RETHINKING HIGH-FIDELITY VIEW SYNTHESIS: SIMPLE HOMOGRAPHY AUGMENTATION FOR VIEWPOINT-ROBUST VLAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models perform strongly across diverse language-conditioned manipulation tasks but remain sensitive to camera viewpoint changes. Recent 3D-aware augmentation methods address this challenge by generating observations from novel viewpoints through learned view synthesis or 3D reconstruction. However, generating these views is computationally expensive and requires additional pretrained models. To address these limitations, we propose **SHARP** (**S**imple **H**omography **A**ugmentation for **R**obust **P**olicies), a lightweight view-augmentation method that improves VLA robustness to unseen third-person camera viewpoints. For VLAs that combine third-person and wrist-camera observations, SHARP applies randomly sampled homographies to third-person images during policy finetuning while preserving the paired wrist observations and action labels. These inexpensive 2D transformations enable online augmentation of demonstrations from a single third-person viewpoint, without additional pretrained synthesis or reconstruction models. Compared with policies finetuned on the same data without augmentation, SHARP improves average success rates across all evaluated viewpoints by 11.8 percentage points on RLBench and 16.4 percentage points on LIBERO. It achieves success rates competitive with the evaluated view-synthesis methods at substantially lower augmentation cost. Comparisons with alternative 2D augmentations highlight the benefit of geometric augmentation for viewpoint robustness. Attention analyses show a shift toward wrist observations, while camera-input analyses suggest that SHARP still benefits from third-person information under certain viewpoint shifts. These results suggest that lightweight geometric augmentation can improve viewpoint robustness in VLAs with stable wrist-camera inputs, even without faithfully reconstructing novel views.

1 INTRODUCTION

Vision-Language-Action (VLA) models (Brohan et al., 2023; Kim et al., 2024; Black et al., 2024) can perform diverse manipulation tasks based on visual observations and language instructions. However, changes in camera viewpoint can substantially degrade their performance (Fei et al., 2025). Such shifts alter the apparent positions, sizes, and visibility of objects, even when the task and physical scene remain unchanged. This sensitivity limits the reuse of demonstrations collected with a fixed camera setup. An important practical goal is to improve viewpoint robustness without collecting additional demonstrations at the target viewpoints.

Prior work addresses viewpoint changes in several ways. Some methods incorporate camera geometry into policy inputs (Jiang et al., 2025), while others transform test-time observations to resemble the training viewpoint (Heo et al., 2026). A promising line of work explores 3D-aware data augmentation, expanding policy training data with observations from different viewpoints. For example, VISTA uses learned novel-view synthesis (Tian et al., 2024), while GenSplat renders additional views from feed-forward 3D Gaussian reconstructions (Wang et al., 2026a). These methods improve viewpoint robustness without modifying the policy architecture or adding inference-time processing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

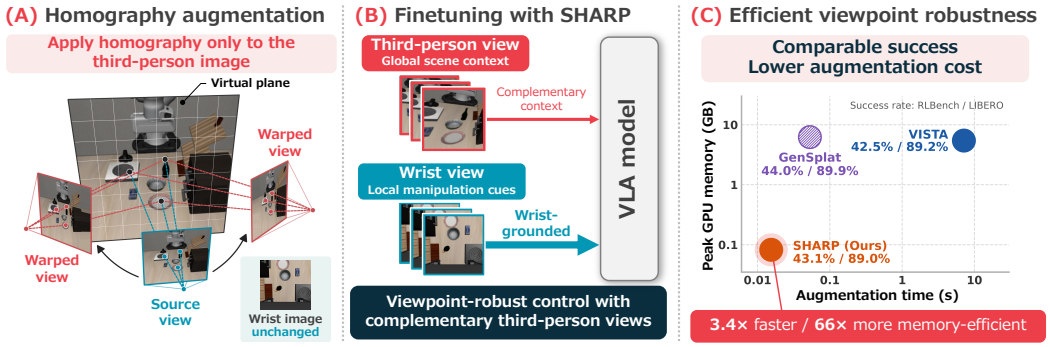

Figure 1: Overview of SHARP. **(A) Homography augmentation.** Random homographies transform source-view third-person images while preserving paired wrist observations and action labels. **(B) Finetuning with SHARP.** Our analyses suggest that SHARP encourages wrist-grounded control complemented by third-person information. **(C) Efficient viewpoint robustness.** SHARP achieves success rates competitive with the evaluated view-synthesis methods while substantially reducing augmentation time and peak GPU memory use.

These augmentation methods, however, require substantial computation for novel-view synthesis, raising the question of whether faithful view synthesis is necessary for viewpoint robustness when stable wrist observations are available. Since the wrist-camera setup remains unchanged under third-person viewpoint shifts, augmentation may improve robustness partly by encouraging greater use of wrist observations. Our analysis of a VISTA-finetuned policy shows increased attention to wrist images (Section 3), consistent with this possibility. We further hypothesize that geometric variation in third-person images can provide additional benefits alongside wrist-grounded control, even without faithful view synthesis. We therefore investigate lightweight 2D transformations as a low-cost approach to introducing such variation and improving viewpoint robustness.

We propose **SHARP** (**S**imple **H**omography **A**ugmentation for **R**obust **P**olicies), a lightweight augmentation method for improving VLA robustness to viewpoint changes. During finetuning, SHARP applies randomly sampled homographies (Hartley & Zisserman, 2004) to third-person images while retaining the paired wrist observations and action labels (Figure 1(A, B)). These lightweight 2D warps combine changes in image position, scale, orientation, and perspective to approximate image variations associated with viewpoint changes. The low cost of these transformations enables fresh augmentations at each training step. The method uses demonstrations collected from a single third-person viewpoint and requires neither target-view demonstrations nor a learned view-synthesis model.

We evaluate SHARP with $\pi_{0.5}$ on RLBench (James et al., 2020) and LIBERO (Liu et al., 2023). At evaluation time, we change the third-person camera viewpoint while keeping the wrist camera's mounting position fixed. Compared with matched finetuning without homography augmentation, SHARP improves average success by 11.8 and 16.4 percentage points on RLBench and LIBERO, respectively. Comparisons with VISTA, with and without generator finetuning, and GenSplat show competitive success rates at lower image-generation cost (Figure 1(C)). In our measurements, SHARP achieves at least a $3.4\times$ reduction in per-image generation time and a $66\times$ reduction in peak GPU memory use relative to the evaluated synthesis pipelines. We further investigate what contributes to SHARP's gains in viewpoint robustness. Our analyses show that homography augmentation outperforms the evaluated alternative augmentation methods, highlighting the value of geometric variation during training. Attention analyses show a shift toward wrist observations, while SHARP outperforms a wrist-only policy across a broad range of viewpoint shifts.

In this work, we make three main contributions: (1) We introduce SHARP, a lightweight online homography augmentation method for improving VLA viewpoint robustness in settings with both wrist and third-person cameras. (2) We show that SHARP achieves success rates competitive with the evaluated view-synthesis methods at substantially lower augmentation cost on RLBench and LIBERO. Its robustness gains extend across three VLA backbones on LIBERO. (3) We analyze the roles of geometric augmentation and the two camera streams. SHARP outperforms the evaluated alternative image augmentations, while camera-input analyses show increased attention to wrist observations and an advantage over a wrist-only policy across a broad range of viewpoint shifts.

2 Problem Formulation

Our objective is to improve the task success of a pretrained VLA policy under third-person camera viewpoint changes without using target-view demonstrations during finetuning. We consider language-conditioned manipulation tasks. At environment timestep t, the observation $o_t = (I_t^{\text{third}}, I_t^{\text{wrist}}, s_t, \ell)$ contains a third-person image, a wrist-camera image, a proprioceptive state s_t, and a task instruction ℓ that is fixed within an episode. The policy $\pi_\theta(a_{t:t+H} \mid o_t)$, parameterized by θ, predicts a sequence of actions $a_{t:t+H} = (a_t, a_{t+1}, \ldots, a_{t+H-1})$ given o_t, where a_t is the action at timestep t and H is the chunk length. Let v_{src} and v_{tgt} denote the source and target third-person camera viewpoints, respectively. We finetune π_θ on a dataset $\mathcal{D}_{\text{src}}$ of demonstrations collected at v_{src} and evaluate the resulting policy at v_{tgt}. The wrist-camera setup, task, embodiment, and action coordinates remain unchanged between viewpoints.

3 Motivating Analysis

Wrist cameras observe the local manipulation region, while third-person cameras can cover parts of the scene outside the wrist camera's field of view (Hsu et al., 2022). We examine how viewpoint augmentation affects a policy's use of these complementary observations.

Setup. We use VISTA (Tian et al., 2024) as an example of learned novel-view synthesis. On LIBERO (Liu et al., 2023), we compare three $\pi_{0.5}$ policies (Physical Intelligence et al., 2025): the released LIBERO checkpoint without additional finetuning (Pretrained), the same checkpoint finetuned on the original source-view demonstrations (No augmentation), and its counterpart finetuned with VISTA augmentation (VISTA). The two finetuned policies use the same source demonstrations and policy-update budget (Section 5.1). VISTA varies third-person images while preserving paired wrist observations and action labels. The analysis uses the source viewpoint v_{src}, with no viewpoint change at evaluation time.

Attention allocation across cameras. To examine visual attention during action prediction, we measure attention from action tokens to each camera's image tokens using identical source-view observations for all three policies (Figure 2). For each observation, we define the wrist-image attention share as $r = A_{\text{wrist}}/(A_{\text{third}} + A_{\text{wrist}})$, where A_{wrist} and A_{third} denote the attention mass from action tokens to wrist-image tokens and third-person image tokens, respectively. Measurement and aggregation details are provided in Appendix E.1. The wrist-image attention share is 74.9% for Pretrained, 74.2% for No augmentation, and 80.8% for VISTA. Thus, VISTA increases the share by 5.9 percentage points relative to Pretrained and by 6.6 percentage points relative to No augmentation.

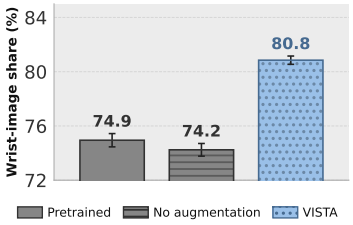

Figure 2: Wrist-image attention share on LIBERO. VISTA increases the share of attention to wrist images.

Hypotheses and research question. The observed attention shift suggests that viewpoint augmentation may change how the policy uses the two camera streams. First, we hypothesize that wrist-grounded control contributes to the robustness gains from viewpoint augmentation. Varying third-person images while preserving paired wrist observations may encourage the policy to use the wrist view as a stable basis for control. Other forms of third-person image augmentation may also encourage greater wrist reliance. Exclusive wrist reliance can be a shortcut (Huang et al., 2026a;b), but greater wrist reliance can still be useful if the policy also uses third-person information.

Second, we hypothesize that geometric variation in third-person images improves viewpoint robustness when stable wrist observations are available. We also examine whether the resulting policy continues to benefit from third-person information.

VISTA improves viewpoint robustness despite imperfections in its synthesized views (Appendix B, Figure 8). These observations and hypotheses motivate us to ask whether simple geometric augmentation of third-person images, while preserving paired wrist observations, can improve viewpoint robustness without faithful novel-view synthesis. We introduce SHARP in Section 4 and evaluate both hypotheses in the subsequent experiments.

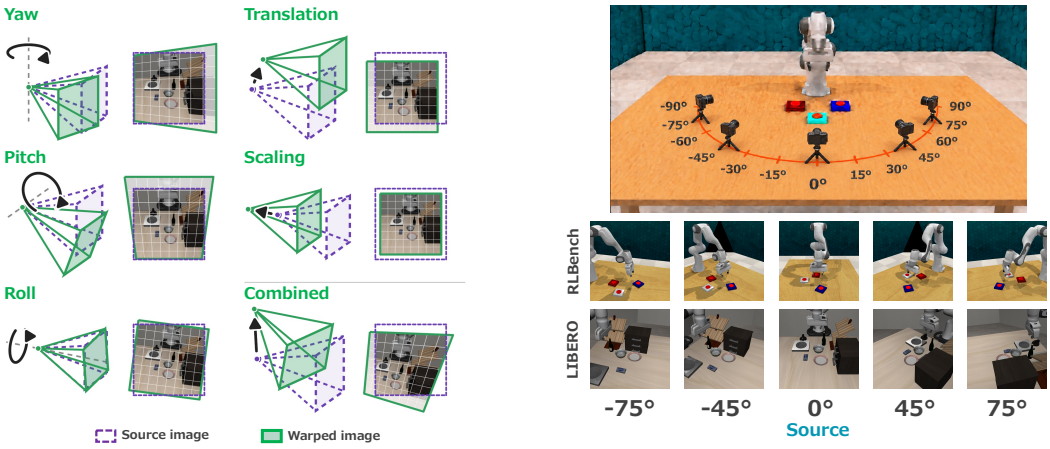

Figure 3: **(Left)** Individual homography components and their combined effect on a source image. Purple dashed and green solid outlines indicate source and transformed configurations. **(Right)** Evaluation camera positions and example observations on RLBench and LIBERO. The source view is at 0°, and the remaining viewpoints are target views.

4 METHOD

SHARP improves viewpoint robustness by applying random homography transformations to third-person images during policy finetuning. Each training step uses freshly sampled transformations while preserving the paired wrist observations and action labels. The method retains the policy's original training objective and requires no changes to its architecture or inference procedure.

4.1 HOMOGRAPHY TRANSFORMATION

We use homographies as flexible, inexpensive 2D transformations to approximate image changes associated with camera viewpoint shifts. A homography is a 2D projective transformation that can combine image translation, scaling, rotation, and perspective distortion (Hartley & Zisserman, 2004). We parameterize the transformation using a virtual camera and a fronto-parallel reference plane. The parameters $\omega = (R, \mathbf{t})$ consist of a rotation matrix R and a translation vector $\mathbf{t} = (t_x, t_y, t_z)^\top$. The rotation matrix R is parameterized by roll, pitch, and yaw, corresponding to rotations about the virtual camera's optical, horizontal, and vertical axes, respectively. Roll rotates the image in its plane, while yaw and pitch introduce perspective distortions. For the translation vector $\mathbf{t}$, we use the lateral components (t_x, t_y) to control image displacement and the depth component t_z to vary image scale. Given a virtual intrinsic matrix K, a plane normal $n = (0, 0, 1)^\top$, and a normalized plane distance $d = 1$, we construct the homography

$$H_\omega = C_R K \left(R + \frac{\mathbf{t} n^\top}{d} \right) K^{-1}, \tag{1}$$

where C_R is an image-plane translation that compensates for the image-center displacement induced by rotation alone, while preserving the effect of the virtual-camera translation $\mathbf{t}$. The virtual camera and reference plane define the augmentation without requiring estimates of scene depth or the physical camera pose. We apply H_ω to an image I through inverse mapping with bilinear interpolation. Pixels mapped outside the source image take the value of the nearest boundary pixel. Figure 3 **(Left)** illustrates yaw, pitch, roll, image translation, and scaling, as well as their combined effect. Additional examples appear in Appendix B (Figure 8).

4.2 POLICY FINETUNING WITH AUGMENTATION

SHARP finetunes a pretrained policy by applying random homographies to source-view third-person images. At each training step, we independently sample $\omega \sim p_{\text{aug}}$ for each third-person image in

the batch. We sample yaw, pitch, and roll from predefined ranges. Translation and scaling are jointly determined by a virtual-camera translation vector sampled with a random direction and magnitude. We chose the default parameter ranges by visually inspecting the warped images to avoid distortions that make the scene unrecognizable. Appendix A.2 provides the sampling distributions, numerical ranges, and implementation details.

For a source observation o_t, let $\widetilde{I}_t^{\text{third}}$ denote the image obtained by applying H_ω to I_t^{third} as described in Section 4.1. The augmented observation is

$$\mathcal{H}_\omega(o_t) = \big(\widetilde{I}_t^{\text{third}}, I_t^{\text{wrist}}, s_t, \ell\big). \tag{2}$$

Only the third-person image is transformed. The paired wrist image, proprioceptive state, and instruction remain unchanged, and the augmented observation is paired with the original action labels. SHARP retains the original training objective of the policy. For $\pi_{0.5}$ (Physical Intelligence et al., 2025), we use its conditional flow-matching loss. Let $\mathcal{L}_{\text{FM}}(\theta, o_t, a_{t:t+H})$ denote this loss for an observation and its corresponding action chunk, averaged over sampled noise and flow timesteps. The finetuning objective is

$$\mathcal{L}_s(\theta) = \mathbb{E}_{(o_t, a_{t:t+H})\sim\mathcal{D}_{\text{src}},\, \omega\sim p_{\text{aug}}}[\mathcal{L}_{\text{FM}}(\theta, \mathcal{H}_\omega(o_t), a_{t:t+H})]. \tag{3}$$

In this work, we use low-rank adaptation (LoRA) (Hu et al., 2021) in the vision encoder, language backbone, and action expert while keeping the pretrained weights frozen.

5 Experiments

The main text focuses on viewpoint generalization without target-view demonstrations. Additional experiments extending SHARP to one-shot viewpoint adaptation are presented in Appendix F.

5.1 Experimental Setup

Benchmarks. We evaluate viewpoint generalization on two simulation benchmarks, RLBench (James et al., 2020) and LIBERO (Liu et al., 2023). We use a subset of 12 RLBench tasks, listed in Appendix A.1. For LIBERO, we evaluate on four suites totaling 40 tasks.

Policy backbones. We use $\pi_{0.5}$ (Physical Intelligence et al., 2025) for our main experiments, initialized from an RLBench-finetuned base model or the public LIBERO checkpoint. To assess whether SHARP's benefits extend beyond $\pi_{0.5}$, we additionally evaluate it with OpenVLA-OFT (Kim et al., 2025) and X-VLA (Zheng et al., 2026) on LIBERO using their public checkpoints. All three backbones receive third-person and wrist-camera images and a language instruction. $\pi_{0.5}$ uses proprioceptive inputs on RLBench but not on LIBERO. OpenVLA-OFT and X-VLA use robot state inputs on LIBERO.

Compared methods. We compare SHARP with source-view finetuning and with learned view-synthesis methods. No augmentation finetunes the policy on the source-view demonstrations without viewpoint augmentation. VISTA (Tian et al., 2024) synthesizes augmented views using ZeroNVS (Sargent et al., 2024), either with its released checkpoint (VISTA) or after benchmark-specific finetuning (VISTA-FT). GenSplat (Wang et al., 2026a) uses feed-forward 3D Gaussian reconstruction and renders additional views from sampled camera poses. We include it as a reference with additional input views, using two simulator-rendered images at azimuth offsets of $\pm 40^\circ$ from v_{src}. SHARP and both VISTA variants use a single source view as input for augmentation. Details of view-synthesis model training and image generation are provided in Appendix B.

Training protocol. No augmentation, SHARP, and both VISTA variants use the same source demonstration dataset $\mathcal{D}_{\text{src}}$ within each benchmark. On RLBench, this dataset contains 100 demonstrations per task and was also used to obtain the initial checkpoint. Because object configurations vary substantially across evaluation episodes in RLBench, we retain the full dataset to preserve the diversity of object configurations during finetuning. On LIBERO, $\mathcal{D}_{\text{src}}$ contains one demonstration per task. GenSplat uses the same number of demonstrations, with the additional input views described above. Within each benchmark and backbone, all finetuned methods start from the same policy checkpoint and use 1,000 finetuning updates. For the main $\pi_{0.5}$ comparison on RLBench and LIBERO, we repeat policy finetuning with three random seeds. The additional backbones, ablations, and camera analyses use one training seed per condition. Optimizer and adapter settings are provided in Appendix A.3.1 for $\pi_{0.5}$ and Appendix A.3.2 for OpenVLA-OFT and X-VLA.

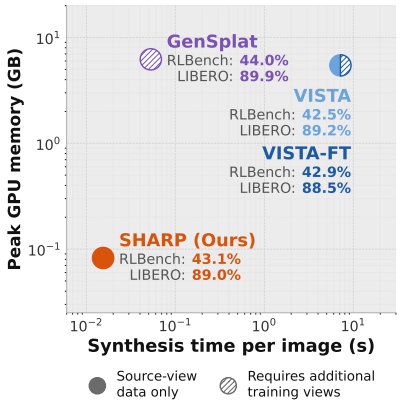

Figure 4: Augmentation cost and success.

Table 1: Effect of SHARP across three policy backbones on LIBERO. SHARP improves success over No augmentation at all four viewpoints shown for each evaluated backbone.

Model	Method	−75°	−45°	+45°	+75°	Avg.
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	No aug.	56.2	77.1	80.0	50.7	66.0
	SHARP	87.4	90.3	88.2	85.7	**87.9**
OpenVLA-OFT (Kim et al., 2025)	No aug.	25.9	55.2	61.6	38.1	45.2
	SHARP	70.0	76.9	79.8	75.3	**75.5**
X-VLA (Zheng et al., 2026)	No aug.	5.0	12.2	8.7	4.4	7.6
	SHARP	68.3	75.4	76.1	69.7	**72.4**

Evaluation protocol. For the main $\pi_{0.5}$ comparison, we evaluate all tasks in each benchmark at the source viewpoint (0°) and at angular offsets of ±15° to ±90° in 15° increments (Figure 3; **Right**). The wrist-camera mounting remains unchanged. All experiments use the full evaluation task set of each benchmark, with 50 episodes per task and viewpoint. We report success rates averaged across tasks and across training seeds. In the main text, we report results at ±45° and ±75° as representative viewpoint shifts. Detailed results are provided in Appendix C.

5.2 Viewpoint Robustness

We evaluate SHARP's viewpoint robustness and augmentation cost, and examine whether its benefits extend across VLA backbones.

Performance and augmentation cost. Averaged across three training seeds, tasks, and all 13 evaluation viewpoints, SHARP achieves success rates of 43.1 ± 1.0% on RLBench and 89.0 ± 0.1% on LIBERO (mean ± standard error), competitive with VISTA and GenSplat at a substantially lower augmentation cost (Figure 4). SHARP generates one augmented image in 0.0154 s, making it approximately 3.4× faster than GenSplat and 460× faster than VISTA. Peak allocated GPU memory for augmentation is 0.082 GB, compared with 6.220 GB for GenSplat and 5.454 GB for VISTA.

The methods also differ in their input and model requirements. GenSplat uses two input views at azimuth offsets of ±40°, whereas SHARP and VISTA use a single source view. VISTA-FT and GenSplat additionally use multi-view rendered images for model finetuning. SHARP directly warps source-view images without a learned synthesis or reconstruction model. Its low computational cost allows a fresh transformation each time a frame is sampled during finetuning. SHARP outperforms a variant that reuses one fixed augmented image per source frame on both benchmarks, suggesting that this training diversity contributes to its performance (Appendix D.1). Additional evaluations under camera elevation angle and distance changes are provided in Appendix C.4.

Generalization across policy models. To assess whether the benefits of SHARP extend beyond $\pi_{0.5}$, we evaluate it on OpenVLA-OFT (Kim et al., 2025) and X-VLA (Zheng et al., 2026). For each model, we compare finetuning without augmentation and with SHARP using the same source demonstrations and training budget. We evaluate at six shifted viewpoints and report four (±45°, ±75°) here. As shown in Table 1, SHARP increases mean success from 45.2% to 75.5% on OpenVLA-OFT and from 7.6% to 72.4% on X-VLA. For reference, the comparison on $\pi_{0.5}$ yields an increase from 66.0% to 87.9%. SHARP improves success at all four viewpoints for all three models, supporting its effectiveness across the evaluated policy backbones.

5.3 What Makes the Augmentation Effective?

Having established SHARP's effectiveness across benchmarks and policy backbones, we next examine which aspects of augmentation contribute to its gains. Specifically, we analyze augmentation type, homography components, and augmentation strength using $\pi_{0.5}$ on LIBERO.

Table 2: Comparison of image augmentation strategies for viewpoint robustness on LIBERO. SHARP achieves the highest success rate at every viewpoint, supporting the benefit of more general geometric transformations. Color-W/S denote weak/strong color augmentation and H-/V-flip denote horizontal/vertical flipping.

Angle	No aug.	Crop	Resize	Color-W	Color-S	Noise	H-flip	V-flip	Masking	Dropout	**SHARP**
$-75°$	55.2	81.7	61.2	55.6	70.8	57.1	74.8	61.4	78.2	71.5	**87.0**
$-45°$	76.8	88.1	81.5	77.2	84.2	77.3	76.7	78.5	87.9	82.6	**91.0**
$+45°$	79.4	86.7	82.9	80.1	83.2	80.5	77.1	81.2	86.0	82.7	**88.0**
$+75°$	51.4	77.1	60.6	51.9	63.5	52.5	63.6	61.3	77.6	65.2	**86.9**
Avg.	65.7	83.4	71.5	66.2	75.4	66.8	73.0	70.6	82.4	75.5	**88.2**

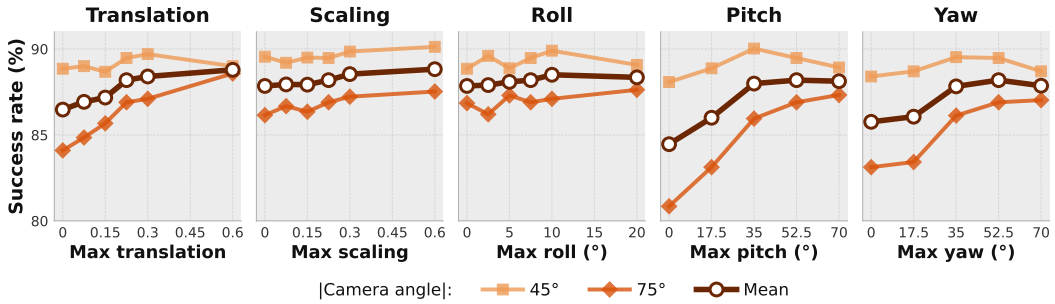

Figure 6: Effect of homography component strength on LIBERO viewpoint robustness. Stronger translation, pitch, and yaw generally help most at $\pm 75°$, while scaling and roll have smaller effects over the tested ranges.

Comparison with other image augmentations. We compare SHARP with alternative third-person image augmentations to examine the hypothesized benefit of geometric variation in policies using both camera streams (Section 3). The controls include crop, resize, color augmentation, noise, horizontal and vertical flips, random masking, and image dropout. All methods use the same source demonstrations, pretrained checkpoint, and training budget, modifying only third-person images while preserving wrist observations and action labels. We evaluate success at four viewpoint shifts ($\pm 45°$ and $\pm 75°$). Augmentation details and examples appear in Appendix E.3.

Table 2 shows that SHARP achieves the highest mean success rate of 88.2%, compared with 83.4% for crop, 82.4% for random masking, and 65.7% without augmentation. Crop and resize also improve viewpoint robustness, indicating that changes in image position and scale already provide useful training variation. SHARP achieves higher success, consistent with a benefit from combining a broader range of geometric transformations.

Contribution of transformation components. We next examine which homography components contribute to viewpoint robustness under the chosen sampling ranges by removing translation, scaling, roll, pitch, or yaw one at a time. Figure 5 shows success-rate changes relative to full SHARP, averaged over positive and negative viewpoint offsets of the same magnitude. Removing individual components has little effect at small viewpoint shifts. At $\pm 90°$, however, removing pitch, yaw, or translation reduces success by 9.0, 7.0, or 5.0 percentage points, respectively. Removing roll or scaling changes success by at most 1.4 percentage points across the evaluated viewpoints.

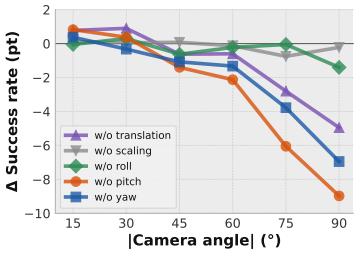

Figure 5: Ablation of SHARP's homography components.

Effect of augmentation strength. To assess how augmentation strength affects viewpoint robustness, we vary the maximum sampled magnitude of each homography component. Increasing translation, pitch, and yaw generally improves success more at $\pm 75°$ than at $\pm 45°$, while scaling and roll have smaller effects over the tested ranges (Figure 6). An additional sweep jointly scaling rotation and translation shows similar benefits at large viewpoint shifts across all four LIBERO suites (Appendix D.3).

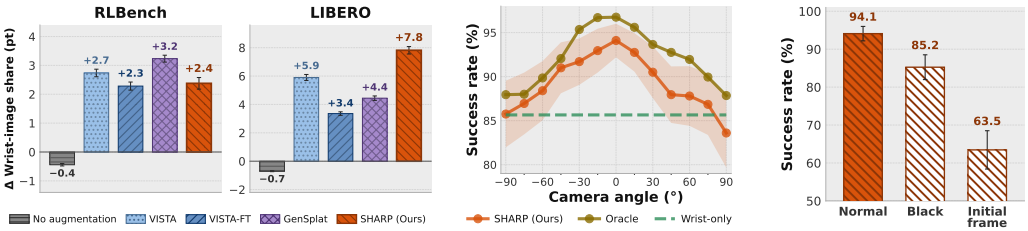

Figure 7: **(Left)** Changes in wrist-image attention share from Pretrained on RLBench and LIBERO. **(Middle)** LIBERO viewpoint robustness of SHARP, a wrist-only policy, and a viewpoint-specific oracle. **(Right)** SHARP's LIBERO success rate with Normal, Black, or Initial frame third-person inputs.

5.4 The Role of Third-Person and Wrist Observations

Building on the analysis in Section 3, we examine how SHARP uses the two camera streams and whether third-person information remains useful.

Attention allocation across cameras. We extend the attention analysis in Section 3 to SHARP on RLBench and LIBERO, using identical source-view observations for all checkpoints (Figure 7; **Left**). We use the wrist-image attention share defined in Section 3 and the aggregation procedure detailed in Appendix E.1. Here, we report the change from Pretrained, $100(r - r_{\text{pretrained}})$, in percentage points. Like the other viewpoint-augmentation methods, SHARP increases the wrist-image attention share on both benchmarks, by 2.4 percentage points on RLBench and 7.8 percentage points on LIBERO relative to Pretrained. In contrast, No augmentation decreases this share by 0.4 and 0.7 percentage points, respectively. This shared pattern suggests that SHARP encourages a shift toward more wrist-grounded control.

Comparison with a wrist-only policy. We finetune a wrist-only policy using the same pretrained checkpoint, 40 source demonstrations, and 1,000 updates, with third-person images blacked out during both training and evaluation. This policy achieves 85.7% mean success across the 40 tasks. Since its inputs are unaffected by changes to the third-person camera, Figure 7 **(Middle)** shows its single evaluation as a horizontal reference. For context, we also include a viewpoint-specific oracle, finetuned separately on simulator-rendered demonstrations from each evaluation angle and evaluated at that same angle. SHARP achieves 88.0–93.0% success at $\pm 15°$, $\pm 30°$, and $\pm 45°$, exceeding the wrist-only reference. The advantage narrows at larger shifts. At $-90°$ and $+90°$, SHARP achieves 85.8% and 83.6%, respectively. The oracle also declines from 96.8% at $0°$ to 88.0% and 87.9% at $-90°$ and $+90°$, respectively, indicating that extreme viewpoints remain challenging even with target-view supervision. SHARP's advantage over the wrist-only policy at small to moderate viewpoint shifts suggests that it benefits from third-person observations in addition to the wrist view.

Third-person input interventions. To further assess whether SHARP benefits from current third-person observations, we compare Normal, Black, and Initial frame inputs on LIBERO at the source viewpoint ($0°$). Normal uses the current third-person image, Black replaces it with zeros, and Initial frame reuses the first policy-query image throughout the episode. The SHARP policy checkpoint is fixed, and all conditions retain current wrist observations. Figure 7 **(Right)** shows that replacing Normal inputs with Black or Initial frame inputs reduces mean success by 8.9 and 30.6 percentage points, respectively, from 94.1%. These results are consistent with SHARP using third-person information alongside wrist observations.

6 Discussion and Limitations

We introduced Simple Homography Augmentation for Robust Policies (SHARP), a simple 2D augmentation strategy that improves viewpoint robustness of VLAs. We discuss how the results relate to our hypotheses and identify remaining limitations and open questions.

Interpreting the gains from viewpoint augmentation. Our camera-input analyses suggest that SHARP encourages wrist-grounded control while retaining useful third-person information under certain viewpoint shifts (Section 5.4). Its gains over alternative augmentations further support the benefit of geometric augmentation when paired wrist observations and action labels are preserved (Section 5.3). Without wrist observations, SHARP and VISTA-FT fall below No augmentation on

LIBERO, while GenSplat provides only a small gain (Appendix D.4). Together, these findings are consistent with our hypotheses that SHARP encourages wrist-grounded control and that geometric augmentation plays an important role in improving viewpoint robustness when stable wrist observations are available. Extending these benefits to policies using only third-person images remains an open challenge.

Generalization to other camera configurations. We vary the third-person viewpoint while keeping the wrist-camera setup fixed. Our analyses suggest that stable wrist observations contribute to the gains, so changes in wrist-camera placement or field of view could affect how effectively the policy combines the two streams. Robustness to changes in wrist-camera placement or field of view, and to other multi-camera setups, remains unverified.

Applicability to real-world settings. Our evaluation is limited to simulation. Real-world observations can differ in lighting and textures, which may affect how policies respond to geometric augmentation. Physical-robot experiments are therefore needed to assess whether SHARP's gains in viewpoint robustness extend to real-world settings.

7 Related Work

Viewpoint robustness in visuomotor policies. Prior work incorporates camera geometry (Jiang et al., 2025) or pretrained geometric features (Abouzeid et al., 2025) into policies. Paired-view objectives encourage camera-frame action equivariance (Zhang et al., 2026) or action-flow consistency (Huang et al., 2026b). Deployment-time approaches adapt visual encoders through frozen dynamics models (Yang et al., 2023) or synthesize training-view observations for frozen policies (Heo et al., 2026; Park & Kim, 2026). Domain updates can also be extracted from models finetuned on source- and target-domain demonstrations of the same task (Kang et al., 2026). For combined wrist and third-person observations, an information bottleneck on the third-person representation improves generalization (Hsu et al., 2022). SHARP addresses viewpoint robustness through homography augmentation during VLA finetuning. It uses demonstrations from a single third-person viewpoint with paired wrist observations and requires neither scene geometry estimation nor additional cross-view objectives.

Data augmentation for visual policy learning. Simple image transformations improve visual reinforcement learning (Laskin et al., 2020; Kostrikov et al., 2021), while expert-student training enables stronger visual perturbations (Fan et al., 2021). Generative augmentation modifies objects and scene appearance (Chen et al., 2023; Yu et al., 2023). NeRF-based augmentation pairs perturbed eye-in-hand views with corrective actions (Zhou et al., 2023). Other approaches augment viewpoints through learned view synthesis (Tian et al., 2024; Chen et al., 2024), 3D Gaussian rendering (Yang et al., 2025; Wang et al., 2026a), or depth-based rendering with diffusion-based video completion (Cai et al., 2026). Under source-view evaluation, conventional 2D augmentations have been reported to often underperform a single-view baseline and fall short of multi-view training (Cai et al., 2026). We show that homography augmentation of third-person images can improve unseen-view robustness in VLAs while retaining paired wrist observations and action labels. SHARP achieves competitive viewpoint robustness at substantially lower augmentation cost than the evaluated view-synthesis methods, without requiring learned synthesis or scene reconstruction.

8 Conclusion

We introduced SHARP, a lightweight homography augmentation method for VLAs using both wrist and third-person images. It augments third-person images during finetuning without target-view demonstrations or a learned view-synthesis model. On RLBench and LIBERO, SHARP achieves success rates competitive with the evaluated view-synthesis methods at substantially lower augmentation cost. Its robustness gains extend across three VLA backbones on LIBERO. Comparisons with alternative 2D augmentations and component ablations support the benefit of geometric augmentation for viewpoint robustness. Our camera-input analyses suggest greater wrist reliance while retaining useful third-person information under certain viewpoint shifts. These results suggest that lightweight geometric augmentation can improve viewpoint robustness in VLAs with stable wrist-camera inputs, even without faithfully reconstructing novel views.

AI USE STATEMENT

We used generative AI tools to assist with implementing methods and translating text. We did not use these tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Generating synthetic datasets and developing or writing mathematical proofs, including providing key ingredients for proofs, were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, draft portions of the manuscript, and improve its readability. The authors reviewed and tested AI-assisted code and reviewed the content of AI-assisted translations and manuscript text. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This study concerns robotic manipulation under visual domain shifts. Failures in real-world deployment could cause harm to people or property, and our results do not establish safety in such settings. Deployment around people or in unconstrained environments would require task-specific safety validation beyond what we report here.

REPRODUCIBILITY STATEMENT

Section 4 describes SHARP, and Section 5.1 specifies the training and evaluation protocols. Appendix A provides benchmark details, homography sampling procedures, model-specific settings, and computational cost measurement protocols. Appendix B documents the view-synthesis baselines, and Appendix E describes the attention measurements and camera-input and augmentation controls.

REFERENCES

Ali Abouzeid, Malak Mansour, Qinbo Sun, Zezhou Sun, and Dezhen Song. GeoAware-VLA: Implicit geometry aware vision-language-action model. *arXiv preprint arXiv:2509.14117*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Boyang Cai, Qiwei Liang, Jiawei Li, Shihang Weng, Zhaoxin Zhang, Tao Lin, Xiangyu Chen, Wenjie Zhang, Jiaqi Mao, Weisheng Xu, et al. Beyond viewpoint generalization: What multi-view demonstrations offer and how to synthesize them for robot manipulation? *arXiv preprint arXiv:2603.26757*, 2026.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. RoVi-Aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. GenAug: Retargeting behaviors to unseen situations via generative augmentation, 2023. URL https://arxiv.org/abs/2302.06671.

Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Animashree Anandkumar. SECANT: Self-expert cloning for zero-shot generalization of visual policies. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3088–3099. PMLR, 2021. URL https://proceedings.mlr.press/v139/fan21c.html.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

Hyeongjun Heo, Seungyeon Woo, Sang Min Kim, Junho Kim, Junho Lee, Yonghyeon Lee, and Young Min Kim. AnyCamVLA: Zero-shot camera adaptation for viewpoint robust vision-language-action models. *arXiv preprint arXiv:2603.05868*, 2026.

Kyle Hsu, Moo Jin Kim, Rafael Rafailov, Jiajun Wu, and Chelsea Finn. Vision-based manipulators need to also see from their hands, 2022. URL https://arxiv.org/abs/2203.12677.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Bingqi Huang, Bingchuan Wei, Yingkai Cai, and Zhaokui Wang. Selective cross-view consistency for world action models: Held-out viewpoint robustness without test-time camera information. *arXiv preprint arXiv:2608.21402*, 2026a.

Bingqi Huang, Bingchuan Wei, Xuan Wang, Yingkai Cai, and Zhaokui Wang. Cross-view action consistency for camera-robust vision-language-action policies. *arXiv preprint arXiv:2608.06965*, 2026b.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Tianchong Jiang, Jingtian Ji, Xiangshan Tan, Jiading Fang, Anand Bhattad, Vitor Guizilini, and Matthew R Walter. Do you know where your camera is? View-invariant policy learning with camera conditioning. *arXiv preprint arXiv:2510.02268*, 2025.

Taewook Kang, Taeheon Kim, Donghyun Shin, and Jonghyun Choi. Domain arithmetic: One-shot VLA adaptation under environmental shifts. *arXiv preprint arXiv:2607.00666*, 2026.

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

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels, 2021. URL https://arxiv.org/abs/2004.13649.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19884–19895. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e615c82aba461681ade82da2da38004a-Paper.pdf.

Weiqi Li, Quande Zhang, Ruifeng Zhai, Liang Lin, and Guangrun Wang. VLA models are more generalizable than you think: Revisiting physical and spatial modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35025–35035, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Yechan Park and HyunJin Kim. GS-VLA: Plug-and-play viewpoint canonicalization for frozen VLA policies via Gaussian splatting. *arXiv preprint arXiv:2608.19066*, 2026.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9420–9429. IEEE, 2024.

Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis. *arXiv preprint arXiv:2409.03685*, 2024.

Sen Wang, Huaiyi Dong, Jingyi Tian, Jiayi Li, Zhuo Yang, Tongtong Cao, Anlin Chen, Shuang Wu, Le Wang, and Sanping Zhou. Efficient camera pose augmentation for view generalization in robotic policy learning. *arXiv preprint arXiv:2603.29192*, 2026a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

Yajat Yadav, Zhiyuan Zhou, Andrew Wagenmaker, Karl Pertsch, and Sergey Levine. Robust fine-tuning of vision-language-action robot policies via parameter merging. In *International Conference on Learning Representations*, volume 2026, pp. 56385–56417, 2026.

Sizhe Yang, Yanjie Ze, and Huazhe Xu. MoVie: Visual model-based policy adaptation for view generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0940. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43b77cef2a83a25aa27d3271d209e4fd-Abstract-Conference.html.

Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with Gaussian splatting enables robust one-shot manipulation. *arXiv preprint arXiv:2504.13175*, 2025.

Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Dee M, Jodilyn Peralta, Brian Ichter, Karol Hausman, and Fei Xia. Scaling robot learning with semantically imagined experience, 2023. URL https://arxiv.org/abs/2302.11550.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595. IEEE, 2018.

Tianyi Zhang, Ziyang Gong, Zhenjie Yang, Zhe Qian, and Haonan Duan. OC-VLA++: Monocular geometry-guided cross-view consistency for viewpoint-robust robotic manipulation. *arXiv preprint arXiv:2608.01066*, 2026.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. NeRF in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. *arXiv preprint arXiv:2301.08556*, 2023. URL https://arxiv.org/abs/2301.08556.

A Implementation and Evaluation Details

A.1 Benchmarks and Pretrained Policies

Our RLBench subset contains 12 tasks: *close box*, *close laptop lid*, *meat on grill*, *open box*, *push buttons*, *stack wine*, *take lid off saucepan*, *toilet seat down*, *sweep to dustpan*, *close jar*, *open drawer*, and *slide block to target*. The pretrained RLBench policy is obtained by finetuning the $\pi_{0.5}$ base model on 100 demonstrations per task for 30,000 steps with a global batch size of 256. On LIBERO, we initialize $\pi_{0.5}$ from its publicly released LIBERO checkpoint.

A.2 Homography Sampling

We sample a virtual-camera rotation and translation to construct the image warp in Equation 1. The coordinate mappings underlying cropping and resizing are special cases of homographies.

Coordinate and sign conventions. Let $\mathbf{X}_{\text{src}}$ and $\mathbf{X}_{\text{dst}}$ denote the coordinates of the same 3D point in the source and destination virtual camera frames, respectively. We define

$$\mathbf{X}_{\text{dst}} = R\mathbf{X}_{\text{src}} + \mathbf{t},$$

where R maps source-frame coordinates into the destination frame and $\mathbf{t}$ is expressed in the destination frame. The destination camera center, expressed in the source frame, is $-R^\top \mathbf{t}$. The reference plane is defined in the source frame by $n^\top \mathbf{X}_{\text{src}} = d$, with $n = (0, 0, 1)^\top$ and $d = 1$. For points on this plane, the coordinate transformation becomes $\mathbf{X}_{\text{dst}} = (R + \mathbf{t}n^\top/d)\mathbf{X}_{\text{src}}$, which gives the positive sign in Equation 1. Both virtual cameras use the same intrinsic matrix K, and C_R applies the image-plane centering correction described below. The resulting H_ω maps source-image coordinates to augmented-image coordinates; image resampling uses its inverse.

Virtual camera and rotation. The virtual camera has a $45°$ field of view and its principal point at the image center, giving focal length $f = W/[2\tan(22.5°)]$ for image width W. Yaw and pitch are sampled independently. Each has a magnitude uniformly drawn from $[0°, 52.5°]$ and an equally likely positive or negative sign. Roll is sampled uniformly from $[-7.5°, 7.5°]$. We combine these rotations as $R = R_z(\text{roll})R_x(\text{pitch})R_y(\text{yaw})$.

Translation and scaling. We sample a direction uniformly on the unit sphere and a normalized magnitude $\|\mathbf{t}\|/d \sim U(0, 0.225)$ to form the translation vector $\mathbf{t}$, where $U(a, b)$ denotes the uniform distribution on $[a, b]$. Its lateral components (t_x, t_y) shift the image, while its depth component t_z controls scaling. These effects therefore come from one sampled vector, rather than independent translation and scaling parameters. Translation is measured relative to the virtual plane distance $d = 1$.

Centering and image warping. Rotation alone can move the image center. We compute this displacement using KRK^{-1} and apply the image-plane correction C_R to bring the center back to its original position. This correction is computed before adding $\mathbf{t}$, so it preserves the sampled translation and scaling. We apply the resulting homography using bilinear interpolation at 224×224 resolution for $\pi_{0.5}$ and 256×256 for OpenVLA-OFT and X-VLA. OpenVLA-OFT then resizes the warped image to 224×224 as part of its input preprocessing. Pixels sampled outside the image take the nearest boundary pixel value.

A.3 Model-Specific Policy Settings

A.3.1 $\pi_{0.5}$ Settings

On both RLBench and LIBERO, SHARP and VISTA use the same optimization settings, with 1,000 updates, batch size 64, and LoRA with rank and scaling parameter 32, training adapters in the vision encoder, language backbone, and action expert while freezing the original weights. We use AdamW as the optimizer with a constant learning rate of 10^{-4} and weight decay of 10^{-2}. The action horizon is 10, and we sample the flow timestep as $\tau = 0.999b + 0.001$, $b \sim \text{Beta}(1.5, 1)$, where $\text{Beta}(\alpha, \beta)$ denotes the beta distribution with shape parameters α and β.

A.3.2 OpenVLA-OFT and X-VLA Settings

Models and training. We initialize OpenVLA-OFT (Kim et al., 2025) from `moojink/openvla-7b-oft-finetuned-libero-spatial-object-goal-10` and X-VLA (Zheng et al., 2026) from `2toINF/X-VLA-Libero`. Each adapted policy is trained for 1,000 updates on one source-view demonstration per task across the 40 LIBERO tasks. Both models use an effective batch size of 64.

OpenVLA-OFT uses rank-32 LoRA on all linear layers with zero adapter dropout, while also training the action head and proprioceptive projector. Its learning rate is constant at 5×10^{-4}. X-VLA uses rank-8 LoRA with scaling parameter 16 on all linear layers, while also training the soft prompt, action encoder, and action decoder. Its learning rate warms up over 100 steps to 10^{-4} and then follows a cosine schedule to 10^{-5}. These settings are shared by SHARP, VISTA, and GenSplat within each model.

Inputs and control. Both models receive third-person and wrist images, a task instruction, and proprioceptive state. OpenVLA-OFT uses 224×224 images and predicts chunks of eight relative end-effector actions, executing the full chunk before replanning. X-VLA uses 256×256 images and absolute end-effector pose control. Source demonstration actions are converted from relative commands to absolute controller targets. It predicts chunks of 30 actions with 10 denoising steps and executes the full chunk before replanning.

Evaluation. We evaluate No augmentation, SHARP, VISTA, and GenSplat policies on all four LIBERO suites at $\pm 15°$, $\pm 45°$, and $\pm 75°$. Table 1 reports success averaged across suites at four of these viewpoints ($\pm 45°$ and $\pm 75°$). Its final column averages those four viewpoints. Tables 8 and 9 provide all six viewpoints, with six-view averages.

A.4 Computational Cost Measurement

We measure image-synthesis cost on an NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU at batch size one (Table 3). Each method processes the same 320 frames sampled across tasks and timesteps from the 40 LIBERO source demonstrations, twice per frame. We report mean runtime and measured peak allocated GPU memory.

Timing includes GPU synchronization, CPU–GPU transfers, and VISTA's LPIPS scoring (Zhang et al., 2018), but excludes model loading and compilation, file I/O, training, evaluation, retries, and input-view acquisition. GenSplat's unused distillation teacher is excluded from memory measurements.

Table 3: Measured single-image synthesis cost on one Blackwell Max-Q GPU, corresponding to Figure 4. Memory is peak allocated GPU memory in GB. SHARP achieves the lowest synthesis time and peak GPU memory usage among the compared methods.

Method	Mean synthesis time (s/image)	Peak memory (GB)
GenSplat	0.0530	6.220
VISTA	7.1411	5.454
SHARP	0.0154	0.082

B View-Augmentation Baselines

B.1 VISTA

Pretrained ZeroNVS. VISTA (Tian et al., 2024) uses the released ZeroNVS (Sargent et al., 2024) checkpoint with no benchmark-specific finetuning. We retain the first generated sample without acceptance filtering, retries, or source-image fallbacks because the released generator is not adapted to the benchmark renderer.

ZeroNVS finetuning. For LIBERO, we use the 90 LIBERO-90 tasks, excluding the 40 policy-evaluation tasks. For RLBench, we collect data from the 94 tasks outside our 12-task evaluation set

and retain those that pass the trajectory and rendering checks. We select 16 states per task and render each state from eight viewpoints to construct the finetuning dataset. The ZeroNVS generators used for image generation are finetuned for 5,000 updates on LIBERO and 2,750 updates on RLBench.

Image generation. Both VISTA variants use the same viewpoint sampling and image-generation settings. For each source frame, we sample an azimuth offset uniformly from $[-90°, 90°]$ and generate an image with 250 DDIM steps, stochasticity parameter $\eta = 1$, and guidance scale 7.5. The calibrated scene scales are 0.8 for LIBERO and 1.3 for RLBench. The camera fields of view are $45°$ and $40°$, respectively. For VISTA-FT, we accept the first image whose AlexNet LPIPS distance (Zhang et al., 2018) to the source image is below 0.55, using the VISTA implementation's $[0, 1]$ input convention. If a sample is rejected, we keep the target pose fixed and change the diffusion seed, allowing at most five attempts in total. If all five fail, we retain the original source image. This yields one stored image per source frame, including fallbacks, while preserving its action, wrist image, state, and instruction.

B.2 GenSplat

Training. We train separate GenSplatPi3 generators for LIBERO and RLBench using the same tasks outside the policy-evaluation set as ZeroNVS. Each training sample contains two images of the same simulator state, captured at azimuth offsets of $\pm 40°$ from the source camera. Starting from pretrained Pi3 (Wang et al., 2026b), the model learns to represent the scene with 3D Gaussians that can be rendered from different viewpoints. Training encourages the rendered images to match the two input images and uses a frozen Pi3 model to guide the predicted geometry. We split training and validation data by trajectory state, train for 10,000 updates, and select the checkpoint with the lowest LPIPS (Zhang et al., 2018) on novel views of validation states.

Image generation. For each policy-training frame, GenSplat reconstructs the scene in a single forward pass from two additional views at $\pm 40°$, using the same input configuration as during generator training. Access to these simulator-rendered views of each policy-training state gives GenSplat an advantage in available visual information over SHARP and both VISTA variants, which use only the source-view image for augmentation. We use the known input-camera poses to map target camera poses into the reconstructed scene's coordinate system. For each frame, we sample an azimuth offset uniformly from $[-90°, 90°]$ relative to the source view and render an image from that viewpoint. The generated image replaces the source image while preserving the action, wrist image, state, and instruction. Figure 8 presents the qualitative comparison.

C Detailed Viewpoint Generalization Results

Table 5 and Table 7 report the full viewpoint sweeps underlying the mean success rates in Figure 4. The sweep contains 13 viewpoints from $-90°$ to $90°$ in $15°$ increments. The five finetuned methods in these RLBench and LIBERO angular sweeps are averaged over three training seeds. For each seed, we average success across tasks and all 13 viewpoints, then compute the standard error (SE). Across the five methods, these SEs range from 0.05 to 0.20 percentage points on LIBERO and from 0.65 to 2.03 on RLBench. For SHARP, the SEs are 0.12 and 0.98 percentage points, respectively, small relative to its mean gains over No augmentation (16.4 and 11.8 percentage points).

C.1 RLBench Per-Task and Task-Averaged Results

Table 4 and Table 5 report per-task and task-averaged results, respectively.

Performance at the source viewpoint. The performance decrease at the source viewpoint ($0°$) is not specific to SHARP. VISTA, VISTA-FT, and GenSplat also underperform No augmentation on several tasks. The three-seed task-averaged success rates are 62.5% for No augmentation, 57.1% for SHARP, 52.6% for VISTA, 55.5% for VISTA-FT, and 54.3% for GenSplat (Table 5). A potential remedy is to mix original source-view demonstrations with augmented demonstrations during finetuning, retaining direct supervision on the source-view observations, as done in prior work (Yu et al., 2023; Chen et al., 2024; Cai et al., 2026). We leave evaluating this approach to future work.

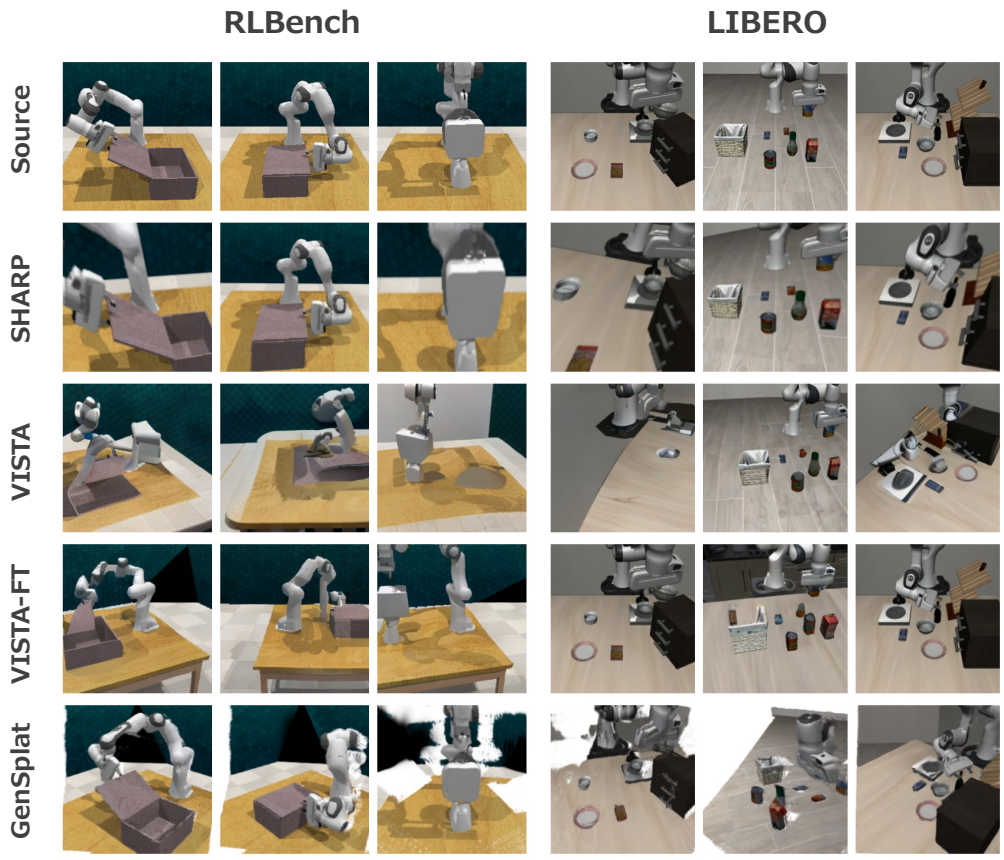

Figure 8: View augmentation on three RLBench examples (left) and three LIBERO examples (right). Each column uses a common source frame (top), followed by outputs from SHARP, VISTA, VISTA-FT, and GenSplat. VISTA and GenSplat produce images that partly resemble real viewpoint changes, whereas SHARP warps the source image in 2D and does not reproduce the full 3D appearance changes caused by moving the camera.

Table 4: Per-task success rates on RLBench (%) over the full viewpoint sweep. Values are means across three training seeds. Columns give the third-person camera angle relative to the source view (0°), with negative and positive angles denoting opposite azimuth directions. **Mean** equally averages all 13 viewpoints.

Method	−90°	−75°	−60°	−45°	−30°	−15°	0°	+15°	+30°	+45°	+60°	+75°	+90°	Mean
Close box														
No augmentation	34.7	38.0	40.7	62.0	68.0	88.7	88.0	54.0	32.0	50.0	48.7	56.0	51.3	54.8
VISTA-FT	59.3	58.7	57.3	48.7	59.3	74.7	80.0	73.3	68.0	60.0	60.7	68.7	64.7	64.1
VISTA	52.7	61.3	51.3	51.3	61.3	70.0	79.3	71.3	57.3	59.3	57.3	62.7	62.0	61.3
GenSplat	35.3	45.3	46.0	48.0	54.0	60.7	74.0	64.0	56.7	44.0	46.0	45.3	43.3	51.0
SHARP	57.3	71.3	67.3	70.7	74.0	90.0	89.3	68.7	58.7	62.0	60.7	64.7	60.7	68.9
Close laptop lid														
No augmentation	44.7	75.3	77.3	85.3	82.0	94.7	100.0	81.3	62.7	56.0	60.0	65.3	64.7	73.0
VISTA-FT	85.3	85.3	82.7	91.3	92.0	93.3	92.0	91.3	84.7	83.3	84.7	84.7	80.7	87.0
VISTA	82.0	83.3	84.0	88.0	92.0	90.0	87.3	89.3	83.3	82.0	84.7	82.7	86.7	85.8
GenSplat	81.3	82.0	86.7	90.7	90.7	89.3	95.3	94.0	88.0	80.7	84.0	82.7	78.0	86.4
SHARP	77.3	87.3	84.0	89.3	92.7	95.3	97.3	92.0	78.7	74.0	71.3	78.7	76.0	84.2
Meat on grill														
No augmentation	1.3	4.0	2.0	2.0	2.0	14.7	18.0	5.3	10.7	2.0	2.7	0.7	2.0	5.2
VISTA-FT	8.7	1.3	6.0	8.0	10.7	13.3	22.7	14.0	13.3	9.3	9.3	11.3	6.7	10.4
VISTA	4.0	8.0	12.0	9.3	8.7	16.7	18.0	10.7	10.0	8.0	6.7	9.3	12.7	10.3
GenSplat	6.7	4.7	7.3	7.3	8.7	11.3	14.0	7.3	10.7	7.3	8.0	8.7	10.7	8.7
SHARP	3.3	4.7	6.0	10.7	7.3	17.3	19.3	13.3	8.0	10.7	8.7	3.3	6.7	9.2
Open box														
No augmentation	12.7	4.0	5.3	10.7	10.0	14.7	32.0	19.3	15.3	14.0	14.7	18.7	12.0	14.1
VISTA-FT	19.3	16.0	19.3	15.3	24.0	22.7	38.7	29.3	23.3	18.7	13.3	19.3	12.7	20.9
VISTA	11.3	14.7	15.3	16.7	20.7	26.0	37.3	26.0	14.0	17.3	18.0	14.7	10.7	18.7
GenSplat	21.3	17.3	23.3	25.3	22.7	20.7	31.3	16.7	12.0	24.0	27.3	22.0	18.7	21.7
SHARP	16.7	14.0	18.0	18.0	12.0	23.3	30.7	26.7	13.3	16.7	21.3	25.3	24.0	20.0
Push buttons														
No augmentation	10.0	15.3	18.7	29.3	38.7	42.0	43.3	44.0	30.0	18.7	16.0	9.3	7.3	24.8
VISTA-FT	30.0	32.7	35.3	42.0	51.3	48.7	50.0	48.7	44.0	39.3	34.0	32.0	28.7	39.7
VISTA	32.7	40.7	40.7	46.0	53.3	48.0	52.0	49.3	44.0	37.3	39.3	36.0	34.0	42.6
GenSplat	32.7	40.0	42.0	46.7	44.7	51.3	48.0	42.7	49.3	41.3	35.3	37.3	32.0	41.8
SHARP	29.3	30.0	36.7	41.3	50.0	49.3	48.7	52.0	44.7	34.0	37.3	31.3	26.7	39.3
Stack wine														
No augmentation	2.7	4.0	3.3	7.3	7.3	18.7	67.3	12.7	12.7	12.7	5.3	4.0	4.0	12.5
VISTA-FT	10.7	9.3	11.3	12.7	18.0	34.0	47.3	28.7	22.7	22.0	21.3	16.0	22.7	21.3
VISTA	7.3	11.3	9.3	12.0	13.3	38.0	43.3	26.0	18.0	20.0	22.0	22.7	24.7	20.6
GenSplat	16.0	18.0	20.0	29.3	20.7	27.3	33.3	19.3	23.3	18.7	18.7	21.3	21.3	22.1
SHARP	8.7	14.7	19.3	17.3	16.7	33.3	44.7	6.7	12.0	11.3	10.0	11.3	10.7	16.7
Take lid off saucepan														
No augmentation	8.0	12.0	22.7	22.7	29.3	58.7	82.7	42.7	38.7	32.7	20.0	18.0	12.7	30.8
VISTA-FT	60.0	61.3	59.3	72.7	73.3	72.7	70.0	68.7	70.0	68.7	72.7	68.0	73.3	68.5
VISTA	57.3	56.7	63.3	72.7	76.0	76.7	69.3	62.0	66.0	70.0	64.7	70.0	70.7	67.3
GenSplat	67.3	65.3	71.3	68.7	70.0	70.7	78.0	80.0	76.7	72.7	70.7	71.3	66.7	71.5
SHARP	52.0	54.0	60.7	68.7	72.0	76.7	74.7	58.7	58.0	55.3	60.7	52.7	60.7	61.9
Toilet seat down														
No augmentation	62.0	68.7	65.3	64.0	66.7	80.7	82.7	82.0	76.7	70.0	64.7	66.7	56.0	69.7
VISTA-FT	85.3	85.3	82.7	82.7	88.0	85.3	87.3	84.7	86.0	83.3	88.0	80.0	86.0	85.0
VISTA	80.0	80.7	83.3	78.0	83.3	88.0	84.0	82.7	86.0	85.3	80.0	84.0	84.0	83.0
GenSplat	70.7	64.7	76.0	76.0	77.3	79.3	87.3	88.7	90.0	92.0	86.7	84.0	76.7	80.7
SHARP	86.7	85.3	79.3	76.0	86.7	84.0	84.0	81.3	84.0	87.3	91.3	86.7	82.7	84.3
Sweep to dustpan														
No augmentation	3.3	29.3	36.0	55.3	75.3	77.3	74.0	65.3	60.0	52.7	16.7	19.3	11.3	44.3
VISTA-FT	51.3	56.0	51.3	62.7	68.7	74.0	68.7	64.7	49.3	42.7	44.7	51.3	56.0	57.0
VISTA	55.3	56.0	54.7	64.7	68.7	72.7	64.0	64.7	52.0	50.7	52.7	55.3	53.3	58.8
GenSplat	66.0	72.7	78.0	86.0	79.3	83.3	91.3	87.3	78.7	74.0	64.0	72.0	63.3	76.6
SHARP	60.7	68.7	64.0	74.7	70.0	71.3	68.7	66.7	59.3	55.3	61.3	60.0	55.3	64.3
Close jar														
No augmentation	0.0	0.7	0.7	2.7	14.0	38.0	54.7	30.0	10.0	2.0	0.0	0.0	0.0	11.7
VISTA-FT	9.3	12.0	10.7	18.7	23.3	31.3	34.0	32.0	30.7	22.0	14.7	10.0	9.3	19.8
VISTA	14.7	16.7	16.0	18.7	24.0	29.3	22.7	24.0	26.0	22.7	18.0	12.7	10.7	19.7
GenSplat	17.3	20.0	24.7	23.3	25.3	32.7	30.0	29.3	24.0	26.0	23.3	16.0	16.7	23.7

continued on the next page

Table 4 (continued)

Method	−90°	−75°	−60°	−45°	−30°	−15°	0°	+15°	+30°	+45°	+60°	+75°	+90°	Mean
SHARP	14.7	20.7	15.3	18.7	32.0	38.7	40.0	32.0	26.7	21.3	10.7	6.7	10.0	22.1
Open drawer														
No augmentation	6.7	13.3	24.7	21.3	32.0	34.7	33.3	19.3	15.3	15.3	9.3	8.7	3.3	18.3
VISTA-FT	28.7	37.3	29.3	39.3	32.7	34.0	32.7	26.7	24.7	16.7	17.3	13.3	16.7	26.9
VISTA	36.7	41.3	34.0	34.7	26.0	27.3	38.7	28.0	18.7	20.0	26.7	16.7	14.7	27.9
GenSplat	24.7	28.0	24.0	23.3	22.0	22.0	24.0	23.3	14.7	18.0	17.3	20.7	27.3	22.3
SHARP	34.7	39.3	41.3	36.0	32.7	29.3	42.0	39.3	28.0	29.3	29.3	27.3	20.7	33.0
Slide block to target														
No augmentation	0.0	0.0	0.7	11.3	17.3	48.7	74.0	47.3	13.3	4.7	0.7	0.0	0.0	16.8
VISTA-FT	0.7	2.7	5.3	13.3	28.0	33.3	42.7	38.0	15.3	4.7	2.7	0.7	0.7	14.5
VISTA	4.0	3.3	5.3	12.0	22.0	33.3	34.7	31.3	16.7	10.7	3.3	0.7	2.0	13.8
GenSplat	2.7	8.7	12.0	14.0	30.0	42.0	45.3	45.3	38.0	23.3	9.3	6.0	4.7	21.6
SHARP	0.0	0.7	4.0	8.7	19.3	38.7	46.0	38.7	22.7	2.7	0.0	0.0	0.0	13.9

Table 5: Task-averaged success rates on RLBench (%) over the full viewpoint sweep, reported as mean ± SE across three training seeds.

Method	−90°	−75°	−60°	−45°	−30°	−15°	0°	+15°	+30°	+45°	+60°	+75°	+90°	Mean
No augmentation	15.5 ±0.7	22.1 ±0.6	24.8 ±0.6	31.2 ±0.8	36.9 ±1.0	50.9 ±1.3	62.5 ±0.4	41.9 ±1.0	31.4 ±1.6	27.6 ±0.5	21.6 ±0.6	22.2 ±0.7	18.7 ±0.6	31.3 ±0.7
VISTA-FT	37.4 ±0.4	38.2 ±1.3	37.6 ±1.1	42.3 ±0.1	47.4 ±1.1	51.4 ±0.6	55.5 ±1.1	50.0 ±1.4	44.3 ±2.4	39.2 ±0.9	38.6 ±0.6	37.9 ±0.7	38.2 ±0.6	42.9 ±0.8
VISTA	36.5 ±1.7	39.5 ±1.2	39.1 ±1.3	42.0 ±1.2	45.8 ±1.6	51.3 ±2.0	52.6 ±1.6	47.1 ±1.6	41.0 ±1.8	40.3 ±1.3	39.4 ±0.4	38.9 ±1.1	38.8 ±0.8	42.5 ±1.1
GenSplat	36.8 ±1.8	38.9 ±1.7	42.6 ±1.2	44.9 ±3.0	45.4 ±2.9	49.2 ±2.6	54.3 ±1.9	49.8 ±2.4	46.8 ±2.1	43.5 ±1.8	40.9 ±2.2	40.6 ±2.1	38.3 ±2.5	44.0 ±2.0
SHARP	36.8 ±1.6	40.9 ±0.8	41.3 ±2.0	44.2 ±1.6	47.1 ±0.8	53.9 ±1.0	57.1 ±2.1	48.0 ±1.7	41.2 ±0.8	38.3 ±0.3	38.6 ±0.7	37.3 ±0.4	36.2 ±0.3	43.1 ±1.0

C.2 LIBERO Results by Benchmark Suite

Table 6 reports results for the Spatial, Object, Goal, and Long suites, each containing ten tasks. Table 7 averages across all four suites.

Table 6: LIBERO success rates (%) by suite, method, and viewpoint. Values are means across three training seeds. Mean averages all 13 viewpoints.

Suite	Method	−90°	−75°	−60°	−45°	−30°	−15°	0°	+15°	+30°	+45°	+60°	+75°	+90°	Mean
Spatial	No augmentation	49.5	60.3	69.0	79.3	95.2	97.7	98.4	99.3	98.1	91.1	76.4	55.4	28.0	76.7
	VISTA	94.3	94.6	95.5	97.4	98.4	98.6	98.7	98.4	98.1	98.1	95.5	96.3	94.6	96.8
	VISTA-FT	91.1	94.2	94.6	97.7	98.5	98.3	98.7	98.3	98.0	95.4	93.9	93.6	90.3	95.6
	GenSplat	94.1	95.7	96.5	97.2	96.9	98.5	97.9	98.3	98.7	98.1	96.3	95.7	93.2	96.7
	SHARP	95.0	97.7	98.3	98.6	98.7	98.9	98.3	98.7	97.4	95.9	94.4	92.3	92.3	96.7
Object	No augmentation	64.3	80.4	91.9	93.7	98.5	98.3	98.9	98.1	97.0	97.1	92.4	74.3	33.2	86.0
	VISTA	98.1	98.6	99.0	98.7	98.9	99.2	98.7	97.1	98.3	98.3	98.3	97.9	97.5	98.3
	VISTA-FT	97.7	98.3	98.7	99.2	99.2	98.7	98.9	98.7	98.4	98.5	96.9	93.7	90.1	97.4
	GenSplat	97.3	97.9	97.9	98.9	98.9	98.7	98.1	96.7	98.1	98.7	97.7	97.9	96.4	97.9
	SHARP	98.9	98.5	99.1	98.7	99.2	98.9	98.4	98.1	98.5	98.3	98.5	97.3	95.7	98.3
Goal	No augmentation	50.3	62.1	76.1	85.5	95.5	97.1	97.3	90.9	79.3	69.4	62.2	57.5	56.9	75.4
	VISTA	88.5	89.3	89.1	92.7	93.1	96.0	94.9	91.2	89.5	89.9	89.3	88.1	84.0	90.4
	VISTA-FT	81.1	90.6	91.7	94.1	95.3	97.6	96.8	92.9	85.9	86.1	87.0	83.5	76.1	89.1
	GenSplat	85.4	91.1	92.6	94.3	94.9	96.0	95.4	93.4	89.8	90.7	90.7	88.1	80.7	91.0
	SHARP	83.5	86.7	89.7	92.9	93.2	95.8	96.7	91.9	87.5	88.5	87.5	84.7	79.1	89.0
Long	No augmentation	9.3	22.0	46.6	50.0	73.5	86.2	92.7	88.3	82.1	62.4	44.3	15.6	8.9	52.5
	VISTA	64.1	67.5	68.7	69.3	73.5	79.8	83.4	78.9	74.8	69.5	67.6	66.3	64.5	71.4
	VISTA-FT	61.1	64.7	67.0	67.8	74.9	83.5	89.1	84.9	81.9	74.2	67.6	61.3	54.9	71.8
	GenSplat	61.3	68.5	71.0	70.3	77.1	82.3	86.3	81.7	79.0	75.2	74.1	68.7	63.9	73.8
	SHARP	64.3	66.8	68.8	71.2	75.4	79.5	84.6	78.7	77.1	70.3	68.9	68.4	63.6	72.1

Table 7: Task-averaged success rates on LIBERO (%) over the full viewpoint sweep, reported as mean ± SE across three training seeds.

Method	−90°	−75°	−60°	−45°	−30°	−15°	0°	+15°	+30°	+45°	+60°	+75°	+90°	Mean
No augmentation	43.4 ±0.4	56.2 ±0.5	70.9 ±0.5	77.1 ±0.3	90.7 ±0.3	94.9 ±0.4	96.8 ±0.1	94.1 ±0.4	89.1 ±0.2	80.0 ±0.4	68.8 ±0.1	50.7 ±0.4	31.7 ±0.7	72.7 ±0.1
VISTA	86.3 ±0.4	87.5 ±0.4	88.1 ±0.3	89.5 ±0.7	91.0 ±0.2	93.4 ±0.2	93.9 ±0.4	91.4 ±0.1	90.1 ±0.3	89.0 ±0.4	87.7 ±0.1	87.1 ±0.7	85.1 ±0.1	89.2 ±0.2
VISTA-FT	82.8 ±0.9	87.0 ±0.5	88.0 ±0.4	89.7 ±0.3	92.0 ±0.1	94.5 ±0.4	95.9 ±0.5	93.7 ±0.1	91.0 ±0.2	88.5 ±0.4	86.4 ±0.5	83.0 ±0.6	77.9 ±0.7	88.5 ±0.2
GenSplat	84.5 ±0.3	88.3 ±0.4	89.5 ±0.4	90.2 ±0.6	91.9 ±0.1	93.9 ±0.1	94.4 ±0.2	92.5 ±0.6	91.4 ±0.1	90.7 ±0.1	89.7 ±0.3	87.6 ±0.1	83.5 ±0.2	89.9 ±0.2
SHARP	85.4 ±0.2	87.4 ±0.2	89.0 ±0.3	90.3 ±0.6	91.6 ±0.3	93.3 ±0.3	94.5 ±0.5	91.9 ±0.4	90.1 ±0.2	88.2 ±0.3	87.3 ±0.4	85.7 ±0.7	82.6 ±1.1	89.0 ±0.1

C.3 OpenVLA-OFT and X-VLA Results by Benchmark Suite

Table 8 and Table 9 extend Table 1 with results by LIBERO suite at all six shifted viewpoints under the evaluation protocol in Appendix A.3.2. In addition to No augmentation and SHARP, these tables include VISTA and GenSplat comparisons for both models. Averaged across the four suites and six shifted viewpoints, SHARP achieves success rates comparable to VISTA and GenSplat on both backbones. On OpenVLA-OFT, SHARP achieves 78.1%, compared with 78.2% for VISTA and 80.0% for GenSplat. On X-VLA, the corresponding success rates are 75.3%, 73.4%, and 74.1%, respectively.

Table 8: OpenVLA-OFT success rates (%) by LIBERO suite and viewpoint. Avg. averages the six shifted viewpoints.

Suite	Method	−75°	−45°	−15°	+15°	+45°	+75°	Avg.
Spatial	No augmentation	23.0	55.2	83.8	94.4	60.4	32.2	58.2
	VISTA	83.2	86.0	86.6	88.0	83.4	82.2	84.9
	GenSplat	74.8	81.4	94.2	95.0	87.4	76.4	84.9
	SHARP	80.4	81.6	85.8	86.0	82.4	80.0	82.7
Object	No augmentation	55.6	90.2	99.2	96.0	86.8	57.4	80.9
	VISTA	96.0	97.6	98.2	96.4	95.4	94.6	96.4
	GenSplat	93.0	95.8	96.2	96.4	96.6	95.4	95.6
	SHARP	93.8	96.6	96.4	97.4	95.0	91.8	95.2
Goal	No augmentation	7.2	40.2	88.0	80.6	63.4	40.6	53.3
	VISTA	53.8	64.2	80.4	78.2	78.8	72.8	71.4
	GenSplat	57.8	73.0	90.4	90.2	87.4	69.2	78.0
	SHARP	55.4	74.0	87.6	80.4	81.6	75.8	75.8
Long	No augmentation	17.8	35.0	66.6	77.2	35.8	22.0	42.4
	VISTA	48.8	54.8	70.4	74.2	62.0	50.2	60.1
	GenSplat	45.2	57.4	76.4	78.0	67.2	45.4	61.6
	SHARP	50.4	55.4	65.2	68.4	60.0	53.4	58.8

Table 9: X-VLA success rates (%) by LIBERO suite and viewpoint. Avg. averages the six shifted viewpoints.

Suite	Method	−75°	−45°	−15°	+15°	+45°	+75°	Avg.
Spatial	No augmentation	2.6	0.6	87.4	66.2	3.4	3.8	27.3
	VISTA	66.4	83.2	84.8	80.4	70.4	63.2	74.7
	GenSplat	52.6	82.8	87.4	76.0	73.8	60.0	72.1
	SHARP	70.4	80.0	86.8	82.8	73.4	63.6	76.2
Object	No augmentation	10.8	16.8	73.8	52.0	4.2	0.0	26.3
	VISTA	79.0	85.0	84.0	77.0	79.4	82.2	81.1
	GenSplat	72.8	94.0	90.4	80.6	89.0	83.6	85.1
	SHARP	93.4	94.4	94.0	95.2	92.8	90.2	93.3
Goal	No augmentation	4.2	18.0	92.2	95.2	16.2	7.8	38.9
	VISTA	79.8	85.4	92.2	90.8	84.8	78.6	85.3
	GenSplat	79.2	83.0	93.4	92.8	89.4	81.0	86.5
	SHARP	76.8	84.6	91.0	92.6	89.8	83.2	86.3
Long	No augmentation	2.4	13.4	69.0	72.8	11.0	5.8	29.1
	VISTA	44.2	53.6	60.6	56.0	54.6	46.0	52.5
	GenSplat	34.4	61.0	63.6	61.2	58.6	38.8	52.9
	SHARP	32.6	42.6	53.8	51.8	48.4	41.8	45.2

C.4 Generalization to Camera Elevation Angle and Distance

We test whether SHARP's robustness extends beyond azimuth changes by varying the third-person camera's elevation angle and distance separately (Figure 9). As in the azimuth evaluation, we rotate the camera around a fixed point. Here, we change its elevation angle by $\pm 15°$ and $\pm 30°$ from the source view. For distance, we move the camera closer to or farther from this point, using 0.7, 0.85, 1.4, or 2.0 times the original distance while keeping its orientation fixed. The wrist-camera setup remains unchanged in both evaluations. We evaluate the existing $\pi_{0.5}$ checkpoints for SHARP, No augmentation, VISTA, and GenSplat without further training, using 50 episodes per task and viewpoint on all 40 LIBERO tasks.

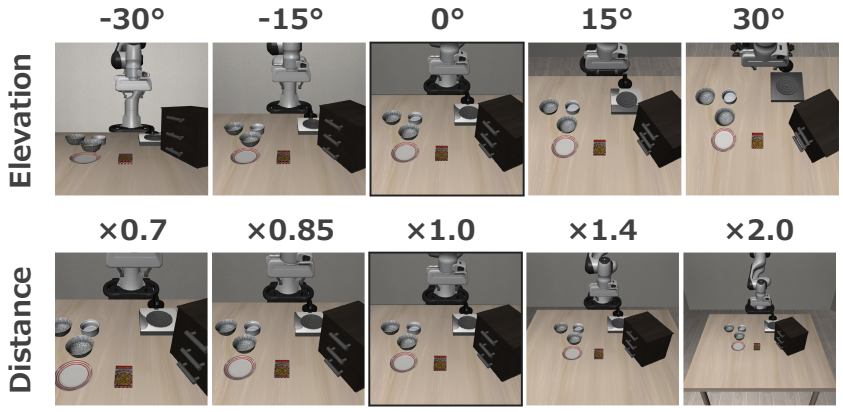

Figure 9: Camera viewpoint changes on LIBERO. Top: changes in elevation angle relative to the source view. Bottom: changes in camera distance, expressed as a multiple of the source-view distance.

Table 10 shows that SHARP achieves mean success rates of 90.5% across the four elevation angle shifts and 92.4% across the four distance changes, improving over No augmentation by 12.7 and 4.9 percentage points, respectively. These results suggest that its robustness extends beyond azimuth changes. SHARP also has higher average success than VISTA and GenSplat on both axes. However, the other methods were not trained with augmentations along these elevation angle and distance changes. VISTA and GenSplat use azimuth-only view augmentation as described in the main text.

Their results therefore serve as reference comparisons for these shifts, rather than comparisons with baselines trained using matching augmentation directions. Explicitly covering multiple types of viewpoint changes with these view-synthesis pipelines would require preparing augmented training data spanning the corresponding camera variations. In contrast, SHARP's low generation cost makes it practical to sample diverse geometric transformations online throughout training. These findings suggest that a single SHARP-trained policy may handle diverse viewpoint changes without the cost of preparing synthesized training data for each type of variation.

Table 10: Success rates (%) under camera elevation angle and distance changes on LIBERO, averaged across 40 tasks. Distance is relative to the source camera's distance to the fixed pivot. Each Avg. averages the four shifted viewpoints on its axis.

	Elevation angle offset					Relative distance				
Method	$-30°$	$-15°$	$+15°$	$+30°$	**Avg.**	0.7	0.85	1.4	2.0	**Avg.**
No augmentation	74.0	89.2	86.0	62.2	77.8	88.6	**95.6**	90.1	75.6	87.5
VISTA	85.1	89.8	90.2	86.4	87.8	90.5	93.3	89.4	87.1	90.1
GenSplat	84.6	91.8	90.3	86.3	88.2	89.5	94.4	89.5	84.6	89.5
SHARP	**88.5**	**93.2**	**92.2**	**88.4**	**90.5**	**93.5**	94.5	**92.2**	**89.5**	**92.4**

D Additional Ablation Experiments

D.1 Benefit of Online Augmentation

SHARP's low image-synthesis cost (Table 3) makes online augmentation practical, allowing it to sample a fresh homography whenever a frame is drawn during training. In comparison, our VISTA and GenSplat pipelines precompute augmented images and reuse them throughout finetuning. We examine how SHARP performs when its number of stored augmented images is matched to these baselines, and whether its low cost contributes to performance by enabling fresh augmentations throughout training.

We compare online SHARP with a fixed-image variant, denoted **SHARP (fixed)**, which stores one homography-augmented third-person image per source frame and reuses it throughout finetuning. This matches the number of stored augmented images per frame used by VISTA and GenSplat. For each frame, we independently sample one transformation from the same distribution as online SHARP (Appendix A.2). Both conditions use the main comparison's source demonstrations, checkpoints, and $\pi_{0.5}$ training settings (Appendix A.3.1), with all other inputs and labels retained. Evaluation follows the full sweep in Appendix C. Percentage-point differences are computed before rounding.

Results. In Table 11, SHARP (fixed) still outperforms No augmentation in mean success on both RLBench (40.6% versus 32.5%) and LIBERO (86.7% versus 72.7%). However, it falls below online SHARP (44.0% and 89.3%, respectively), with gaps of 3.3 and 2.6 percentage points. On LIBERO, the gap increases to 6.1 percentage points at $-90°$ and 5.7 percentage points at $+90°$. With this matched number of stored augmented images, SHARP (fixed) also falls below VISTA on both RLBench (40.6% versus 44.0%) and LIBERO (86.7% versus 89.2%), and below GenSplat on LIBERO (89.5%).

These results highlight low augmentation cost as a key strength of SHARP. This low cost makes it practical to generate a fresh geometric transformation each time a source frame is sampled, exposing the policy to more diverse augmented observations throughout training while retaining the original action labels. The improvement over SHARP (fixed) suggests that this online diversity contributes to SHARP's competitive performance relative to learned view synthesis. Thus, SHARP's efficiency provides both a computational advantage and a practical means of improving policy performance through online augmentation.

Table 11: No augmentation, one fixed homography image per source frame, and online resampling with $\pi_{0.5}$ (success rate in %). Mean averages all 13 viewpoints.

Method	−90°	−75°	−60°	−45°	−30°	−15°	0°	+15°	+30°	+45°	+60°	+75°	+90°	Mean
RLBench														
No augmentation	16.5	23.2	26.0	32.5	38.2	53.5	63.2	42.3	34.7	28.2	22.5	23.2	19.2	32.5
SHARP (fixed)	35.7	38.8	40.2	43.0	48.8	52.7	54.3	43.0	38.7	33.5	33.8	33.8	32.0	40.6
SHARP (online)	38.7	41.2	43.8	44.8	47.7	55.5	59.8	49.5	41.7	38.5	38.2	36.7	35.8	44.0
LIBERO														
No augmentation	44.0	55.2	70.0	76.8	91.0	95.0	97.0	93.9	89.4	79.4	68.9	51.4	33.2	72.7
SHARP (fixed)	79.7	83.5	86.6	88.6	90.3	92.8	94.0	92.0	89.4	85.6	84.7	81.6	77.9	86.7
SHARP (online)	85.8	87.0	88.4	91.0	91.7	93.0	94.1	92.8	90.5	88.0	87.8	86.9	83.6	89.3

D.2 Homography Components

Table 12 provides the per-angle success rates underlying Figure 5. Removing pitch, yaw, or translation causes the largest performance drops, concentrated at large viewpoint shifts. Removing roll or scaling changes mean success only slightly. For small viewpoint shifts, success rates differ little between full SHARP and variants with one transformation component removed.

Table 12: Homography component ablation across the LIBERO angular sweep (task-averaged success rate in %). Full uses all components. Each ablated variant removes the named component while retaining all others. Avg. averages all 12 shifted angles. Each condition uses one training seed.

Method	−90°	−75°	−60°	−45°	−30°	−15°	+15°	+30°	+45°	+60°	+75°	+90°	Avg.
SHARP (w/o translation)	80.7	85.2	88.8	89.3	93.1	93.9	93.4	90.9	88.5	86.2	83.1	78.8	87.6
SHARP (w/o scaling)	85.8	87.1	88.5	89.9	92.0	93.0	93.0	90.4	89.2	87.5	85.3	83.2	88.7
SHARP (w/o roll)	84.4	87.6	88.9	89.4	91.8	93.0	92.6	91.0	88.4	86.9	86.2	82.2	88.5
SHARP (w/o pitch)	79.2	84.2	87.9	89.3	92.1	94.9	92.5	90.9	86.9	84.1	77.6	72.2	86.0
SHARP (w/o yaw)	79.7	84.1	87.4	89.3	91.3	93.7	92.8	90.3	87.6	86.2	82.2	75.8	86.7
SHARP (full)	85.8	87.0	88.4	91.0	91.7	93.0	92.8	90.5	88.0	87.8	86.9	83.6	88.9

D.3 Warp Strength

The strength ablation in the main text (Figure 6) varies the maximum magnitude of one homography component at a time while keeping the other components at their default settings. In this section, we complement that analysis by varying the overall warp strength. We jointly multiply the sampling ranges for virtual-camera rotation and translation by the same factor $s \in \{0.25, 0.5, 0.75, 1\}$. Here, $s = 1$ is the full SHARP augmentation in Appendix A.2, while $s = 0.5$, for example, halves these ranges and produces milder warps. All other training settings follow Appendix A. We evaluate each policy on LIBERO at 0° to 90° in 15° steps. No augmentation denotes source-view finetuning without added viewpoint augmentation.

Results. Figure 10 reports the average across all tasks, and Figure 12 separates the four suites. Consistent with the component-wise ablation in the main text, stronger augmentation is especially beneficial at large viewpoint shifts. At 90°, success rises from 35.2% with $s = 0.25$ to 83.6% with $s = 1$. This improvement with each increase in strength holds in all four suites, with Long remaining the most challenging. Stronger warps do not uniformly improve performance near the source view. Their main benefit is robustness to larger shifts.

D.4 Third-Person-Camera-Only Comparison

We examine whether the benefits of SHARP and learned view-synthesis augmentation persist when wrist images are unavailable, testing the role of stable wrist observations in our hypothesis. All four policies in Figure 11 start from a wrist-free checkpoint obtained by full finetuning for 25,000 updates, with zero wrist pixels and a disabled wrist-image mask. No augmentation, SHARP, VISTA-FT, and GenSplat are then finetuned for 1,000 updates using the same training hyperparameters as their counterparts with wrist-camera input (Appendix A.3.1).

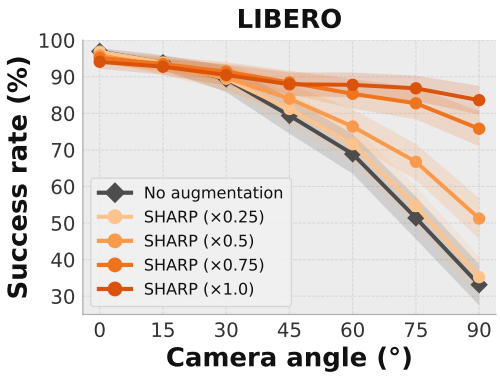

Figure 10: Warp-strength ablation on LIBERO. Increasing the overall homography strength improves success rates particularly at large viewpoint shifts.

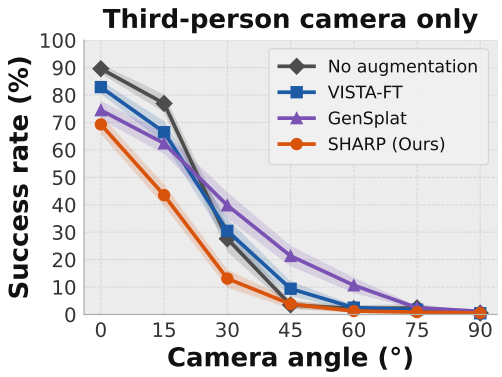

Figure 11: Viewpoint generalization on LIBERO using only the third-person camera images, with wrist images omitted during both finetuning and evaluation.

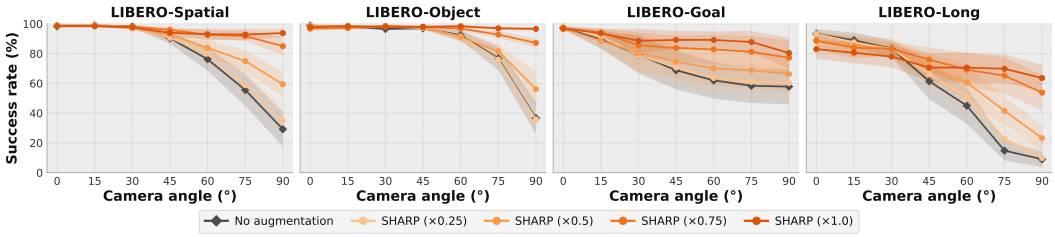

Figure 12: Warp-strength ablation by LIBERO suite. Each panel reports the mean over ten tasks at each angle for No augmentation and SHARP with $s \in \{0.25, 0.5, 0.75, 1\}$.

Results. In Figure 11, SHARP does not improve average success over No augmentation under this third-person-only training setup: it achieves 18.9% across the seven evaluated viewpoints, compared with 29.0% for No augmentation. VISTA-FT also falls below No augmentation at 27.7%, while GenSplat reaches 30.3%, with gains concentrated at viewpoint shifts from 30° to 60°. SHARP, VISTA-FT, and GenSplat all perform substantially worse than their two-camera counterparts at the corresponding viewpoints (Table 7).

Together with the analyses in Section 5.4, these results suggest that the gains from geometric augmentation rely on stable wrist observations. The augmentations produced by SHARP, VISTA-FT, and GenSplat do not necessarily reproduce actual viewpoint changes faithfully. Stable wrist observations may help the policy tolerate these imperfect augmentations. Without them, mismatches between augmented training images and observations at new viewpoints may have a greater impact on control, which could explain the lower performance when only third-person images are available. These results do not contradict our hypothesis that geometric augmentation improves viewpoint robustness when stable wrist observations are available (Section 3).

E Ablation Protocols

E.1 Attention Measurement

For the camera-attention analyses in Figure 2 and Figure 7 **(Left)**, we measure how much attention action tokens assign to third-person and wrist-image tokens. The LIBERO analysis uses the same 600 source-view observations for Pretrained and the finetuned policies. Figure 2 compares Pretrained, No augmentation, and VISTA. Figure 7 **(Left)** reports changes from Pretrained for No augmentation, VISTA, VISTA-FT, GenSplat, and SHARP on RLBench and LIBERO. We sample five frames from each of three demonstrations per LIBERO task: one at the start and the others at 25%, 50%, 75%, and 95% of the demonstration's length in timesteps. Let x_τ denote the action-

chunk variable supplied to the action expert at flow time τ. The action probe fixes $x_\tau = \mathbf{0}$ and $\tau = 0.5$. For each observation and camera stream, we sum attention over its image tokens and average over action queries, all heads, and 18 layers to obtain A_{third} and A_{wrist}. We then compute the wrist-image attention share $r = A_{\text{wrist}}/(A_{\text{third}} + A_{\text{wrist}})$ for each observation, average r first within each task and then across tasks, and report the share in %. For changes from Pretrained, we compute $100(r - r_{\text{pretrained}})$ on paired observations before averaging within and across tasks, and report the result in percentage points. Error bars show SE across task means.

E.2 Camera-Input Controls

Wrist-only reference and oracle. For the comparison in Figure 7 **(Middle)**, we train a wrist-only reference and a target-view oracle using the same pretrained checkpoint and finetuning hyperparameters as the other $\pi_{0.5}$ LIBERO experiments (Appendix A.3.1). The wrist-only reference uses the same source demonstrations. It blacks out the third-person image during both training and evaluation while retaining wrist images and instruction inputs. Unlike the inference-time Black intervention, this policy is trained separately. It is evaluated once, yielding the horizontal reference in the figure. For the oracle, we render demonstration observations in the simulator at each target camera angle and finetune a separate policy on the resulting target-view data. Each policy is evaluated at the same camera angle used for training, providing a reference for performance when target-view demonstrations are available.

E.3 Augmentation Controls

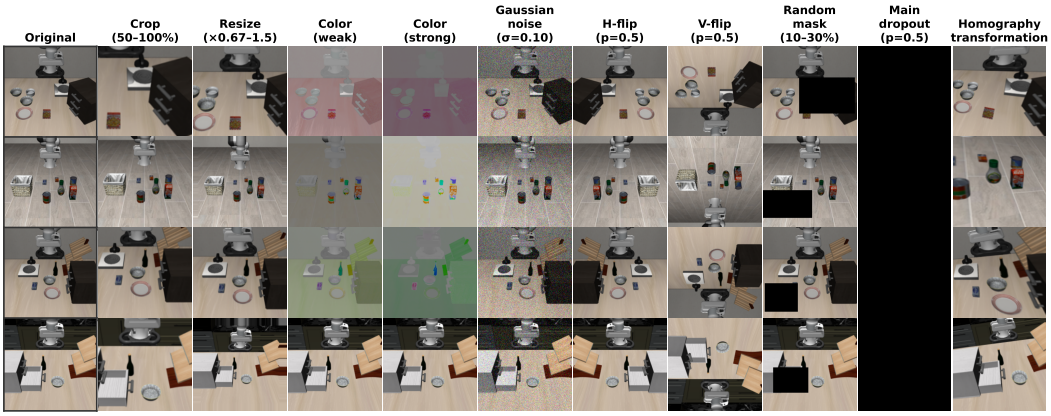

Figure 13: Sample third-person images for the augmentation-type ablation in Table 2. Each row shows the same source frame under the different augmentations, with the original image in the first column and SHARP's homography transformation in the last. The dropout and flip columns show transformed samples. Each transformation is applied with probability 0.5 during training.

For the augmentation comparison in Table 2, all finetuned policies use the same source demonstrations, checkpoint, optimizer, LoRA configuration, and training budget (Appendix A). These augmentations modify only the third-person training images, retaining wrist images and action labels. No augmentation finetunes on the source images without additional perturbations. Crop extracts a randomly positioned square region, with its side length sampled uniformly between 50% and 100% of the original, and resizes it to the original resolution. Resize scales the image about its center by a factor sampled log-uniformly between $2/3$ and 1.5, filling uncovered borders by repeating edge pixels. Color (weak) and Color (strong) apply color augmentation at two intensity levels. Noise adds independent zero-mean Gaussian noise to each pixel and color channel, with standard deviation 0.10 on the $[0, 1]$ intensity scale, and clips the result to this range. Horizontal flip and Vertical flip apply their respective reflection with probability 0.5. Random mask covers 10–30% of each image with a randomly placed black rectangle. Dropout replaces the entire third-person image with a black image with probability 0.5. SHARP applies the homography augmentation specified in Appendix A.2. These perturbations are sampled independently for each frame. At evaluation, all

Table 13: Target-demo task groups. Short descriptions identify each task. Each cell contributes one demonstration.

Scene	Group 1	Group 2	Group 3
Living room	Alphabet soup and tomato sauce into basket	Alphabet soup and cream cheese box into basket	Cream cheese box and butter into basket
Kitchen	Turn on stove and place moka pot on it	Bowl into bottom drawer and close it	Yellow/white mug into microwave and close it
Study	Book into back compartment of caddy	Book into back compartment of caddy	Book into back compartment of caddy
Floor	Alphabet soup into basket	Ketchup into basket	BBQ sauce into basket
Tabletop	Bowl between plate and ramekin onto plate	Open middle drawer	Bowl in top drawer onto plate

controls receive both current camera images without these perturbations. Figure 13 illustrates each image transformation.

F One-Shot Viewpoint Adaptation

Although SHARP requires no target-view demonstrations, a small amount of target-view data may be available at deployment and provide useful information about the new viewpoint. This motivates examining whether homography augmentation can also improve performance when combined with one-shot adaptation. In this section, we extend SHARP to this setting and compare it with existing adaptation methods.

F.1 Problem Setting

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ denote a set of tasks that a pretrained policy can perform under the source camera viewpoint. Let $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tgt}}$ denote successful demonstrations recorded at v_{src} and v_{tgt}, respectively, with the wrist camera, embodiment, and action coordinates unchanged. The source-only setting assumes $\mathcal{D}_{\text{tgt}} = \emptyset$.

Following DART (Kang et al., 2026), one-shot adaptation uses a single target demonstration of task $T_j \in \mathcal{T}$ to recover performance across $\mathcal{T}$ under the same viewpoint change. DART uses only source and target demonstrations of T_j. SHARP jointly finetunes on $\mathcal{D}_{\text{src}}$ and $\mathcal{D}_{\text{tgt}}$ with either the same source demonstrations as DART or one source demonstration per task in $\mathcal{T}$. Access to source demonstrations across tasks is a natural assumption when the data used to train the source policy are retained. Reusing these demonstrations provides broader source-task coverage without requiring additional target-view data collection.

We retain the source loss $\mathcal{L}_s$ from Equation 3 and add the target-view loss

$$\mathcal{L}_t(\theta) = \mathbb{E}_{(o_t, a_{t:t+H}) \sim \mathcal{D}_{\text{tgt}}}[\mathcal{L}_{\text{FM}}(\theta, o_t, a_{t:t+H})]. \tag{4}$$

Each update draws half of its samples from $\mathcal{D}_{\text{src}}$ and half from $\mathcal{D}_{\text{tgt}}$ and minimizes $\mathcal{L}_s + \mathcal{L}_t$. Homography augmentation applies only to the source images.

F.2 Experimental Setup

F.2.1 LIBERO Task Groups

Following the scene-wise LIBERO protocol of Kang et al. (2026), each task group contains five target-view demonstrations, one per scene category (Table 13). The source set is shared across task groups.

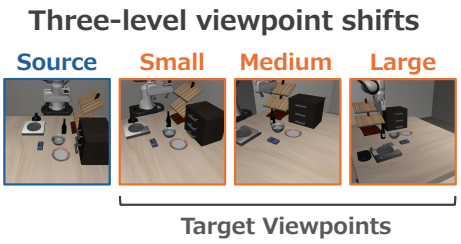

Figure 14: LIBERO source view and the Small, Medium, and Large target viewpoints used for one-shot adaptation.

Table 14: Third-person camera pose changes from the source view on LIBERO. The displacement vector is $\Delta\mathbf{p} = (\Delta x, \Delta y, \Delta z)^\top$. Position offsets and displacement norms are in meters. Rotation reports the range of relative rotation angles across scenes, in degrees.

View	Δx	Δy	Δz	$\|\Delta\mathbf{p}\|$	Rotation (°)
Original	0.0	0.0	0.0	0.000	0.0
Small	0.0	0.3	−0.1	0.316	19.4–33.4
Medium	−0.2	0.7	−0.2	0.755	46.5–70.6
Large	−1.2	1.0	−0.2	1.575	107.4–128.0

F.2.2 Camera Configurations

Following DART (Kang et al., 2026), we use scene-specific Small, Medium, and Large shifts (Figure 14), with third-person camera offsets in Table 14 and corresponding orientation changes.

F.2.3 Baselines

Pretrained evaluates the publicly released $\pi_{0.5}$ LIBERO checkpoint without additional finetuning. **One-shot FT** uses full-parameter or LoRA (Hu et al., 2021) finetuning on a single target-view demonstration per scene. **FLA** (Li et al., 2026) adapts the visual backbone using the same demonstration budget. **RETAIN** (Yadav et al., 2026) merges the model finetuned on the target domain with the pretrained model to preserve pretrained capabilities. **DART** (Kang et al., 2026) extracts a transferable domain-specific update from paired source- and target-domain finetuning updates.

All adaptation methods use the same five target-view demonstrations. One-shot FT, FLA, and RETAIN use only these demonstrations, while DART and SHARP additionally use the source-view demonstrations described in Appendix F.1.

F.2.4 Training and Evaluation Details

For One-shot FT (LoRA) and SHARP, we apply rank-32 LoRA to the vision encoder, language backbone, and action expert, following the configuration in Appendix A.3.1.

FLA instead uses rank-16 adaptation of the visual backbone. DART forms its domain update from separate full finetuning runs on the five source and five target demonstrations.

All one-shot runs train for 1,000 updates, except FLA, which trains for 1,500. Full finetuning, the DART finetuning runs, and target-only LoRA use batch size 64, and FLA uses 32. SHARP uses a batch size of 64, with 32 target-view and 32 source-view samples per update. Its ablation without homography augmentation uses the same batch composition. SHARP results use the same homography distribution as the source-only setting, specified in Appendix A.2.

We tune the RETAIN and DART coefficients on LIBERO-Long under the Medium viewpoint shift using task group 1, then fix them across all evaluations.

We evaluate each adapted policy on all 40 LIBERO tasks, including the 35 tasks without target demonstrations. We average adapted-policy results over the three task groups of Appendix F.2.1.

F.3 Results

F.3.1 Comparison with Baselines

In Table 15, SHARP with 40 source demonstrations achieves **85.3%** average success, exceeding the strongest baseline, DART, by 8.2 percentage points, with its largest gain under the Large shift. At matched demonstration budgets, five-demo SHARP exceeds DART by 5.4 percentage points on average, although it performs worse on Small. Increasing SHARP's source set to 40 demonstrations adds 2.8 percentage points. These results show that combining one-shot adaptation with homog-

Table 15: Success rates across viewpoint-shift scales on LIBERO (%). Results for adapted policies are averaged over three task groups. All adapted policies use the same five target demonstrations, one per scene. SHARP (5 source demos) uses the same source demonstrations as DART, while SHARP (40 source demos) uses one per task. Both SHARP variants use the same homography range as the source-only setting. Green denotes the average gain over the strongest baseline in percentage points.

Method	Small	Medium	Large	Avg. Success
Pretrained	89.2	64.4	11.8	55.1
One-shot FT (full)	41.1	27.0	21.9	30.0
One-shot FT (LoRA)	89.7	79.2	60.8	76.6
FLA (Li et al., 2026)	91.3	76.4	49.3	72.3
RETAIN (Yadav et al., 2026)	86.6	70.9	43.8	67.1
DART (Kang et al., 2026)	91.7	78.2	61.3	77.1
SHARP (5 source demos)	87.7	82.8	76.9	82.5 (+5.4)
SHARP (40 source demos)	91.2	85.1	79.7	**85.3** (+8.2)

Table 16: Effect of the additional homography transformation in one-shot adaptation. Both rows use 40 source demonstrations, five target demonstrations, 32+32 samples per update, and identical training settings.

Method	Small	Medium	Large	Avg.
SHARP (w/o homography transformation)	89.1	76.0	54.3	73.1
SHARP	91.2	85.1	79.7	**85.3**

raphy augmentation of source-view demonstrations outperforms the existing adaptation methods in average success rate.

F.3.2 Contribution of the Homography Transformation

To test whether homography augmentation contributes beyond access to more source data, we compare SHARP with and without it using the same 40 source and five target demonstrations (Table 16). Simply adding source demonstrations without homography augmentation reduces average success from 76.6% for target-only One-shot FT (LoRA) (Table 15) to 73.1%, with lower success at all three shift scales. Combining the source demonstrations with homography augmentation raises average success to 85.3%, exceeding One-shot FT (LoRA) by 8.7 percentage points. The largest gain over training on source and target data without homography occurs under the Large shift (54.3% to 79.7%). These results show that source-data reuse improves one-shot adaptation when combined with homography augmentation in this setting.

F.3.3 Results by Benchmark Suite

Table 17 breaks down Table 15 by suite and viewpoint shift.

Table 17: LIBERO success rates by suite and viewpoint shift (%). Target-adapted policies report means over three task groups. Pretrained is evaluated without additional finetuning. Average equally weights the four suites, matching Table 15.

Method	Novel viewpoints (success rate, %)				
	Spatial	Object	Goal	Long	Average
Small viewpoint shift					
Pretrained	92.2	95.6	85.4	83.6	89.2
One-shot FT (full)	63.5	35.4	39.5	25.8	41.1
One-shot FT (LoRA)	97.5	95.0	83.5	82.9	89.7
FLA (Li et al., 2026)	97.8	96.5	84.7	86.3	91.3
RETAIN (Yadav et al., 2026)	95.7	90.1	79.7	80.9	86.6
DART (Kang et al., 2026)	97.6	97.5	83.5	88.2	**91.7**
SHARP (5 demos)	97.2	96.7	79.1	77.9	87.7
SHARP (40 demos)	97.4	97.6	84.5	85.2	91.2
Medium viewpoint shift					
Pretrained	68.2	84.8	74.4	30.0	64.4
One-shot FT (full)	38.6	28.5	23.7	17.3	27.0
One-shot FT (LoRA)	84.2	96.1	72.3	64.1	79.2
FLA (Li et al., 2026)	83.7	95.0	68.6	58.2	76.4
RETAIN (Yadav et al., 2026)	78.6	86.5	62.1	56.3	70.9
DART (Kang et al., 2026)	82.9	96.1	76.0	57.7	78.2
SHARP (5 demos)	91.5	96.7	76.7	66.3	82.8
SHARP (40 demos)	91.0	96.7	82.5	69.9	**85.1**
Large viewpoint shift					
Pretrained	10.2	0.2	31.0	5.6	11.8
One-shot FT (full)	29.5	29.3	13.6	15.4	21.9
One-shot FT (LoRA)	65.5	84.1	45.4	48.4	60.8
FLA (Li et al., 2026)	47.9	58.1	45.1	46.0	49.3
RETAIN (Yadav et al., 2026)	37.8	66.1	32.1	39.2	43.8
DART (Kang et al., 2026)	60.3	85.4	46.7	52.9	61.3
SHARP (5 demos)	83.6	95.7	71.2	56.9	76.9
SHARP (40 demos)	90.3	92.7	73.5	62.2	**79.7**