

InfiNoVA: ∞ Novel View Augmentation for Viewpoint Invariant Robot Policies

Sai Puneeth Reddy Gottam, Elmar Rueckert[†] and Vedant Dave[†]

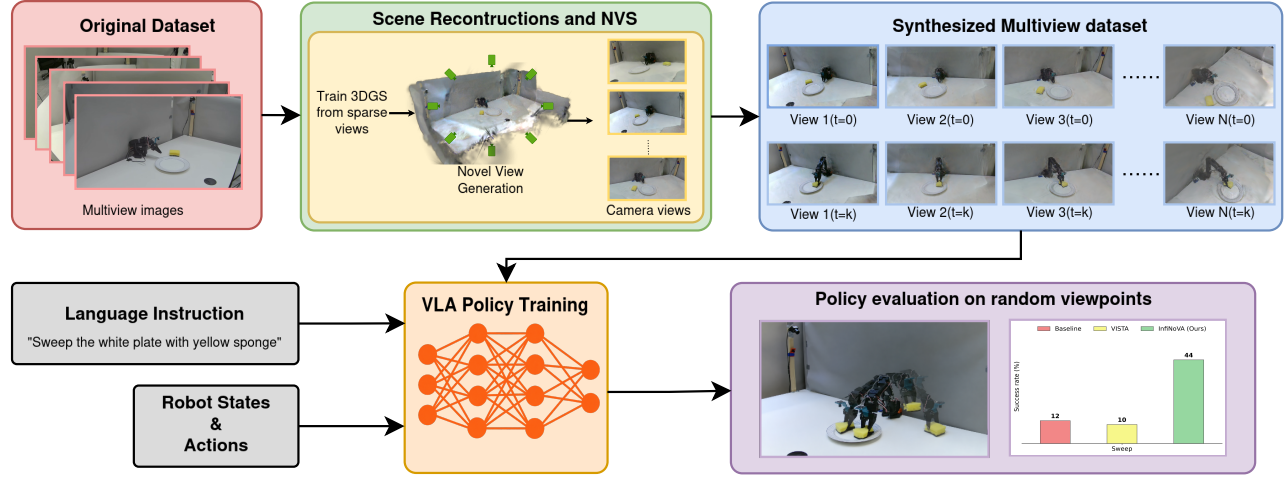


Fig. 1: Our framework (InfiNoVA) creates 3D scene from Gaussian splatting and renders novel views for training VLA policies.

Abstract—Vision-Language-Action (VLA) policies often rely strongly on the camera viewpoints seen during training, causing substantial performance degradation when deployed from unseen perspectives. Collecting demonstrations from sufficiently diverse physical viewpoints is expensive and still provides only sparse coverage of the viewpoint space. We introduce InfiNoVA, a data-augmentation framework that converts synchronized multi-camera demonstrations into a dense distribution of geometrically consistent training views. InfiNoVA reconstructs each manipulation trajectory as a time-varying 3D Gaussian representation and renders novel observations from sampled camera poses while preserving the original state-action correspondence. This explicit scene representation improves frame-level fidelity and temporal consistency while reducing task-critical hallucinations observed in generative novel-view synthesis. Across four real-world manipulation tasks, policies trained with InfiNoVA achieve $5.4\times$ higher average success under unseen randomized viewpoints than both VISTA-based augmentation and the unaugmented policy. InfiNoVA further achieves $1.7\times$ higher success than training directly on all five physical camera views. These results show that dense, geometrically grounded viewpoint augmentation provides a practical route toward camera-robust robot policies without modifying the underlying policy architecture. Project page: <https://infi-nova.github.io>.

I. INTRODUCTION

Vision-Language-Action models (VLAs) have rapidly emerged as a promising interface between large-scale visual-language pretraining and robotic control. By conditioning action prediction on visual observations and language instructions, recent models such as RT-2 [1], OpenVLA [2], π_0 [3], and SmoVLA [4] demonstrate that pretrained multimodal representations can be adapted to robotic manipulation, enabling policies that interpret task intent, perceive objects, and produce actions from raw sensory input.

However, learning such robotic manipulation policies requires large and diverse datasets. This data must capture not only visual observations and language instructions, but also a wide range of object interactions, camera viewpoints, and environmental variations. Collecting such diverse demonstrations in the real world is costly, time-consuming, and difficult to scale. Such data collection often requires human teleoperation, multiple expert demonstrations, repeated task execution, and extensive effort to cover multiple environmental configurations. As a result, most datasets cover only a limited subset of possible viewpoints, lighting conditions, backgrounds, object configurations, and distractor patterns. While simulated environments offer a scalable alternative for collecting enormous amount of data, policies trained in simulation struggle to transfer directly to real hardware due to the sim-to-real gap, caused by discrepancies in visual fidelity, contact dynamics, and unmodeled physical variations [5].

[†]Equal Advising.

All authors are associated with the Chair of Cyber Physical Systems, Montanuniversität Leoben, Austria.

Corresponding author: sai.gottam@unileoben.ac.at



Fig. 2: Multi-view reconstruction and novel-view sampling. Red cameras are used for reconstruction; green cameras denote sampled novel views.

Consequently, policies remain brittle when deployed under real-world conditions that differ from their training setups, particularly because viewpoint and appearance changes can substantially alter the visual evidence available to a VLA without changing the task itself. Recent evaluations [6], [7], [8], [9], [10], [11] show that current VLAs degrade substantially under shifts in camera pose suggesting that they overly rely on camera-specific visual correlations rather than the task structure itself. In particular, LIBERO-Plus [6] reports an average relative performance drop of approximately 69.6% under small camera-viewpoint perturbations across multiple VLA policies, highlighting camera shift as a major source of deployment failure.

A more direct way to scale viewpoint diversity is to physically add cameras. Multi-camera rigs capture the same demonstration from several fixed viewpoints simultaneously, giving policies genuine exposure to different camera poses during training. This improves robustness over single-view setups [12], but scaling further quickly becomes expensive as each additional viewpoint requires extra hardware, calibration, and synchronization, and even a well-instrumented rig only samples a finite, fixed set of positions. A policy trained on such multi-camera data can still overfit to those specific projections, and may fail when the deployment camera lies between, outside, or slightly shifted from the training views, a limitation we further examine in our multi-camera study (Sec. IV-G).

To bypass physical hardware limits, recent work synthesizes new views. Novel-view synthesis (NVS) augmentation [13], [14], [15] removes the hardware bottleneck entirely, generating additional viewpoints directly from existing demonstrations. However, approaches that synthesize views from a single source observation become increasingly challenging as the target viewpoint moves away from the input view, where previously occluded geometry, unseen object surfaces, and background regions must be inferred. Explicit 3D reconstruction methods [16] avoid this by building a full 3D scene representation, offering stronger geometric control, but rely on additional scene-editing and alignment stages to insert or manipulate objects within the reconstruction

The central premise of our work is that a VLA should not learn actions from a small set of fixed, camera-conditioned inputs, but from a family of geometrically valid observations that preserve the same underlying state-action relation. Decoupling action learning from static camera perspectives enables the policy to generalize seamlessly across varying viewpoints. We therefore treat synchronized multi-camera observations not merely as additional training inputs, but as the basis for constructing a continuous viewpoint augmentation space. Given a manipulation trajectory recorded from five cameras, we reconstruct a time-varying 3D Gaussian representation using InstantSplat++ [17], whose pose-free formulation avoids requiring pre-calibrated camera extrinsics for scene reconstruction. The resulting renderable sequence remains temporally aligned with the original trajectory and action labels, allowing additional observations to be generated without relabeling the demonstration. Because the reconstruction is grounded in multiple synchronized real views, we can sample a dense set of virtual camera poses varying in azimuth, elevation, and radial distance within the observed workspace (Fig. 2). This transforms a small set of physical camera observations into a substantially broader distribution of geometrically consistent training viewpoints, without requiring additional cameras or relying solely on single-view extrapolation.

II. RELATED WORK

A. Spatial Grounding in Vision-Language-Action Models

Vision-language-action models (VLAs) extend vision-language pretraining to robotic control, mapping visual observations and language instructions to action sequences and demonstrating strong semantic generalization and cross-task transfer [1], [2], [3], [4], [18], [19]. However, recent evaluations reveal substantial performance degradation under visual and camera-pose perturbations, highlighting limited spatial grounding across viewpoints [9], [10]. Recent methods address this limitation by incorporating geometric information through 3D representations, point clouds, pretrained geometric features, or explicit camera extrinsics [20], [21], [22], [23], [24]. While these approaches improve viewpoint generalization, they often require architectural modifications, additional geometric inputs, or known/estimated camera calibration. CamVLA [11] instead learns a camera-to-base transformation to avoid deployment-time calibration, but remains limited to a single third-person camera and struggles under large viewpoint changes. In contrast, our work addresses viewpoint robustness at the data level by augmenting existing demonstrations with geometrically consistent novel views, without modifying the policy architecture or requiring camera calibration at deployment.

B. Generative Data Augmentation for Robot Generalization

Generative augmentation has been widely used to improve robot-policy robustness without collecting additional demonstrations. Early domain-randomization methods varied colors, textures, and simulator appearance [25], [26], while recent approaches use image-text models, diffusion,

segmentation, and scene-level generation to diversify objects, backgrounds, distractors, and scene context [27], [28], [29], [30], [31], [32]. These methods primarily broaden the appearance and scene distribution while keeping the underlying camera projection fixed. In contrast, our approach expands viewpoint diversity by rendering the same underlying state-action trajectory from novel camera poses while retaining alignment with the original action labels.

C. Novel-View Synthesis for Robust Robotic Policies

Novel-view synthesis (NVS) provides a direct mechanism for augmenting robot demonstrations with unseen viewpoints. VISTA [13] uses single-image diffusion-based NVS to synthesize alternate training views, but its viewpoint range is limited by hallucination of disoccluded geometry and unseen scene content. AnyCamVLA [15] instead performs NVS at test time to transform deployment observations into the policy’s training viewpoint, but relies on camera calibration or pose estimation and on accurate real-time synthesis. Yang et. al. [16] uses Gaussian splatting for viewpoint and appearance augmentation, but requires scene decomposition and geometric alignment for comprehensive object and embodiment editing. Unlike these existing methods, we instead perform training-time viewpoint expansion from real multi-camera observations, reconstructing a scene-specific 3D representation with InstantSplat++. This provides geometrically grounded novel views across a wider range of poses while avoiding single-image hallucination, test-time calibration requirements, and the scene-decomposition overhead of comprehensive scene-editing pipelines.

III. APPROACH

A. Problem Formulation

We consider robotic manipulation policies conditioned on visual observations and language instructions. A demonstration trajectory is represented as $\tau = (o_t, l, a_t)_{t=1}^T$, where o_t is the visual observation, l the language instruction, and a_t the expert action at timestep t . Our goal is to learn a Vision-Language-Action policy $\pi(a_t | o_t, l)$ that generalizes across camera viewpoints.

Training demonstrations typically cover only a small set of camera poses, causing policies to become sensitive to viewpoint-specific visual correlations. We therefore augment an original demonstration dataset $\mathbb{D}$ with novel-view observations to obtain $\mathbb{D}_{nv}$. For each sample (o_t, l, a_t) , we generate an observation $(\tilde{o}_t)$ from a different camera pose while preserving the underlying scene state and action semantics, such that $(\tilde{o}_t, l, a_t)$ remains a valid training sample.

We generate these observations from a reconstructed 3D Gaussian representation of each demonstration, enabling geometrically consistent novel-view rendering without collecting additional robot trajectories. The resulting dataset exposes the policy to a broader camera-view distribution and encourages generalization to viewpoints unseen during training.

B. 3D scene reconstruction and Novel View Synthesis

As shown in the Fig 2, our framework reconstructs the tabletop manipulation scene from synchronized multi-view observations. We use a five-camera setup to capture the workspace from different viewpoints, providing complementary visual evidence of the robot arm, gripper, tabletop, and manipulated objects. For each timestep t , we collect the synchronized set of observations $O_t = \{o_t^1, o_t^2, o_t^3, o_t^4, o_t^5\}$, where o_t^i denotes the image captured by the i -th camera at timestep t . The multi-view image set O_t is given as input to Instantsplat++ [17], which reconstructs a Gaussian Splatting representation of the scene $G_t = f_{GS}(O_t)$.

We use Instantsplat++ because it provides fast feed-forward sparse-view reconstruction and does not require externally provided camera parameters. This pose-free property is useful for robotic demonstration data, where accurate camera calibration may not always be available. The reconstructed representation G_t captures the task-relevant scene, including the tabletop workspace, robot arm, gripper, and manipulated objects. After reconstructing the scene at timestep t , we render novel views from a set of virtual camera poses $C = \{c_1, c_2, \dots, c_K\}$ placed around the workspace $\tilde{o}_t^k = R(G_t, c_k)$, where R denotes the rendering function and $\tilde{o}_t^k$ is the rendered observation from virtual camera pose c_k . Repeating this process for all timesteps produces novel-view trajectories that preserve the original robot motion and object configuration. Therefore, each rendered observation can be paired with the original language instruction and action label.

The multi-view setting is important for reducing hallucinations in novel-view synthesis. Single-view approaches are highly under constrained because they must infer unseen geometry, depth, and occluded regions from only one image. This can lead to unrealistic or task-inconsistent outputs, particularly around the gripper, robot arm, and manipulated objects. By using five synchronized camera views, our reconstruction receives stronger geometric cues and produces more consistent novel-view renderings. We adopt a Gaussian Splatting-based approach for rendering realistic images while preserving 3D consistency. In contrast, diffusion-based novel-view generation methods may introduce visual artifacts, change object geometry, or generate images that appear synthetic in robot manipulation settings. Such methods may also require scene-specific fine-tuning, making them difficult to generalize across different tasks and environments. Our reconstruction-based pipeline avoids this issue by producing novel views directly from the observed multi-view scene.

C. VLA Policy Training

After generating novel-view observations, we use the resulting dataset to train a Vision-Language-Action policy. The policy receives a visual observation and a natural language instruction as input and predicts the robot action required to complete the task. The key advantage of our new augmented dataset is that each generated observation preserves the original task state, allowing us to reuse the same language instruction and action label from the source demonstration.

Let $\mathbb{D}$ denote the original demonstration dataset, and $\mathbb{D}_{\text{nv}}$ the dataset created through InstantSplat++ based novel-view rendering. The final training dataset is defined as $\mathbb{D}_{\text{train}} = \mathbb{D} \cup \mathbb{D}_{\text{nv}}$. Each sample in $\mathbb{D}_{\text{train}}$ has the form (o_t, l, a_t) , where o_t may be an original observation, or a rendered novel view image.

We train the VLA policy using imitation learning. For continuous-action policies, the model is optimized to minimize the difference between the predicted action and the expert action: $\mathcal{L}_{\text{act}} = \sum (o_t, l, a_t) \in \mathbb{D}_{\text{train}} |\pi(o_t, l) - a_t|_2^2$. For action-token-based VLA models, the same training data can be used with a cross-entropy objective over discretized action tokens, $\mathcal{L}_{\text{token}} = - \sum (o_t, l, a_t) \in \mathbb{D}_{\text{train}} \log p\pi(a_t | o_t, l)$.

During training, each mini-batch contains a mixture of original and augmented observations. Since the same action may be paired with multiple visual variants of the same task state, the policy is encouraged to learn visual representations that are invariant to camera viewpoints. This reduces reliance on spurious cues such as fixed camera pose, tabletop color, or background texture, and encourages the model to focus on task-relevant information such as object identity, object location, gripper pose, and language-conditioned intent. Our augmentation framework is model-agnostic and can be applied to any VLA architecture. It does not require changing the policy architecture or action space; instead, it improves the visual diversity of the demonstration data used for training. At inference time, the policy receives only the current visual observation and language instruction, and no reconstruction module is required.

D. Experimental setup

We evaluate our approach on four manipulation tasks: *Pick-and-Place Cube*, *Sweep*, *Sort*, and *Stack*, chosen to span different levels of occlusion, object geometry, and manipulation precision. For each task, we collect 70 demonstrations using a synchronized five-camera setup mounted around an SO-101 manipulator. From these multi-view observations, we reconstruct a per-timestep 3D Gaussian representation using InstantSplat++.

We define camera azimuth with respect to the center of the manipulation workspace, with 0° corresponding to the frontal viewing direction. The five physical cameras span approximately a 180° arc around the workspace, and novel camera poses are restricted to $\theta \in [-\pi/2, \pi/2]$. We additionally vary the radial distance from the workspace center within $r \in [0.6, 1.2]$ m. Elevation is sampled from the table-plane level up to a top-down view, corresponding to $\phi \in [0, \pi/2]$. This sampling strategy avoids viewpoints behind the robot while introducing diversity in viewing angle, scale, and framing. We sample up to 15 novel third-person views per timestep. These views are then combined to construct up to 100 virtual episodes from each original demonstration, with camera combinations sampled independently across demonstrations.

We use SmolVLA as the policy architecture for all experiments, with three visual inputs: one wrist-mounted camera and two third-person cameras. For each virtual episode,

we retain the original wrist-camera observations and action labels, while replacing the two third-person observations with synthesized views from our InfiNoVA pipeline. This results in 7,000 episodes for each of the tasks. The single-view baseline is trained using the original reference-camera observations, while the VISTA baseline is trained using data generated by VISTA. For a fair comparison, all policies use the same SmolVLA training configuration, are trained for the same number of epochs with a batch size of 64, and receive no method-specific hyperparameter tuning.

IV. RESULTS

A. Baselines

For novel-view rendering, we compare InfiNoVA against methods spanning single-view generation, multi-view diffusion, explicit 3D generation, and Gaussian Splatting. **VISTA** [13] is the closest robotics-specific baseline, as it uses novel-view synthesis for viewpoint-robust policy augmentation; for a fair comparison, we fine-tune its NVS model on our environment. **GLD** [33] provides a multi-view diffusion baseline, contrasting InfiNoVA’s explicit Gaussian reconstruction with generative multi-view synthesis. **Hunyuan3D** [34] represents an explicit 3D-generation alternative, while **DepthSplat** [35] provides the closest reconstruction baseline by combining sparse multi-view inputs with feed-forward Gaussian Splatting.

For policy evaluation, we compare against a **Base Policy** (SmolVLA) trained on the original camera views and evaluated under unseen randomized viewpoints, and **VISTA-Aug**, trained with VISTA-generated views. We use VISTA as the policy-level synthesis baseline because it is the only competing method that remains sufficiently realistic and task-consistent for downstream augmentation; the others either exhibit substantial visual and geometric artifacts or incur impractical computational cost (Section IV-C).

B. Quality of Novel-view Renderings

We first evaluate whether multi-view 3D reconstruction provides higher-quality and more reliable novel-view augmentation than existing synthesis approaches (Fig. 3). We compare all methods using the same source observations and target viewpoints, randomly sampling 100 observations from each of the four tasks, for a total of 400 evaluation samples. Each generated view is compared against the corresponding real camera observation using PSNR, SSIM [36], and LPIPS [37]. As reported in Table I, InfiNoVA achieves the best performance across all three metrics, with a PSNR of 16.65, SSIM of 0.841, and LPIPS of 0.388. Compared with the three strongest baselines, InfiNoVA improves PSNR by 4.3–25.5% and SSIM by 2.6–11.7%, while reducing LPIPS by 15.5–25.0%. These results demonstrate consistently higher pixel-level, structural, and perceptual fidelity across the evaluated novel views.

The qualitative results in Fig. 3 further expose the main failure modes of the competing methods. Hunyuan3D often reconstructs incomplete workspaces with fragmented objects, while DepthSplat preserves the coarse layout but

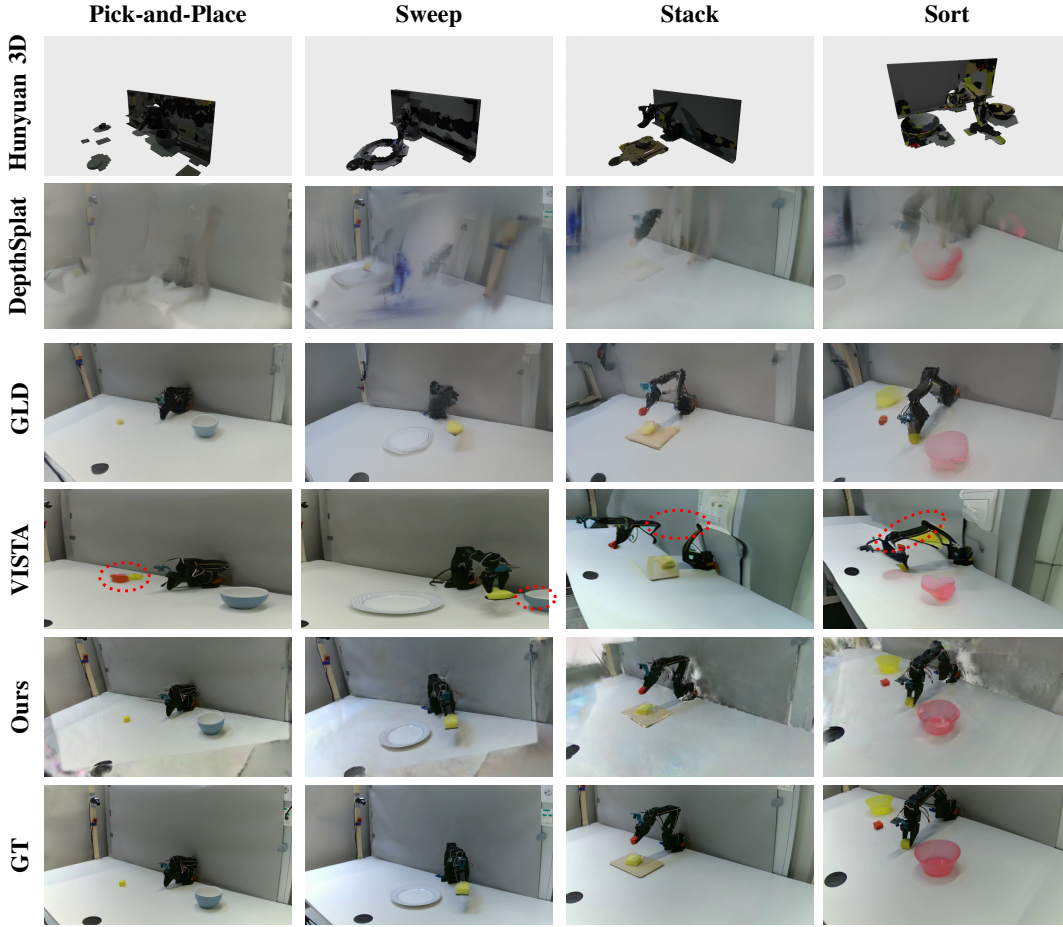


Fig. 3: Qualitative comparison of novel-view synthesis across the four manipulation tasks. Rows correspond to competing synthesis methods. Red ellipses highlight representative task-critical artifacts in VISTA.

TABLE I: Quantitative comparison of novel-view rendering quality. Higher PSNR and SSIM indicate better reconstruction fidelity, while lower LPIPS indicates better perceptual similarity.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Hunyuan 3D [34]	7.48	0.693	0.596
DepthSplat [35]	13.27	0.809	0.517
GLD [33]	15.97	0.753	0.485
VISTA [13]	15.13	0.820	0.459
InfiNoVA (Ours)	16.65	0.841	0.388

produces blurred robot and object geometry. GLD gives the strongest visual quality among the baselines, but incurs substantially higher computational cost (Table II). VISTA generates realistic-looking images but introduces task-critical hallucinations, highlighted by the red ellipses in Fig. 3. In Pick-and-Place, it incorrectly introduces two sponges that belong to the Stack task, while in Sweep it hallucinates the bowl observed in Pick-and-Place. In Stack, part of the robot arm is missing, and in Sort the arm geometry is visibly distorted.

Because VISTA is fine-tuned on the robot-task data but is not explicitly constrained by the task identity or complete 3D state of the current episode, ambiguous or disoccluded

regions can be filled using dataset-level visual priors, leading to cross-task content leakage. This is particularly problematic for policy augmentation because the hallucinated frame retains the original action label. In contrast, InfiNoVA renders from an explicit 3D Gaussian Splat reconstructed from synchronized multi-view observations of the current scene, so object configurations and disoccluded regions are grounded in observed geometry, preserving the robot, manipulated objects, and scene structure more reliably and without borrowing task-specific content from other demonstrations.

C. Rendering Time

Table II compares the computational cost of novel-view synthesis, including reconstruction and rendering, for the strongest methods from Section IV-B. Although GLD produces competitive renderings, its computational cost is prohibitively high, making trajectory-scale generation impractical for temporal and downstream policy evaluation; we therefore exclude it from those experiments. Averaged across tasks, InfiNoVA is approximately 3.6 $\times$ faster than VISTA and 82.8 $\times$ faster than GLD, while also providing higher rendering fidelity and temporal consistency (Section IV-B). Overall, InfiNoVA achieves the strongest quality–efficiency trade-off among the evaluated methods.

TABLE II: Novel-view synthesis time comparison (in seconds).

Task	GLD	VISTA	InfiNoVA
Pick-and-Place	25615	1115	318
Sweep	35119	1529	437
Sort	39026	1693	434
Stack	45056	1961	560
Average/episode	36204	1575	437

D. Temporal and Geometric Consistency

We next evaluate whether the generated trajectories preserve the temporal evolution and task state of the original demonstrations. Beyond frame-wise rendering quality, a valid augmentation should reproduce the motion observed in the real trajectory without introducing artificial changes across consecutive frames. To quantify this, we randomly sample 10 trajectories from each of the four tasks, and evaluate them using two complementary feature-space temporal metrics. 1) *FeaCD* [38] evaluates ResNet-18 [39] features from consecutive frames and their inter-frame feature differences are treated as temporal change vectors, and cosine distance measures whether the generated and ground-truth trajectories evolve in the same direction in feature space. 2) We additionally introduce *ReStraV- $\Delta\theta$* , a reference-based adaptation of ReStraV [40]. ReStraV represents video frames using DINOv2 [41] features and characterizes temporal trajectories through their local curvature. Given frame representations z_t , we first compute consecutive displacements $\Delta z_t = z_{t+1} - z_t$ and the local trajectory angle

$$\theta_t = \arccos \left(\frac{\Delta z_t^\top \Delta z_{t+1}}{\|\Delta z_t\|_2 \|\Delta z_{t+1}\|_2} \right). \quad (1)$$

Rather than using curvature to distinguish real and generated videos as in the original formulation, we compute these angles independently for the synthesized and ground-truth trajectories and report their mean absolute angular deviation,

$$ReStraV - \Delta\theta = \frac{1}{T-2} \sum_{t=1}^{T-2} |\hat{\theta}_t - \theta_t^{GT}|. \quad (2)$$

This measures whether the synthesized trajectory follows the same local geometry through DINOv2 feature space as the real trajectory. As reported in Table III, InfiNoVA achieves the best performance on both metrics, with a FeaCD of 0.915 and a ReStraV- $\Delta\theta$ of 3.49°. Compared with the next-best baseline, DepthSplat, this corresponds to reductions of 5.6% in FeaCD and 32.8% in angular deviation. Relative to VISTA, the reductions are 7.0% and 51.3%. These gains show that InfiNoVA better preserves both the frame-to-frame direction of semantic change and the higher-order temporal geometry of the real trajectory, yielding a more coherent progression of the underlying manipulation state.

The qualitative results in Fig. 4 further expose these temporal failures. The three observations are sampled from the same Pick-and-Place episode at timestep t_{90} , t_{494} , and t_{512} . At t_{90} , VISTA completely removes the yellow cube that is present in the corresponding scene. At t_{494} , the bowl

TABLE III: Temporal consistency averaged across the four manipulation tasks. Lower values indicate better agreement with the ground-truth trajectories.

Method	FeaCD ↓	ReStraV- $\Delta\theta$ (°) ↓
Hunyuan 3D [34]	0.9780	6.09
DepthSplat [35]	0.969	5.19
VISTA [13]	0.984	7.16
InfiNoVA (Ours)	0.915	3.49

disappears despite remaining part of the same episode. More severely, at t_{512} a yellow cube appears at the location of the missing bowl while the robot is simultaneously holding another cube, resulting in two instances of an object that should occur only once. Such inconsistencies are particularly harmful for policy augmentation because each synthesized observation retains the action label of the corresponding real state. Removing, duplicating, or replacing task-relevant objects therefore pairs the expert action with an observation describing a different physical state, introducing contradictory state-action supervision. InfiNoVA instead preserves these scene elements across the sampled timesteps, remaining consistent with the progression of the real episode.

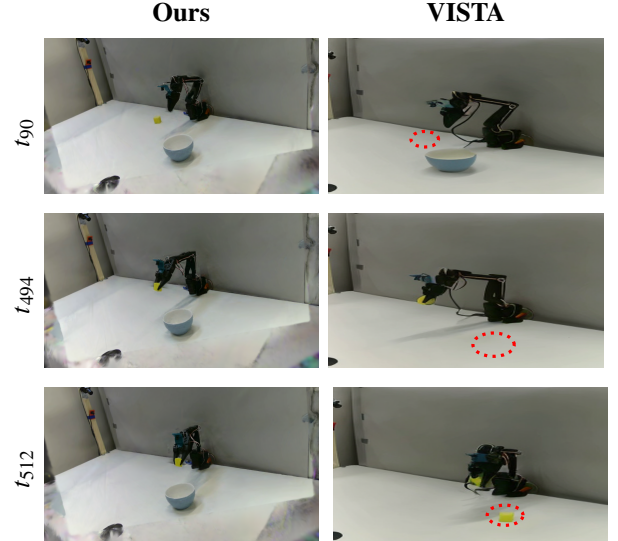


Fig. 4: Temporal consistency across three timesteps (t_{90} , t_{494} , and t_{512}) sampled from the same Pick-and-Place episode. Red ellipses highlight VISTA’s temporal inconsistencies.

E. Rendering Precision

We next evaluate whether novel-view synthesis produces repeatable outputs under identical input conditions for three best methods based on their rendering quality. Repeatability is relevant for policy training because each synthesized observation retains the original action label, while uncontrolled changes in robot or object state can weaken state-action correspondence. We sample $N = 5$ observations and generate each target view $K = 5$ times, keeping the source images and target camera pose fixed while varying the random seed. For observation n , let $\hat{I}_n^{(k)} = G(x_n, c_n; \xi_k)$ denote the output

generated from source observations x_n , target camera pose c_n , and random seed ξ_k . Following the LPIPS-based diversity measure of [42], we quantify rendering precision using the average pairwise perceptual variation,

$$V_{\text{LPIPS}} = \frac{2}{NK(K-1)} \sum_{n=1}^N \sum_{i=1}^{K-1} \sum_{j=i+1}^K \text{LPIPS}(\hat{I}_n^{(i)}, \hat{I}_n^{(j)}), \quad (3)$$

which averages over all 10 distinct output pairs per observation and subsequently across observations. Lower values indicate greater repeatability, with identical outputs yielding zero variation. As reported in Table IV, InfiNoVA achieves lower pairwise perceptual variation than the compared stochastic synthesis baselines, indicating greater consistency across repeated generations under unchanged conditioning. We also report mean LPIPS against the corresponding real target-view observations, since a constant but incorrect output can achieve perfect repeatability. Together, these metrics assess both consistency and fidelity, supporting the reliability of our augmentations.

TABLE IV: Values report average pairwise LPIPS and mean LPIPS against ground truth over 5 generations for each of 5 observations.

Method	Pairwise LPIPS ↓	LPIPS to GT ↓
GLD [33]	0.5428	0.4810
VISTA [13]	0.2765	0.5002
InfiNoVA (Ours)	0.1406	0.466

F. Policy Evaluation

We evaluate whether novel-view augmentation improves policy robustness to camera viewpoint shifts across four manipulation tasks: Pick-and-Place, Stack, Sweep, and Sort. Figure 5 reports success rates over 100 rollouts per result. Baseline (Reference) is trained and evaluated on the original reference viewpoint, whereas Baseline (Random) evaluates the same policy from randomized viewpoints to measure degradation under viewpoint shift. VISTA and InfiNoVA are trained on datasets augmented with views synthesized by their respective frameworks and evaluated using the same randomized-view protocol as Baseline (Random).

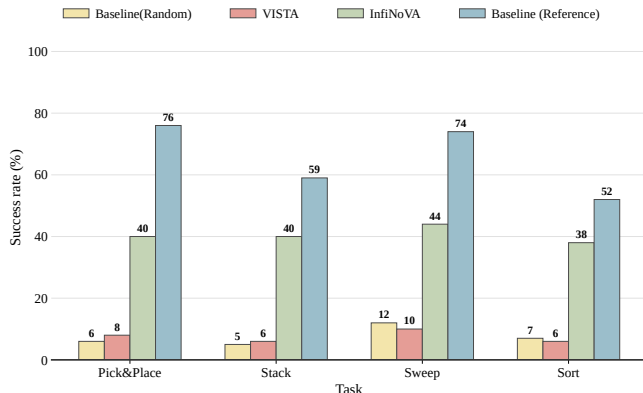


Fig. 5: Policy success rates across four manipulation tasks.

Averaged across the four tasks in Fig. 5, the Baseline (Reference) achieves an average success rate of 65.3%. When the same single-view policy is evaluated from randomized camera viewpoints, average success drops to 7.5%, corresponding to an 88.5% relative degradation and highlighting its strong dependence on the training viewpoint.

VISTA-based augmentation provides no overall recovery, with an average success rate of 7.5%. In contrast, InfiNoVA achieves 40.5% average success, corresponding to a 5.4× improvement over VISTA under viewpoint perturbations. Relative to the single-view policy under camera shift, InfiNoVA recovers approximately 57% of the performance lost with respect to the in-distribution reference.

The remaining gap to the Baseline (Reference) reflects the harder evaluation setting: the reference policy is trained and tested from the same fixed viewpoint, whereas InfiNoVA is evaluated across randomized camera poses. These results demonstrate substantially improved generalization to unseen viewpoints while retaining the original policy architecture.

G. Beyond Multi-Camera Training

To evaluate the benefit of viewpoint synthesis from simply adding more physical cameras, we compare three policies. The single-view baseline is trained using demonstrations from only the reference camera view. The multi-view baseline is trained on the same demonstrations recorded from all five physical camera views, without NVS augmentation. Our method augments these demonstrations with densely synthesized novel views. All policies are evaluated over 100 Pick-and-Place trials using the same unseen camera poses with randomized viewpoints and distances. The single-view baseline achieves 76% success on its training view but drops to 6% on random views. Multi-view training improves performance to 24%, while InfiNoVA increases success to 40%. These results show that dense viewpoint augmentation provides stronger generalization than training on a sparse set of physical views.

V. CONCLUSION

We presented InfiNoVA, a novel-view augmentation framework that converts sparse synchronized multi-camera demonstrations into a dense distribution of geometrically grounded training views. By reconstructing manipulation trajectories as time-varying Gaussian Splat representations, InfiNoVA preserves task-relevant scene structure and state-action correspondence while varying the camera viewpoint. Across four real-world manipulation tasks, InfiNoVA produces higher-fidelity, more temporally consistent, and more repeatable novel views than existing synthesis approaches. These improvements translate directly to downstream robustness as under unseen randomized viewpoints, InfiNoVA-trained policies achieve 5.4× higher average success than both VISTA-based augmentation and the unaugmented policy. This shows that viewpoint diversity alone is insufficient; synthetic views must also preserve the underlying task state to avoid contradictory state-action supervision. InfiNoVA further achieves 1.7× higher success than training directly on all five physical camera views, demonstrating the benefit of

dense synthesized viewpoint coverage beyond sparse multi-camera training.

The current study focuses on a controlled real-world setting with synchronized multi-camera observations, where the available viewpoints provide strong geometric coverage of the manipulation workspace. While our experiments demonstrate robust generalization across randomized unseen viewpoints, extending the sampling range further beyond the observed workspace would be an interesting direction for future study. Our evaluation considers four manipulation tasks on a single robot platform with SmolVLA, providing a consistent setting for isolating the effect of viewpoint augmentation. Evaluating the approach across additional embodiments, policy architectures, and more diverse long-horizon tasks would further establish its generality. Finally, the current pipeline reconstructs scene representations offline at each timestep; future work could explore temporally shared reconstruction or more compact scene representations to further improve scalability.

REFERENCES

- [1] B. Zitkovich, *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” in *Conf. on Robot Learning*. PMLR, 2023, pp. 2165–2183.
- [2] M. J. Kim, *et al.*, “Openvla: An open-source vision-language-action model,” in *Proceedings of The 8th Conf. on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, *et al.*, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 2679–2713.
- [3] K. Black, *et al.*, “pi-0: A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [4] M. Shukor, *et al.*, “Smolvla: A vision-language-action model for affordable and efficient robotics,” *arXiv preprint arXiv:2506.01844*, 2025.
- [5] X. B. Peng, *et al.*, “Sim-to-real transfer of robotic control with dynamics randomization,” in *2018 IEEE international Conf. on robotics and automation (ICRA)*. IEEE, 2018, pp. 3803–3810.
- [6] S. Fei, *et al.*, “Libero-plus: In-depth robustness analysis of vision-language-action models,” *arXiv preprint arXiv:2510.13626*, 2025.
- [7] Z. Wang, *et al.*, “Vlatest: Testing and evaluating vision-language-action models for robotic manipulation,” *Proceedings of the ACM on Software Engineering*, 2025.
- [8] J. Gao, *et al.*, “A taxonomy for evaluating generalist robot manipulation policies,” *IEEE Robotics and Automation Letters*, 2026.
- [9] A. Xie, *et al.*, “Decomposing the generalization gap in imitation learning for visual robotic manipulation,” in *2024 IEEE International Conf. on Robotics and Automation (ICRA)*, 2024, pp. 3153–3160.
- [10] W. Pumacay, *et al.*, “The COLOSSEUM: A benchmark for evaluating generalization for robotic manipulation,” in *RSS 2024 Workshop: Data Generation for Robotics*, 2024. [Online]. Available: <https://openreview.net/forum?id=aiSZqS5wkB>
- [11] W. Li, *et al.*, “From fixed to free cameras: Calibration-free view-robust vision-language-action model,” *arXiv preprint arXiv:2607.05396*, 2026.
- [12] J. Yang, *et al.*, “Mobi- π : Mobilizing your robot learning policy,” in *Proceedings of The 9th Conf. on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, *et al.*, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 3516–3536. [Online]. Available: <https://proceedings.mlr.press/v305/yang25b.html>
- [13] S. Tian, *et al.*, “View-invariant policy learning via zero-shot novel view synthesis,” in *Conf. on Robot Learning*. PMLR, 2025, pp. 1173–1193.
- [14] A. Zhou, *et al.*, “Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis,” in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 17907–17917.
- [15] H. Heo, *et al.*, “Anycamvla: Zero-shot camera adaptation for viewpoint robust vision-language-action models,” *arXiv preprint arXiv:2603.05868*, 2026.
- [16] S. Yang, *et al.*, “Novel Demonstration Generation with Gaussian Splatting Enables Robust One-Shot Manipulation,” in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025.
- [17] Z. Fan, *et al.*, “Instantsplat: Sparse-view gaussian splatting in seconds,” 2025. [Online]. Available: <https://arxiv.org/abs/2403.20309>
- [18] A. O’Neill, *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0,” in *2024 IEEE International Conf. on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.
- [19] O. M. Team, *et al.*, “Octo: An open-source generalist robot policy,” *arXiv preprint arXiv:2405.12213*, 2024.
- [20] D. Qu, *et al.*, “Spatialvla: Exploring spatial representations for visual-language-action model,” *arXiv preprint arXiv:2501.15830*, 2025.
- [21] L. Sun, *et al.*, “Geovla: Empowering 3d representations in vision-language-action models,” *arXiv preprint arXiv:2508.09071*, 2025.
- [22] A. Abouzeid, *et al.*, “Geoaware-vla: Implicit geometry aware vision-language-action model,” *arXiv preprint arXiv:2509.14117*, 2025.
- [23] T. Jiang, *et al.*, “Do you know where your camera is? View-invariant policy learning with camera conditioning,” in *IEEE International Conf. on Robotics and Automation (ICRA)*, 2026.
- [24] T. Zhang, *et al.*, “Grounding actions in camera space: Observation-centric vision-language-action policy,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 22, 2026, pp. 18 782–18 790.
- [25] J. Tobin, *et al.*, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *2017 IEEE/RSJ international Conf. on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.
- [26] R. Alghonaim *et al.*, “Benchmarking domain randomisation for visual sim-to-real transfer,” in *2021 IEEE international Conf. on robotics and automation (ICRA)*. IEEE, 2021, pp. 12 802–12 808.
- [27] Q. Chen, *et al.*, “GenAug: Retargeting behaviors to unseen situations via Generative Augmentation,” in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.
- [28] T. Yu, *et al.*, “Scaling Robot Learning with Semantically Imagined Experience,” in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.
- [29] Z. Chen, *et al.*, “Semantically controllable augmentations for generalizable robot learning,” *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1705–1726, 2025.
- [30] E. Teoh, *et al.*, “Green screen augmentation enables scene generalisation in robotic manipulation,” *arXiv preprint arXiv:2407.07868*, 2024.
- [31] C. Yuan, *et al.*, “Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation,” in *2025 IEEE/RSJ International Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 7622–7629.
- [32] X. Wang, *et al.*, “Roboaug: One annotation to hundreds of scenes via region-contrastive data augmentation for robotic manipulation,” *arXiv preprint arXiv:2602.14032*, 2026.
- [33] W. Jang, *et al.*, “Repurposing geometric foundation models for multi-view diffusion,” in *European Conf. on Computer Vision (ECCV)*, 2026.
- [34] T. H. Team, “Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material,” 2025.
- [35] H. Xu, *et al.*, “Depthspat: Connecting gaussian splatting and depth,” in *2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 16 453–16 463.
- [36] Z. Wang, *et al.*, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, 2004.
- [37] R. Zhang, *et al.*, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conf. on computer vision and pattern recognition*. IEEE, 2018, pp. 586–595.
- [38] M. Gong, *et al.*, “Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration,” in *2025 IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2025, pp. 14 326–14 335.
- [39] K. He, *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conf. on computer vision and pattern recognition*, 2016, pp. 770–778.
- [40] C. Internò, *et al.*, “Ai-generated video detection via perceptual straightening,” *Advances in neural information processing systems*, 2026.
- [41] M. Oquab, *et al.*, “Dinov2: Learning robust visual features without supervision,” 2024. [Online]. Available: <https://arxiv.org/abs/2304.07193>
- [42] U. Ojha, *et al.*, “Few-shot image generation via cross-domain correspondence,” in *2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 10 738–10 747.