

[9] T. Lin, G. Li, Y. Zhong, Y. Zou, Y. Du, J. Liu, E. Gu, and B. Zhao, "Evo-0: Vision-language-action model with implicit spatial understanding," *arXiv preprint arXiv:2507.00416*, 2025.

[10] F. Li, W. Song, H. Zhao, J. Wang, P. Ding, D. Wang, L. Zeng, and H. Li, "Spatial forcing: Implicit spatial representation alignment for vision-language-action model," *arXiv preprint arXiv:2510.12276*, 2025.

[11] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[12] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang *et al.*, "Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning," *arXiv preprint arXiv:2504.18904*, 2025.

[13] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *arXiv preprint arXiv:2306.03310*, 2023.

[14] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[15] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[16] O. X.-E. Collaboration, A. O'Neill, Rehman *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[17] H. Liu, X. Li, P. Li, M. Liu, D. Wang, J. Liu, B. Kang, X. Ma, T. Kong, and H. Zhang, "Towards generalist robot policies: What matters in building vision-language-action models," 2025.

[18] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," *arXiv preprint arXiv:2412.10345*, 2024.

[19] 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.

[20] 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.

[21] P. Li, Y. Chen, H. Wu, X. Ma, X. Wu, Y. Huang, L. Wang, T. Kong, and T. Tan, "Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models," *arXiv preprint arXiv:2506.07961*, 2025.

[22] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," *arXiv preprint arXiv:2503.07511*, 2025.

[23] L. Sun, B. Xie, Y. Liu, H. Shi, T. Wang, and J. Cao, "Geovla: Empowering 3d representations in vision-language-action models," *arXiv preprint arXiv:2508.09071*, 2025.

[24] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[25] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa, "Continuous 3d perception model with persistent state," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10 510–10 522.

[26] Q. Team, "Qwen2.5: A party of foundation models," September 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[28] J. Yao, H. Gu, X. Chen, J. Wang, and Z. Cheng, "Open vocabulary monocular 3d object detection," 2024. [Online]. Available: https://arxiv.org/abs/2411.16833

[29] D. Team, "Dinov2: Learning robust visual features without supervision," 2023.

[30] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023.

[31] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[32] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke *et al.*, "Evaluating real-world robot manipulation policies in simulation," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[33] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.

[34] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[35] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," *arXiv preprint arXiv:2107.14483*, 2021.

[36] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[37] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[38] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

[39] H. Li, Y. Zuo, J. Yu, Y. Zhang, Z. Yang, K. Zhang, X. Zhu, Y. Zhang, T. Chen, G. Cui *et al.*, "Simplevla-rl: Scaling vla training via reinforcement learning," *arXiv preprint arXiv:2509.09674*, 2025.

[40] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2024. [Online]. Available: https://arxiv.org/abs/2309.13037