# Efficient Camera Pose Augmentation for View Generalization in Robotic Policy Learning

Sen Wang[1], Huaiyi Dong[1], Jingyi Tian[1], Jiayi Li[1], Zhuo Yang[1], Tongtong Cao[2], Anlin Chen[2], Shuang Wu[2], Le Wang[1], Sanping Zhou[1,✉]

Xi'an Jiaotong University, Noah's Ark Lab
Code is available at: https://github.com/SanMumumu/GenSplat

**Abstract.** Prevailing 2D-centric visuomotor policies exhibit a pronounced deficiency in novel view generalization, as their reliance on static observations hinders consistent action mapping across unseen views. In response, we introduce GenSplat, a feed-forward 3D Gaussian Splatting framework that facilitates view-generalized policy learning through novel view rendering. GenSplat employs a permutation-equivariant architecture to reconstruct high-fidelity 3D scenes from sparse, uncalibrated inputs in a single forward pass. To ensure structural integrity, we design a 3D-prior distillation strategy that regularizes the 3DGS optimization, preventing the geometric collapse typical of purely photometric supervision. By rendering diverse synthetic views from these stable 3D representations, we systematically augment the observational manifold during training. This augmentation forces the policy to ground its decisions in underlying 3D structures, thereby ensuring robust execution under severe spatial perturbations where baselines severely degrade.

**Keywords:** Feed-Forward 3D Gaussian Splatting · 3D-Prior Distillation · View Generalization Policy Learning

## 1 Introduction

While recent advances in large-scale imitation learning have demonstrated considerable promise in Embodied AI [3, 4, 11, 48, 65, 68], their efficacy remains strictly contingent upon expansive, diverse datasets [9, 23, 30, 53]. This data dependency exposes a fundamental vulnerability in the prevailing observation-to-action paradigm: current 2D-centric policies exhibit a pronounced deficiency in novel view generalization. Specifically, because these models inherently rely on perspective-dependent visual representations, even minor extrinsic $SE(3)$ perturbations disrupt the spatial grounding of action-relevant cues [1, 14, 32, 35]. Mitigating this spatial fragility through the exhaustive re-collection of demonstrations for unseen viewpoints imposes prohibitive marginal costs on teleoperation, thereby hindering scalable real-world deployment [23, 39]. This dilemma naturally motivates a critical inquiry: *How can we augment the observational manifold to induce view generalization without additional human collection?*

✉ Corresponding author.

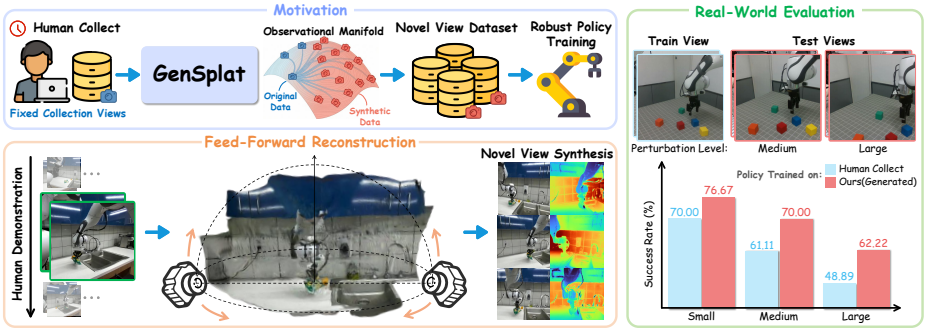


**Fig. 1: The diagram of the proposed GenSplat.** Given expert demonstrations, GenSplat employs a feed-forward 3D reconstruction pipeline to render geometrically consistent novel viewpoints under controlled perturbations, explicitly boosting camera pose diversity. Experiments demonstrate that policies trained on GenSplat-augmented data achieve competitive performance across different perturbation levels, substantially outperforming those trained solely on human-collected demonstrations.

To address this, prior efforts have leveraged novel view synthesis (NVS) to artificially expand camera pose diversity, broadly falling into two paradigms: 2D diffusion-based generation [40, 42, 52, 59, 67] and 3D Gaussian-based reconstruction [26, 51, 58, 61]. **Generative methods** employ diffusion models for plug-and-play viewpoint synthesis, bypassing the need for explicit 3D reconstruction or rigid camera calibration [5, 45, 50, 59]. Although recent adaptations incorporate 3D-aware conditioning [28, 40], their optimization objectives remain strictly confined to the 2D image plane. Consequently, these generated augmentations lack a unified metric workspace, frequently violating epipolar geometry and multi-view visibility constraints. This manifests as severe scale drift, erroneous parallax, and implausible occlusions, which are fatal for precise robotic manipulation. Moreover, underutilizing the multi-camera topology of robotics datasets restricts both geometric fidelity and synthesis controllability.

In contrast, **Gaussian-based methods** parameterize explicit 3D scenes using anisotropic Gaussian primitives [22, 33]. Rendering from this shared, underlying geometry inherently guarantees strict cross-view physical consistency by explicitly preserving exact occlusion ordering, multi-view parallax, and metric scale, all while enabling real-time synthesis. However, classical 3DGS pipelines suffer from restrictive operational bottlenecks: achieving high-fidelity reconstruction typically necessitates dense multi-view coverage, precise camera pose calibration, and prohibitively expensive per-scene optimization [12, 26, 51, 58, 61]. Crucially, real-world robotic deployments are fundamentally characterized by extreme viewpoint sparsity, entirely uncalibrated observation setups, and the necessity for instant adaptation across diverse environments. Motivated by this domain gap, we propose a feed-forward reconstruction paradigm capable of instant, high-fidelity scene synthesis from sparse, uncalibrated observations, circumventing the limitations of traditional per-scene optimization.

In this work, we introduce GenSplat, a feed-forward 3D Gaussian Splatting framework designed to expand the observational manifold via novel view synthesis, thereby inducing robust view generalization in visuomotor policies. Circumventing the prohibitively expensive per-scene tuning of classical pipelines, GenSplat leverages recent advances in feed-forward 3DGS [19, 60] to parameterize 3D scenes from sparse, uncalibrated inputs in a single forward pass. This enables efficient, physically grounded multi-view synthesis, systematically expanding the critical dimension of viewpoint diversity required by modern imitation learning architectures [3, 6, 41]. However, directly applying naive feed-forward mechanisms to the extreme sparsity of robotic datasets [9, 23, 30, 53] often precipitates severe topological fragmentation and inconsistent occlusions [22, 33]. While such photometric artifacts might be tolerated in pure visual rendering, they severely corrupt the geometric consistency mandatory for reliable visuomotor supervision. To overcome this structural collapse, we propose a 3D-prior distillation objective that extracts geometric knowledge from a 3D foundation model [57] to explicitly regularize the reconstruction landscape. Consequently, GenSplat yields metrically consistent viewpoint augmentations that systematically enhance downstream policy robustness without incurring additional data collection costs. In summary, our primary contributions are:

(1) We propose GenSplat, a feed-forward 3DGS framework that enables geometrically consistent view augmentation for robotic datasets with sparse and uncalibrated camera observations. The design effectively eliminates the need for sensor calibration, physical viewpoint adjustment, and per-scene optimization.

(2) We introduce a 3D-prior distillation strategy that transfers geometric knowledge from a pretrained 3D foundation model to regularize reconstructions, ensuring that synthesized novel views are geometrically consistent and physically reliable for policy learning.

(3) Extensive real-world experiments demonstrate that policies trained with GenSplat-augmented data, achieve strong generalization to unseen camera poses. GenSplat significantly enhances policy performance while demonstrating superior computational efficiency compared to prior methods.

## 2 Related Work

**Robot Data Collection Pipelines.** The pursuit of generalizable visuomotor policies has catalyzed the curation of large-scale, multi-task robotic datasets [9, 23, 30, 53]. These corpora are typically collected via teleoperation or kinesthetic teaching across diverse environments, pairing observations and natural-language instructions with continuous actions [18, 62]. While pretraining on such data yields robust initialization [3, 24, 36, 65, 66], transferring these models to novel physical deployments necessitates target-domain finetuning. Crucially, hardware constraints typically restrict deployment-time collection to a sparse set of fixed extrinsic configurations. Consequently, expanding the observational support to novel viewpoints mandates repeated human demonstrations, imposing an unscalable teleoperation bottleneck [18, 34, 62]. GenSplat fundamentally bypasses

this limitation by leveraging feed-forward 3DGS to expand the observational manifold, thereby enriching viewpoint diversity without incurring marginal data collection costs.

**3D Gaussian Splatting in Robotics.** 3DGS has emerged as a practical tool for reconstructing robot workspaces and synthesizing photorealistic observations to support policy learning [16, 17, 26, 51, 58, 61]. Several representative systems illustrate this trajectory: Real2Render2Real reconstructs real-world workspaces as Gaussian scenes and exports renderable digital twins for closed-loop policy training and evaluation [61]; RE$^3$SIM adopts a real-to-sim pipeline that preserves geometric structure and occlusion relationships to enable realistic manipulation benchmarking [16]; RoboGSim builds upon Gaussian reconstructions to provide controllable cameras, lighting, and object layouts, facilitating studies on cross-view generalization [26]; and EnerVerse scales synthetic data generation by composing reusable Gaussian assets with world models to produce diverse, physically plausible scenes [17]. Recent pipelines further advance novel-view synthesis while retaining metric-scale accuracy in workspace reconstruction [31, 51, 58]. Collectively, these pioneering works demonstrate that 3DGS enables physically consistent, multi-view rendering and high-fidelity digital twins for robotic applications. Despite strong geometric fidelity, these pipelines typically require dense, calibrated multi-view capture and costly per-scene optimization, which limits practicality for the sparse and uncalibrated conditions common in robotic manipulation datasets [9, 23, 30, 53]. GenSplat bypasses these bottlenecks via a feed-forward 3DGS paradigm, directly parameterizing 3D structures from sparse, unposed inputs in a single forward pass. This enables scalable expansion of the observational manifold without rigid calibration or per-scene optimization.

**View Generalization Policy Learning.** Robust physical deployment necessitates strong $SE(3)$ viewpoint generalization, a challenge as fundamental as task-level generalization [5, 32, 45, 50, 59]. Despite the strong in-distribution convergence of modern Vision-Language-Action (VLA) models [3, 24, 65, 66, 68] and Diffusion Policies [6], representations anchored to static camera extrinsics exhibit severe generalization gaps under spatial perturbations. While prior efforts attempt to induce such generalization via architectural inductive biases or algorithmic regularizations [2, 15, 20, 43, 49, 55], their efficacy remains strictly upper-bounded by the spatial density of the training manifold. Rather than modifying the core policy architecture [1, 25, 27, 64], we expand the observational manifold to densify the $SE(3)$ spatial support, inherently inducing view-generalized spatial groundings directly from metrically consistent synthetic observations.

## 3 Method

We detail **GenSplat**, a feed-forward 3D Gaussian Splatting (3DGS) framework designed to expand the observational manifold via novel view synthesis from sparse, uncalibrated visual demonstrations. By reconstructing metrically consistent 3D workspaces without rigid camera calibration, GenSplat densifies the continuous $SE(3)$ spatial support of the training data, inherently inducing view-

generalized action mapping for downstream policies. The overarching architectural pipeline is illustrated in Fig. 2.

### 3.1 Problem Formulation

We consider a robotic imitation learning dataset $\mathcal{D} = \{(\zeta_i, l_i)\}_{i=1}^{N_\mathcal{D}}$, where each demonstration $\zeta$ comprises a sequence of observations $\mathcal{O} = \{I_t\}_{t=1}^{T}$, paired actions $\mathcal{A} = \{\boldsymbol{a}_t\}_{t=1}^{T}$, and a language instruction $l$. Our goal is to train visuomotor policies that generalize to novel camera views without additional data collection. To this end, we construct a per-frame, pose-free camera augmentation pipeline, denoted as $\mathcal{G}_\theta$. Given $V$ camera views $\{I_j\}_{j=1}^{V}$ at a time step $t$, $\mathcal{G}_\theta$ represents the 3D scene by a set of Gaussian primitives $\mathcal{S}_t$, predicts the camera parameters for each view $\boldsymbol{p}_t^v$, and renders images from sampled target camera poses. The resulting augmented dataset is $\tilde{\mathcal{D}} = \{(\tilde{\zeta}_i, l_i)\}_{i=1}^{N}$, with $\tilde{\zeta}$ replacing each original frame $(I_t, \boldsymbol{a}_t)$ by $(\hat{I}_t^{\Delta}, \boldsymbol{a}_t)$, where $\hat{I}_t^{\Delta}$ is a novel view rendered from $\mathcal{S}_t$ under a regularly sampled camera transform $\Delta$. The policy $\pi(\boldsymbol{a}_t \mid \mathcal{D}_{\text{train}})$ is trained on the combined dataset $\mathcal{D}_{\text{train}} = \mathcal{D} \cup \tilde{\mathcal{D}}$, inherently inducing view-generalized action mapping.

### 3.2 Model Architecture

To address the reconstruction instability and artifacts caused by improper reference viewpoint selection in traditional methods [21, 54, 56], we design a feed-forward permutation-equivariant architecture. This formulation processes sparse, uncalibrated inputs symmetrically, ensuring robust 3D representation without explicit order dependence. To overcome the absence of reliable camera extrinsics in large-scale robotic datasets, the framework integrates geometric distillation from a pre-trained foundation model [57], providing essential structural regularization for pose-free novel view synthesis.

**Permutation-Equivariant Architecture.** Each image is first processed by DINOv2 [29] to generate patch tokens $\{\boldsymbol{h}_i\}_{i=1}^{V} \in \mathbb{R}^{L \times C}$. To establish a multi-view context without imposing a rigid temporal or spatial order, these tokens are aggregated through alternating intra-frame and global self-attention layers [54]. This symmetric aggregation mechanism ensures the extracted representations are strictly permutation-equivariant. The tokens are subsequently routed to task-specific transformer decoders. For the $i$-th image, a camera pose decoder $D_{\text{c}}$ predicts the local-to-global spatial transformation $\boldsymbol{p}_i = (\boldsymbol{R}_i, \boldsymbol{T}_i) \in \text{SE}(3)$, where $\boldsymbol{R}_i \in \text{SO}(3)$ denotes the camera rotation and $\boldsymbol{T}_i \in \mathbb{R}^3$ is the translation vector [10]. Concurrently, a point map decoder $D_{\text{p}}$ regresses a dense 3D point map $\boldsymbol{P}_i \in \mathbb{R}^{H \times W \times 3}$ representing the local camera-coordinate geometry, along with a per-pixel confidence map $\boldsymbol{C}_i \in \mathbb{R}^{H \times W \times 1}$.

**Gaussian Parameter Prediction.** We explicitly decouple spatial localization from the prediction of rendering attributes. The spatial centers of the Gaussian primitives $\boldsymbol{\mu}_g \in \mathbb{R}^3$ are derived via a deterministic coordinate transformation: $\boldsymbol{\mu}_g = \boldsymbol{R}_i \boldsymbol{P}_i + \boldsymbol{T}_i$. Parameterizing the scene geometry through dense

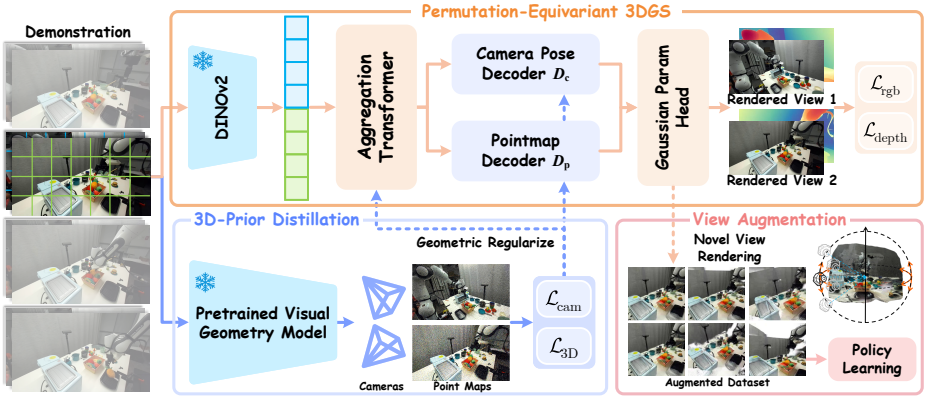


**Fig. 2: Overview of *GenSplat*.** Our feed-forward 3DGS framework reconstructs a 3D scene from sparse, uncalibrated robotic observations. GenSplat employs a permutation-equivariant transformer architecture to predict camera poses, dense point maps, and Gaussian parameters, while leveraging pre-trained visual geometry models for 3D-prior distillation supervision. The reconstructed scenes enable geometrically consistent novel view synthesis to expand the observational manifold, significantly improving the viewpoint generalization of robotic policies.

point maps rather than depth maps inherently preserves the local metric structure and mitigates topological distortions [37, 44, 46, 56], as visualized in Fig. 7.

For the remaining rendering attributes, a Dense Prediction Transformer (DPT) head $H_{\text{G}}$ is utilized [38]. This head integrates the deep equivariant tokens $F_{\text{DPT}}(\boldsymbol{h}_i)$ with high-resolution shallow features $F_{\text{RGB}}(I_i)$ to regress opacity $\sigma_g$, quaternion rotation $\boldsymbol{r}_g$, anisotropic scales $\boldsymbol{s}_g$, and spherical harmonic coefficients $\boldsymbol{c}_g$. Formulated as:

$$\begin{cases} (\boldsymbol{P}_i, \boldsymbol{C}_i) = D_{\text{p}}(\boldsymbol{h}_i), \ \boldsymbol{\mu}_g = \text{trans}(\boldsymbol{P}_i, \boldsymbol{p}_i), \\ (\sigma_g, \boldsymbol{r}_g, \boldsymbol{s}_g, \boldsymbol{c}_g) = H_{\text{G}}(F_{\text{DPT}}(\boldsymbol{h}_i) + F_{\text{RGB}}(I_i)). \end{cases} \tag{1}$$

### 3.3 Pose-Free Pre-training via 3D-Prior Distillation

Directly deploying feed-forward 3D Gaussian Splatting in the robotic domain is frequently hindered by severe ghosting and rendering artifacts. These issues stem from the inability of pure photometric optimization to preserve strict geometric consistency [60], a critical prerequisite for reliable robotic manipulation. Compounding this challenge, existing large-scale robotic datasets are typically characterized by noisy or entirely missing camera pose annotations.

To address this, we introduce a 3D-prior distillation strategy that extracts geometric knowledge from a pre-trained visual geometry foundation model [57], providing essential structural supervision for pose-free novel view synthesis. Specifically, to parameterize these spatial priors, we regress per-pixel 3D coordinates in the camera frame using pseudo-ground-truth from the foundation model, as detailed in Fig. 2. The point map loss $\mathcal{L}_{\text{3D}}$ is then computed as the

mean squared Euclidean distance between the predicted geometry $\hat{\boldsymbol{P}}$ and the target $\boldsymbol{P}$:

$$\mathcal{L}_{\text{3D}} = \frac{1}{HW} \sum_{u,v} \left\| \hat{\boldsymbol{P}}(u,v) - \boldsymbol{P}(u,v) \right\|_2^2, \tag{2}$$

where $(u, v)$ denotes the spatial pixel coordinates. As empirically visualized in Fig. 6, this explicit 3D supervision significantly reduces novel-view artifacts, such as floaters and boundary tearing, in Gaussian rendering. It markedly outperforms both depth-based 2.5D supervision [19,47] and RGB-only baselines [13, 60]. The resulting multi-view geometric coherence yields clean, low-variance novel-view training data. This metric stability is essential for reliably expanding the observational manifold and enhancing view-generalized execution for downstream policies.

To construct a consistent global coordinate frame without absolute tracking data, we distill camera-pose priors to obtain relative affine-invariant transformations. For the relative pose between views $i$ and $j$, rotational alignment is supervised via the geodesic distance $\ell_{\text{rot}}$ on the $SO(3)$ manifold [7], and translational alignment is constrained using a robust Huber loss $\mathcal{H}_\delta$ applied to the relative translation vectors:

$$\mathcal{L}_{\text{cam}} = \sum_{i \neq j} \Big( \ell_{\text{rot}}(\hat{\boldsymbol{R}}_{i \leftarrow j}, \boldsymbol{R}_{i \leftarrow j}) + \mathcal{H}_\delta(\hat{\boldsymbol{T}}_{i \leftarrow j} - \boldsymbol{T}_{i \leftarrow j}) \Big). \tag{3}$$

This affine-invariant pose supervision exploits the inherent low-dimensional manifold of real-world camera trajectories [57], regularizing geometry, suppressing drift and degeneracies, and stabilizing novel-view synthesis from sparse inputs. The complete loss $\mathcal{L}_{\text{3D-prior}}$ is defined as the weighted sum of the point map loss and the camera pose loss:

$$\mathcal{L}_{\text{3D-prior}} = \lambda_{\text{3D}} \mathcal{L}_{\text{3D}} + \lambda_{\text{cam}} \mathcal{L}_{\text{cam}}. \tag{4}$$

Ultimately, the proposed distillation strategy yields substantial advantages for both scene reconstruction and policy learning. Fundamentally, point maps serve as a 2D parameterization of 3D features, coupling metric precision with structural regularity to provide an optimal supervisory signal. Unlike traditional depth-based pseudo-labels that exhibit sharp boundary discontinuities and cause tearing artifacts [19], dense point maps explicitly enforce spatial smoothness and topological coherence. Distilling these priors imposes robust structural constraints, accelerating convergence and producing the high-fidelity representations essential for reliable robotic manipulation.

### 3.4 Policy Deployment

We instantiate our view-generalized framework across two representative imitation learning architectures: $\pi_0$ [3] and Diffusion Policy [6]. Both models are optimized over the expanded observational manifold $\mathcal{D}_{\text{train}}$. Formally, the policy $\pi_\theta$ processes a visual observation $\boldsymbol{o}$ alongside the robot's proprioceptive state to

predict a sequence of future kinematic actions $\boldsymbol{a}$. The learning objective follows a standard behavioral cloning paradigm, maximizing the log-likelihood of expert demonstrations under the parameterized policy:

$$\max_{\theta} \sum_{(\boldsymbol{o},\boldsymbol{a})\in\mathcal{D}_{\text{train}}} \log \pi_\theta(\boldsymbol{a} \mid \boldsymbol{o}), \tag{5}$$

where $\theta$ denotes the learnable parameters of the policy network. Comprehensive training configurations and architectural details are deferred to Appendix A.1.

### 3.5 Training Setup

**Training Objectives.** GenSplat is optimized end-to-end by minimizing a synergistic combination of photometric, geometric, and distillation objectives:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{3D-prior}}. \tag{6}$$

For each input view, the photometric rendering pipeline is strictly supervised via a combination of an $\ell_1$ penalty and a perceptual LPIPS regularizer [63]:

$$\mathcal{L}_{\text{rgb}} = \lambda_1 \|I - \hat{I}\|_1 + \lambda_2\, \texttt{LPIPS}(I, \hat{I}), \tag{7}$$

where $\hat{I}$ denotes the novel view synthesized via the differentiable splatting process. To mitigate the subtle ambiguities that arise during multi-view alignment and aggregation [19], we introduce a depth alignment loss that enforces alignment between the Gaussian primitives and the predicted point map, promoting geometric stability and mitigating view inconsistency artifacts:

$$\mathcal{L}_{\text{depth}} = \frac{1}{\|M\|_1} \left\| M \odot \left(\boldsymbol{P}_{\dots,\,2} - \hat{D}\right) \right\|_2^2, \tag{8}$$

where $M$ is a binary mask indicating valid pixels, $\hat{D}$ denotes the rendered depth, and the predicted depth is obtained from the z-component of the point map.

**Implementation Details.** The aggregation transformer contains 36 alternating attention blocks that interleave frame-wise and global attention, and camera pose, point map decoders are lightweight 5-layer Transformer modules that use self-attention only. All these layers together with the prediction decoder introduced in Sec. 3.2 are initialized from [57] so that the model inherits permutation equivariant geometric priors and attains stable early training. Moreover, to narrow the gap between general scenes and robot manipulation, we pretrain on a large-scale dataset collected in the wild [23]. We temporally subsample every episode at 1 frame per second, which yields 271k training images that cover a wide range of environments and tasks. We train for 30k iterations with a batch size of 16 and maintain an exponential moving average of the model weights to stabilize evaluation. The learning rate follows a cosine annealing schedule with a 1k step linear warmup, and the peak learning rate equals $2e^{-4}$. Following [19], parameters initialized from [57] use a learning rate of 0.1 times the base value, and we enable FlashAttention [8], gradient checkpointing, and bfloat16 precision to reduce memory usage and accelerate training.

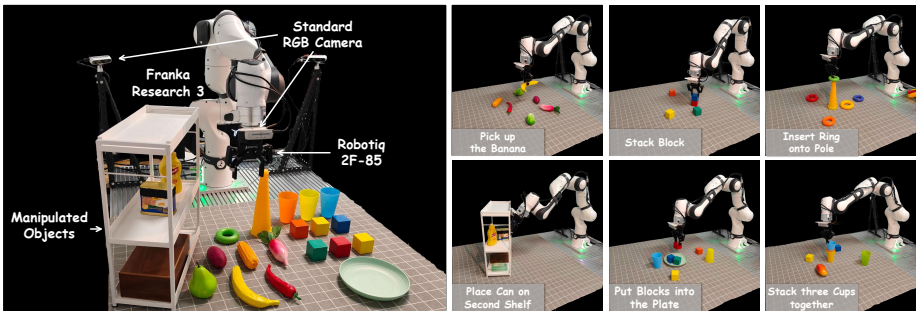


**Fig. 3: Overview of the experiment setup and tasks.** We design six manipulation tasks for real-world evaluation.

## 4 Experiments

We conduct extensive experiments on a real-world robotic platform to evaluate the performance of our proposed method, GenSplat. Our evaluation is structured around four central research questions (RQs) designed to assess our contributions systematically:

RQ1: How effectively does expanding the observational manifold via GenSplat improve policy robustness to out-of-distribution camera perturbations?
RQ2: How does GenSplat compare against diffusion-based and gaussian-based NVS methods in terms of rendering quality, policy learning, and efficiency?
RQ3: How does the core 3D-prior distillation component impact geometric consistency, rendering fidelity, and the success rate of downstream policies?
RQ4: What is the impact of increasing the number of rendered views per trajectory on view generalization and data efficiency?

**Experimental Setup.** Our hardware suite (Fig. 3) comprises a 7-DoF Franka Research 3 arm equipped with a Robotiq 2F-85 gripper. Visual observations are captured by three standard RGB cameras: two providing external views and one mounted on the robot's wrist. We evaluate our framework across six manipulation tasks. For each task, we collect 100 expert demonstrations via teleoperation. To ensure rigorous evaluation, the dataset incorporates rich spatial and visual diversity through systematic variations in object positions and colors.

### 4.1 RQ1: Robustness to Camera Perturbations

This experiment quantitatively investigates whether densifying the spatial support via GenSplat systematically improves policy robustness against camera perturbations of varying magnitudes.
**Viewpoint re-rendering strategy.** We employ GenSplat to reconstruct the scene and synthesize novel-view trajectories while strictly preserving temporal alignment and kinematic action labels. For each demonstration, we define the pivot point as the midpoint of the shortest line segment connecting the optical

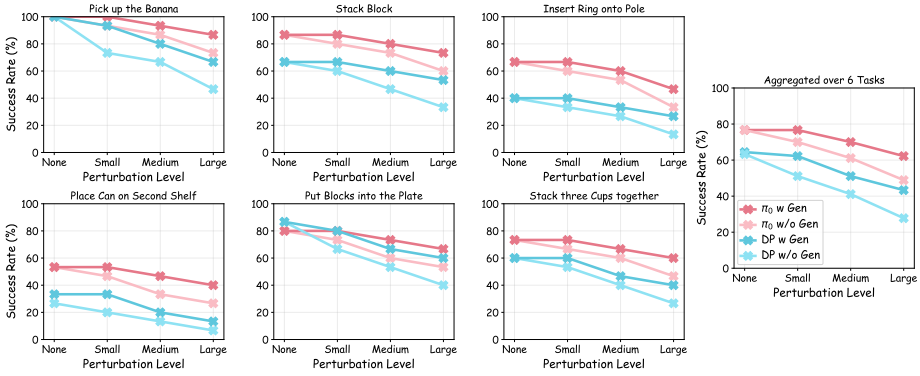


**Fig. 4: Policy robustness to camera pose perturbations.** We evaluate 30 episodes per task on 6 real-world tasks under increasing camera perturbations and report mean success rates (in %). Policies trained with GenSplat augmented novel views data consistently outperform those trained only on source views for both $\pi_0$ and Diffusion Policy.

axes of the two external cameras. Following [45,50], novel views are generated by rotating the camera about a designated world axis ($\mathcal{X}$ or $\mathcal{Y}$) by $\theta \in [-30°, 30°]$, followed by a forward translation $\mu \in [0.5, 6]$ cm along the camera's $+\mathcal{Z}$ axis. This translation crucially prevents severe rotations from pushing the viewpoint outside the densified spatial support (*i.e.*, beyond valid splat coverage). For each original demonstration, we sample exactly one unique perturbation pair $(\theta, \mu)$ to generate a single augmented trajectory.

**Perturbation analysis protocol.** We deploy the trained policies under three controlled spatial perturbation regimes during physical rollouts: (1) **Small**: Rotation $\theta_{\max} \in [3°, 6°]$ and translation $\mu \in [0.5, 2]$ cm. (2) **Medium**: Rotation $\theta_{\max} \in [8°, 15°]$ and translation $\mu \in [2, 4]$ cm. (3) **Large**: Rotation $\theta_{\max} \in [18°, 30°]$ and translation $\mu \in [4, 6]$ cm. We compare baselines trained strictly on source data against our policies trained on the expanded manifold.

As visualized in Fig. 4, GenSplat augmentation yields consistent and substantial improvements in policy robustness across all six manipulation tasks and all perturbation levels, mitigating the sharp degradation typically caused by out-of-distribution camera viewpoints. Aggregated across all tasks, GenSplat yields substantial relative performance improvements across all perturbation levels for $\pi_0$: +**6.7**% for small, +**8.9**% for medium, and +**13.3**% for large perturbations. The benefit is most pronounced for DP under large perturbations, where source-view-only policies suffer significant collapse, and GenSplat achieves a remarkable **56.0**% relative gain (from 27.78% to **43.33**%). Crucially, the augmented policies maintain parity with the baselines in the unperturbed (source) setting, confirming that our geometrically consistent novel views do not compromise the original data distribution. These results establish GenSplat as an effective, architecture-agnostic strategy for training viewpoint generalizable policies.

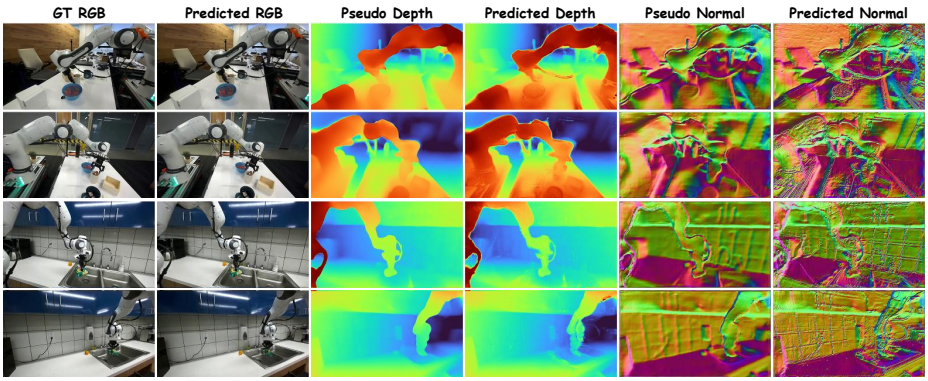


**Fig. 5: Qualitative 3D geometry reconstruction on the DROID dataset.** We visualize pairs of reference targets (ground truth or pseudo-labels) against GenSplat's predicted RGB, depth, and normal maps. GenSplat accurately recovers structural high-frequency details and maintains strict spatial alignment without explicit camera calibration. Extended visual comparisons are provided in Appendix B.

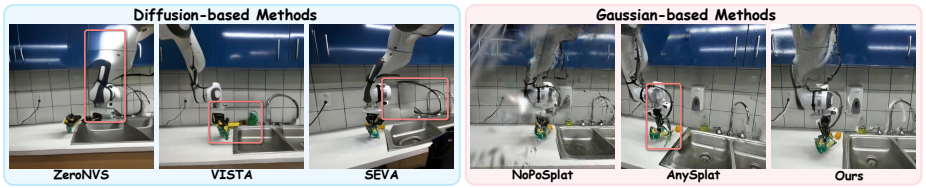


**Fig. 6: Qualitative comparison of NVS.** We compared GenSplat with diffusion- and Gaussian-based methods under the same input example. GenSplat shows a clear advantage in 3D geometric consistency modeling while effectively eliminating floating artifacts and visual artifacts. Note: VISTA is built on ZeroNVS and is fine-tuned on a 3k trajectory subset of DROID [23], yielding significant improvements in NVS.

### 4.2 RQ2: Comprehensive Comparison with SOTA

We address RQ2 by evaluating GenSplat across three critical axes: (1) rendering quality, (2) downstream policy performance, and (3) synthesis efficiency. We compare GenSplat against leading NVS methods including ZeroNVS [40], VISTA [50], SEVA [67], InstantSplat [13], NoPoSplat [60] and AnySplat [19].

**Geometric Consistency & Rendering Quality.** We first validate the underlying 3D structural fidelity. As illustrated in Fig. 5, GenSplat accurately infers dense depth and normal maps. While the distilled pseudo-labels inherently suffer from over-smoothing, our unified photometric and geometric optimization effectively recovers sharp object boundaries and fine-grained topologies. By leveraging RGB photometric supervision with 3D-priors strictly acting as structural regularizers, GenSplat successfully reconstructs high-frequency details.

Building upon this robust geometry, GenSplat demonstrates a significant advantage in synthesizing high-fidelity novel views. As depicted in Fig. 6, diffusion-based baselines frequently violate strict epipolar geometry, resulting in severe spatial hallucinations and metric distortions (*e.g.*, the bent faucet). Conversely,

| Method / Task | Avg. Success↑ | Pick up Banana | Stack Block | Insert Ring | Place Can | Put Blocks into Plate | Stack Cups | Synthesis Speed |
|---|---|---|---|---|---|---|---|---|
| Baseline | 48.9 | 73.3 | 60.0 | 33.3 | 26.7 | 53.3 | 46.7 | - |
| VISTA [50] | 37.8 | 76.7 | 43.3 | 36.7 | 6.7 | 40.0 | 23.3 | 2.81 s |
| SEVA [67] | 57.7 | 83.3 | 70.0 | 43.3 | 30.0 | 63.3 | 56.7 | 50.00 s |
| InstantSplat [13] | 55.0 | 80.0 | 66.7 | 40.0 | 33.3 | 60.0 | 50.0 | 85.20 s |
| NoPoSplat [60] | 36.7 | 73.3 | 40.0 | 33.3 | 10.0 | 36.7 | 26.7 | **0.05** s |
| AnySplat [19] | 56.1 | 80.0 | 66.7 | 43.3 | 30.0 | 63.3 | 53.3 | 0.16 s |
| GenSplat(ours) | **62.2** | **86.7** | **73.3** | **46.7** | **40.0** | **66.7** | **60.0** | 0.14 s |

**Table 1: Policy Performance and Synthesis Efficiency.** We compare various NVS methods by training $\pi_0$ on augmented datasets and report mean success rate and synthesis speed. Each task was evaluated 30 episodes under the **Large** perturbation level. Synthesis speed evaluated on the same computing device.

conventional Gaussian-based methods lacking prior distillation are plagued by severe floater artifacts and boundary tearing [60], fundamentally degrading visual quality. In stark contrast, GenSplat renders scenes with photorealism and absolute structural integrity. This geometrically consistent, high-fidelity synthesis is paramount, providing the clean, low-variance multi-view data necessary for reliable downstream policy learning.

**Downstream Policy Performance.** As detailed in Tab. 1, GenSplat achieves the highest average success rate of **62.2**% under large camera perturbations, significantly outperforming all competing NVS paradigms. Crucially, the results reveal that low-quality data augmentation actively harms policy learning: both the diffusion-based VISTA [50] (37.8%) and the standard feed-forward NoPoSplat [60] (36.7%) severely degrade performance below the unaugmented Baseline (48.9%). Their pronounced geometric distortions and rendering artifacts act as noise, contaminating the visuomotor mapping (see Fig. 6). While SEVA [67] (57.7%) and InstantSplat [13] (55.0%) offer improvements, their sparse-view reconstructions still exhibit localized holes and inconsistencies that bottleneck execution accuracy. Compared to the closest feed-forward baseline, AnySplat [19] (56.1%), GenSplat provides a substantial absolute gain of +**6.1**%. This confirms that our 3D-prior distillation explicitly enforces the spatial smoothness necessary to generate physically reliable, high-fidelity training signals.

**Synthesis Efficiency.** For scalable data augmentation in robotics, computational efficiency is a hard constraint. GenSplat operates at a highly efficient 0.14 seconds per frame, completely bypassing the prohibitive computational bottlenecks of iterative diffusion sampling and per-scene optimization. While NoPoSplat [60] is marginally faster, this speed incurs the unacceptable cost of catastrophic policy degradation. Ultimately, GenSplat achieves remarkable generation speed without sacrificing geometric consistency, positioning our framework as a highly scalable engine for learning view-generalized robotic policies.

### 4.3 RQ3: Ablation of 3D-Prior Distillation

To address RQ3, we conduct an ablation study isolating the contributions of the permutation-equivariant (P.E.) architecture [57] and the core 3D-prior distillation module. Tab. 2 details the quantitative impact of these components

| VGGT | P.E. [57] | 3D-Prior | PSNR↑ | LPIPS↓ | 30°, 6cm | 60°, 12cm |
|---|---|---|---|---|---|---|
| ✓ | ✗ | ✗ | 22.35 | 6.01 | 66.0% | 30.0% |
| ✗ | ✓ | ✗ | 25.87 | 5.03 | 70.0% | 40.0% |
| ✗ | ✓ | ✓ | **26.53** | **3.72** | **74.0%** | **46.0%** |

**Table 2: Ablation of GenSplat components.** We evaluate the impact of the permutation-equivariant (P.E.) architecture and 3D-prior distillation on novel view rendering quality (PSNR, LPIPS) and policy average success rates under **Large** (30°, 6cm) and **Extreme** (60°, 12cm) camera perturbations.

on both novel view synthesis fidelity and downstream policy robustness under challenging viewpoint shifts. Replacing the baseline VGGT with the P.E. architecture improves reconstruction. By mitigating order-dependent artifacts from uncalibrated inputs, it boosts policy success from 66.0% to 70.0% under large perturbations, and 30.0% to 40.0% under extreme shifts. Furthermore, integrating 3D-prior distillation provides essential geometric regularization for pose-free settings, achieving the best PSNR and LPIPS. This strict structural consistency translates directly to physical robustness, maximizing policy success to **74.0%** (large) and **46.0%** (extreme). By explicitly suppressing floating artifacts, the 3D prior prevents geometric collapse, delivering the high-fidelity manifold data necessary to induce view-generalized behaviors.

As illustrated in Fig. 7, we utilize 3D point maps as structural pseudo-labels, unlike AnySplat [19], which relies on scalar depth maps. Depth-based supervision often fails to maintain rigorous multi-view consistency, rendering the model incapable of reconstructing flat regions and leading to severe structural collapse of the desk. By directly encoding explicit 3D coordinates, point maps offer higher precision and enforce spatial smoothness [44, 56]. This effectively prevents geometric collapse, recovering a remarkably accurate and flat desk surface. More visual examples are in Appendix B.2.

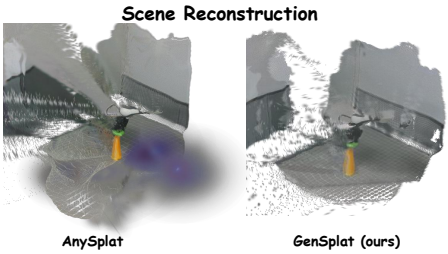


**Fig. 7: Qualitative scene reconstruction comparison.** AnySplat (left) exhibits geometric irregularities on flat surfaces. GenSplat (right) reconstructs accurate, smooth 3D geometry.

### 4.4 RQ4: Impact of the Number of Rendered Views

For RQ4, we quantify the trade-off between the number of synthesized novel views and policy generalization. Fixing the source trajectories at 100, we systematically vary the view augmentation factor $N \in \{1, 2, 3, 4\}$ and evaluate the resulting Diffusion Policies [6]. According to Fig. 8, under a **Medium** perturbation regime, increasing the number of rendered views consistently improves policy success rates. Notably, the most substantial performance leap occurs immediately upon introducing the first augmented view ($N = 1$).

**Analysis of Data Efficiency.** Critically, performance gains saturate rapidly. Tab. 3 extends this analysis across all perturbation severities on the `Stack Block`

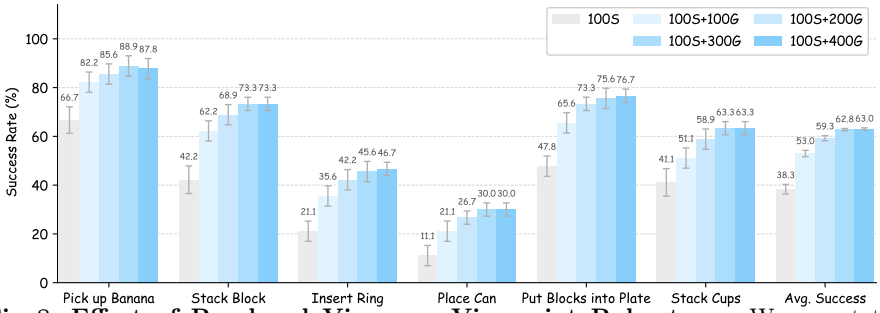


**Fig. 8: Effect of Rendered Views on Viewpoint Robustness.** We report the mean success rate and standard deviation across three independent trials, with 30 execution episodes per task in each run.

| Task | Stack Block | | | | Stack Cups | | | |
|---|---|---|---|---|---|---|---|---|
| Perturbation | None | Small | Mediun | Large | None | Small | Mediun | Large |
| 100S (Baseline) | 68.0 | 58.0 | 42.0 | 30.0 | 60.0 | 50.0 | 40.0 | 26.0 |
| 100S+100G | 70.0 | 66.0 | 62.0 | 48.0 | 62.0 | 58.0 | 50.0 | 44.0 |
| 100S+200G | 68.0 | 70.0 | 68.0 | 58.0 | 60.0 | 62.0 | 58.0 | 50.0 |
| 100S+300G | **72.0** | **74.0** | **74.0** | **64.0** | **64.0** | 64.0 | **64.0** | **56.0** |
| 100S+400G | 70.0 | 74.0 | 72.0 | 62.0 | 62.0 | **66.0** | 64.0 | 54.0 |

**Table 3: Cross-Perturbation Analysis.** We report the success rates for Diffusion Policy with varying GenSplat view augmentations, evaluated on two tasks across perturbation levels from None to Large, based on 50 episodes per task.

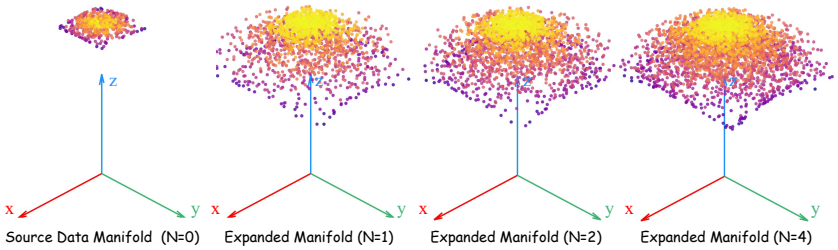


**Fig. 9: Densification of camera extrinsics on the observational manifold.**

and `Stack Cups` tasks, revealing a consistent scaling law: after a sharp initial leap at $N = 1$, marginal benefits diminish and saturate near $N = 3$. Since $N = 4$ yields negligible improvements, a moderate augmentation density ($N = 3$) is proven sufficient to unlock robust view generalization, effectively circumventing the overhead of excessive rendering.

**Geometric Interpretation.** To geometrically explain this saturation, Fig. 9 visualizes a single camera's extrinsics across 600 real-world demonstrations. The initial augmentation ($N = 1$) sharply expands the sparse source distribution ($N = 0$), driving the primary performance leap. Subsequent views ($N \in \{2, 4\}$) merely densify this expanded spherical manifold without enlarging its boundary. Once topological continuity is achieved (near $N = 3$), the policy interpolates robustly, rendering further augmentations computationally redundant.

## 5 Conclusion and Discussion

We introduced GenSplat, a feed-forward 3D Gaussian Splatting framework synthesizing geometrically consistent novel views from sparse, uncalibrated observations. By explicitly incorporating 3D-prior distillation, GenSplat eliminates the structural artifacts inherent to pure photometric optimization. Driven by its feed-forward design, this high-fidelity and efficient synthesis provides a scalable data augmentation paradigm, expanding the observational manifold to resolve the critical bottleneck of policy view generalization. Future work will explore integrating dynamic wrist-camera observations to capture fine-grained, occlusion-aware geometries during active manipulation.

## Acknowledgement

This work was supported in part by the National Key Research and Development Project under Grant 2024YFB4708100, the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM504, National Natural Science Foundation of China under Grants 62572384, U24A20325, 62125305, 62503384, and Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80.

## References


1. Abouzeid, A., Mansour, M., Sun, Z., Song, D.: GeoAware-VLA: Implicit geometry aware vision-language-action model. arXiv preprint arXiv:2509.14117 (2025)
2. Bai, Y., Wang, Z., Liu, Y., Chen, W., Chen, Z., Dai, M., Zheng, Y., Liu, L., Li, G., Lin, L.: Learning to see and act: Task-aware view planning for robotic manipulation. arXiv preprint arXiv:2508.05186 (2025)
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: $\pi_0$: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. arXiv preprint arXiv.2410.24164 (2024)
4. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
5. Chen, L.Y., Xu, C., Dharmarajan, K., Irshad, M.Z., Cheng, R., Keutzer, K., Tomizuka, M., Vuong, Q., Goldberg, K.: Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. CoRL (2024)
6. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion Policy: Visuomotor policy learning via action diffusion. IJRR (2023)
7. Chowdhury, S.B.R., Monath, N., Dubey, A., Ahmed, A., Chaturvedi, S.: Unsupervised opinion summarization using approximate geodesics. arXiv preprint arXiv:2209.07496 (2022)
8. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: NeurIPS (2022)
9. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C.: Robonet: Large-scale multi-robot learning. arXiv preprint arXiv:1910.11215 (2019)
10. Dong, S., Wang, S., Liu, S., Cai, L., Fan, Q., Kannala, J., Yang, Y.: Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In: CVPR. pp. 16739–16752 (2025)
11. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: PaLM-E: An embodied multimodal language model. In: ICML. pp. 8469–8488. PMLR (2023)
12. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In: ACM SIGGRAPH. pp. 1–11 (2024)
13. Fan, Z., Cong, W., Wen, K., Wang, K., Zhang, J., Ding, X., Xu, D., Ivanovic, B., Pavone, M., Pavlakos, G., et al.: InstantSplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. CoRL (2024)
14. Gao, J., Xie, A., Xiao, T., Finn, C., Sadigh, D.: Efficient data collection for robotic manipulation via compositional generalization. RSS (2024)
15. Goyal, A., Blukis, V., Xu, J., Guo, Y., Chao, Y.W., Fox, D.: RVT2: Learning precise manipulation from few demonstrations. RSS (2024)
16. Han, X., Liu, M., Chen, Y., Yu, J., Lyu, X., Tian, Y., Wang, B., Zhang, W., Pang, J.: Re$^3$Sim: Generating high-fidelity simulation data via 3d-photorealistic real-to-sim for robotic manipulation. arXiv preprint arXiv:2502.08645 (2025)
17. Huang, S., Chen, L., Zhou, P., Chen, S., Jiang, Z., Hu, Y., Liao, Y., Gao, P., Li, H., Yao, M., et al.: Enerverse: Envisioning embodied future space for robotics manipulation. arXiv preprint arXiv:2501.01895 (2025)

18. Iyer, A., Peng, Z., Dai, Y., Guzey, I., Haldar, S., Chintala, S., Pinto, L.: Open teach: A versatile teleoperation system for robotic manipulation. CoRL (2024)
19. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: AnySplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM SIGGRAPH (2025)
20. Jiang, T., Ji, J., Tan, X., Fang, J., Bhattad, A., Guizilini, V., Walter, M.R.: Do you know where your camera is? view-invariant policy learning with camera conditioning. arXiv preprint arXiv:2510.02268 (2025)
21. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: MapAnything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
22. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139–1 (2023)
23. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945 (2024)
24. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: OpenVLA: An open-source vision-language-action model. CoRL (2024)
25. Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., Zeng, L., Li, H.: Spatial forcing: Implicit spatial representation alignment for vision-language-action model. arXiv preprint arXiv:2510.12276 (2025)
26. Li, X., Li, J., Zhang, Z., Zhang, R., Jia, F., Wang, T., Fan, H., Tseng, K.K., Wang, R.: RoboGSim: A real2sim2real robotic gaussian splatting simulator. arXiv preprint arXiv:2411.11839 (2024)
27. Lin, T., Li, G., Zhong, Y., Zou, Y., Du, Y., Liu, J., Gu, E., Zhao, B.: Evo-0: Vision-language-action model with implicit spatial understanding. arXiv preprint arXiv:2507.00416 (2025)
28. Lu, R., Chen, Y., Ni, J., Jia, B., Liu, Y., Wan, D., Zeng, G., Huang, S.: Movis: Enhancing multi-object novel view synthesis for indoor scenes. In: CVPR. pp. 26767–26778 (2025)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
30. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: ICRA. pp. 6892–6903. IEEE (2024)
31. Pan, C., Liang, L., Bauer, D., Cousineau, E., Burchfiel, B., Feng, S., Song, S.: One demo is worth a thousand trajectories: Action-view augmentation for visuomotor policies. CoRL (2025)
32. Pang, J.C., Tang, N., Li, K., Tang, Y., Cai, X.Q., Zhang, Z.Y., Niu, G., Sugiyama, M., Yu, Y.: Learning view-invariant world models for visual robotic manipulation. In: ICLR (2025)
33. Papantonakis, P., Kopanas, G., Kerbl, B., Lanvin, A., Drettakis, G.: Reducing the memory footprint of 3d gaussian splatting. PACM CGIT **7**(1) (May 2024), https://repo-sam.inria.fr/fungraph/reduced_3dgs/
34. Patil, V., Hutter, M.: Radiance fields for robotic teleoperation. In: IROS. pp. 13861–13868. IEEE (2024)

35. Pumacay, W., Singh, I., Duan, J., Krishna, R., Thomason, J., Fox, D.: The colosseum: A benchmark for evaluating generalization for robotic manipulation. RSS (2024)
36. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: SpatialVLA: Exploring spatial representations for visual-language-action model. RSS (2025)
37. Ramamonjisoa, M., Du, Y., Lepetit, V.: Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In: CVPR. pp. 14648–14657 (2020)
38. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: CVPR. pp. 12179–12188 (2021)
39. Sadeghi, F., Toshev, A., Jang, E., Levine, S.: Sim2real viewpoint invariant visual servoing by recurrent control. In: CVPR. pp. 4691–4699 (2018)
40. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single image. In: CVPR. pp. 9420–9429 (2024)
41. Saxena, V., Bronars, M., Arachchige, N.R., Wang, K., Shin, W.C., Nasiriany, S., Mandlekar, A., Xu, D.: What matters in learning from large-scale datasets for robot manipulation. ICLR (2025)
42. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. NeurIPS **37**, 80220–80243 (2024)
43. Seo, Y., Kim, J., James, S., Lee, K., Shin, J., Abbeel, P.: Multi-view masked world models for visual robotic manipulation. In: ICML. pp. 30613–30632. PMLR (2023)
44. Shi, D., Wang, W., Chen, D.Y., Zhang, Z., Bian, J.W., Zhuang, B., Shen, C.: Revisiting depth representations for feed-forward 3d gaussian splatting. arXiv preprint arXiv:2506.05327 (2025)
45. Shi, L., Shi, Y., Xu, X., Liu, T., Xi, J., Chen, C.: Nvspolicy: Adaptive novel-view synthesis for generalizable language-conditioned policy learning. arXiv preprint arXiv:2505.10359 (2025)
46. Sun, L., Bian, J.W., Zhan, H., Yin, W., Reid, I., Shen, C.: Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. IEEE TPAMI **46**(1), 497–508 (2023)
47. Sun, X., Jiang, H., Liu, L., Nam, S., Kang, G., Wang, X., Sui, W., Su, Z., Liu, W., Wang, X., et al.: Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. arXiv preprint arXiv:2508.03643 (2025)
48. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 (2024)
49. Tian, J., Wang, L., Zhou, S., Wang, S., Li, J., Sun, H., Tang, W.: PDFactor: Learning tri-perspective view policy diffusion field for multi-task robotic manipulation. In: CVPR. pp. 15757–15767 (2025)
50. Tian, S., Wulfe, B., Sargent, K., Liu, K., Zakharov, S., Guizilini, V., Wu, J.: View-invariant policy learning via zero-shot novel view synthesis. CoRL (2024)
51. Torne, M., Jain, A., Yuan, J., Macha, V., Ankile, L., Simeonov, A., Agrawal, P., Gupta, A.: Robot learning with super-linear scaling. RSS (2024)
52. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: ECCV. pp. 313–331. Springer (2024)

53. Walke, H.R., Black, K., Zhao, T.Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A.W., Myers, V., Kim, M.J., Du, M., et al.: Bridgedata v2: A dataset for robot learning at scale. In: CoRL. pp. 1723–1736. PMLR (2023)
54. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025)
55. Wang, S., Wang, L., Zhou, S., Tian, J., Li, J., Sun, H., Tang, W.: FlowRAM: Grounding flow matching policy with region-aware mamba framework for robotic manipulation. In: CVPR. pp. 12176–12186 (2025)
56. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3R: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024)
57. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: $pi^3$: Scalable permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
58. Yang, S., Yu, W., Zeng, J., Lv, J., Ren, K., Lu, C., Lin, D., Pang, J.: Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. RSS (2025)
59. Yang, S., Ze, Y., Xu, H.: Movie: Visual model-based policy adaptation for view generalization. NeurIPS **36**, 21507–21523 (2023)
60. Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.H., Peng, S.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. ICLR (2025)
61. Yu, J., Fu, L., Huang, H., El-Refai, K., Ambrus, R.A., Cheng, R., Irshad, M.Z., Goldberg, K.: Real2render2real: Scaling robot data without dynamics simulation or robot hardware. CoRL (2025)
62. Zhang, D., Yuan, C., Wen, C., Zhang, H., Zhao, J., Gao, Y.: KineDex: Learning tactile-informed visuomotor policies via kinesthetic teaching for dexterous manipulation. CoRL (2025)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
64. Zhang, T., Duan, H., Hao, H., Qiao, Y., Dai, J., Hou, Z.: Grounding actions in camera space: Observation-centric vision-language-action policy. arXiv preprint arXiv:2508.13103 (2025)
65. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: CVPR. pp. 1702–1713 (2025)
66. Zheng, R., Liang, Y., Huang, S., Gao, J., Daumé III, H., Kolobov, A., Huang, F., Yang, J.: TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. ICLR (2025)
67. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489 (2025)
68. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL. pp. 2165–2183. PMLR (2023)