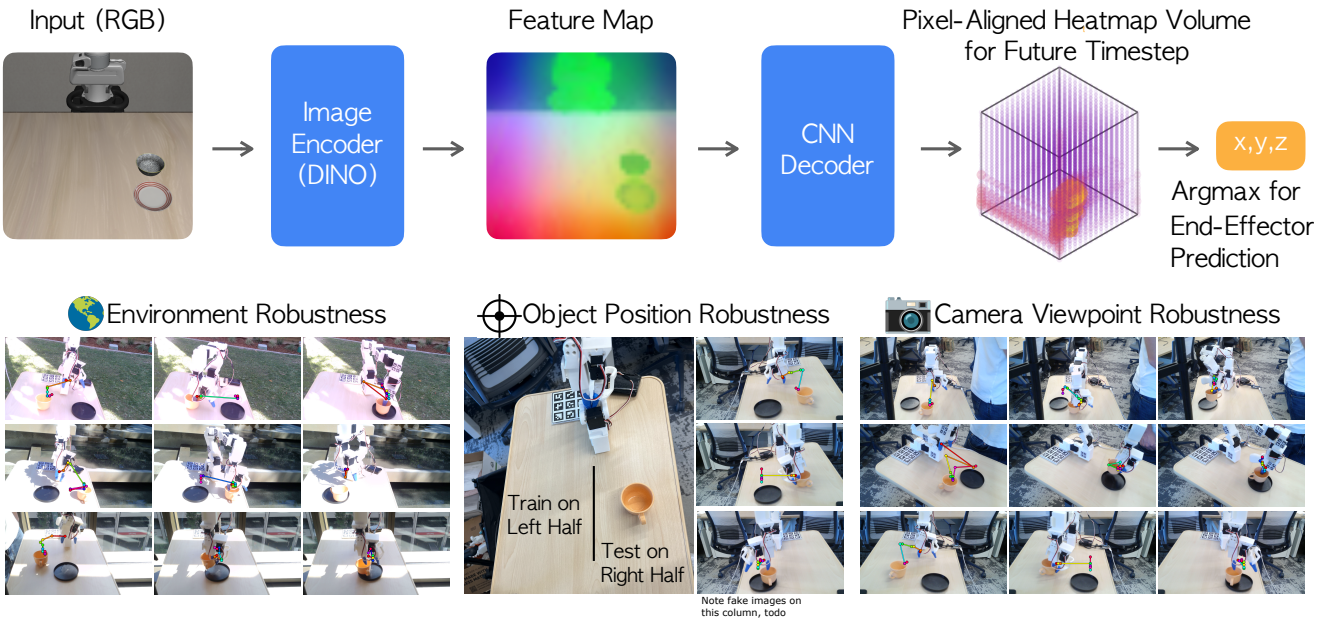

Figure 1: PARA overview. PARA reformulates end-effector action prediction as a per-pixel heatmap volume over the image. The same formulation transfers across object position, camera viewpoint, and environment changes.

PARA: Pixel-Aligned Robot Actions for Spatially Robust Manipulation

Anonymous Authors

Abstract

Visuomotor policies trained via behavioral cloning are brittle: modest changes in object placement or camera viewpoint cause dramatic failures, even when the task is unchanged. We identify action parameterization as a key contributor—regressing end-effector coordinates from a pooled image embedding discards spatial structure and couples the policy to viewpoint-specific cues. We propose PARA (Pixel-Aligned Robot Actions), which predicts actions as dense image-space classifications: a 2D heatmap identifies where the end-effector should project in the image, and per-pixel height-bin logits determine how high above the support surface. The 3D target is recovered by intersecting the camera ray with the predicted height plane. On a real SO-100 robot arm with only 20 demonstrations per task, PARA achieves 97% on pick-and-place (vs. 9% for coordinate regression), 97% on towel folding (vs. 11%), and 95% on table wiping (vs. 0%), while transferring zero-shot to new viewpoints (52% vs. 0%) and new environments (94% vs. 0%). In controlled LIBERO simulation, PARA achieves 54% on spatial extrapolation where coordinate regression scores 1%, and 61% on zero-shot viewpoint transfer (vs. 24%). Pixel alignment also makes video diffusion models effective action backbones (92% vs. 0% for global regression on identical features) and enables cross-embodiment transfer via point-track pretraining (66% vs. 10% from scratch at 10 demos).

1 Introduction

Visuomotor policies trained with behavioral cloning are notoriously brittle: translating an object a few centimeters or repositioning the camera can cause complete failure, even when the underlying task is unchanged. Foundation models such as DINOv2 now produce spatially rich, shift-equivariant features, yet most policy architectures discard this structure immediately: a global CLS token or spatial average is fed to an MLP that regresses end-effector poses in world coordinates. This forces the network to implicitly solve correspondence, geometry, and control in a single unstructured output space, encouraging shortcut solutions tied to absolute positions.

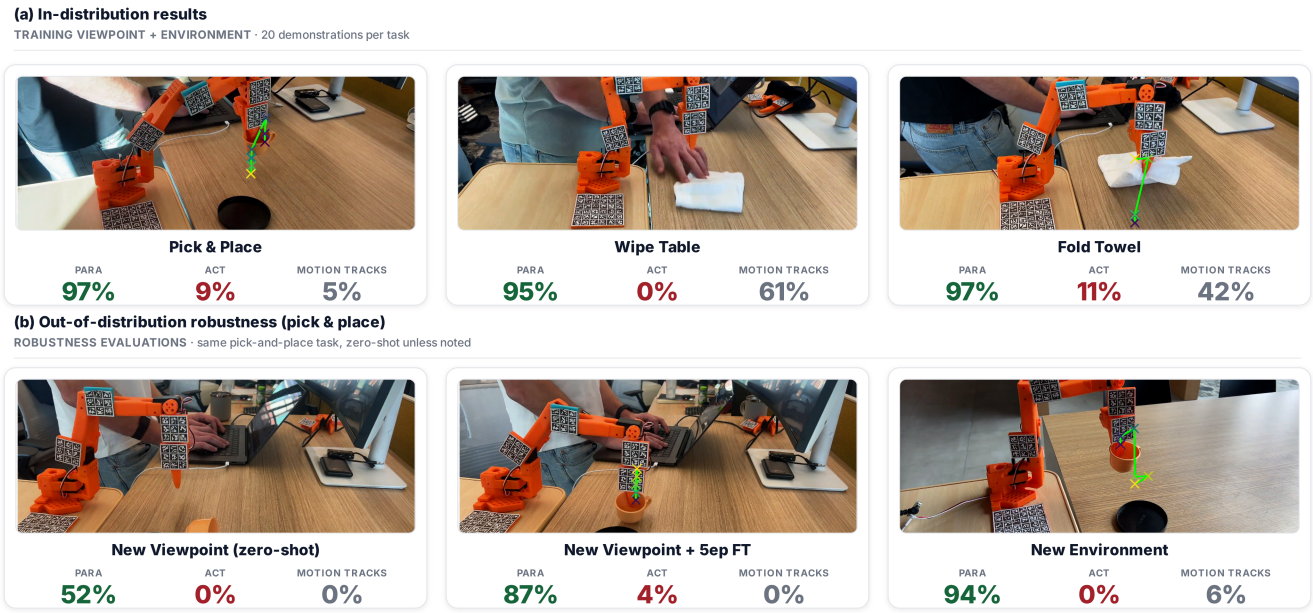

Figure 2: Real robot results (SO-100, 20 demos per task). (a) In-distribution performance: PARA achieves 95–97% across three tasks; ACT and Motion Tracks fail on precise manipulation. Motion Tracks achieves 61% on wipe table (a coarse trajectory task). (b) OOD robustness on pick-and-place: PARA transfers to new viewpoints (52% zero-shot, 87% with 5 fine-tuning demos) and new environments (94%); both baselines collapse.

Manipulation actions are fundamentally local in the image: "place the teacup" corresponds to a specific pixel region regardless of viewpoint. PARA (Pixel-Aligned Robot Actions) restores this locality by predicting a dense heatmap over pixels indicating where the end-effector should project, then classifying height bins at that pixel. The 3D target is recovered by intersecting the camera ray with the predicted height plane (Figure 1). This decomposition inherits the spatial equivariance of the encoder: translating the object translates the heatmap, and changing the viewpoint changes the pixel but not the height.

On a real SO-100 robot arm, PARA achieves 95–97% across three tasks with only 20 demonstrations, while coordinate regression (ACT) scores 0–11% and motion-track regression scores 5–61%. PARA transfers zero-shot to new viewpoints (52% vs. 0%) and new environments (94% vs. 0%). In controlled simulation, PARA outperforms ACT by 20–53 percentage points across six OOD axes. Pixel alignment also makes video diffusion models effective action backbones (92% vs. 0% for global regression on the same features) and enables cross-embodiment transfer via point-track pretraining (66% vs. 10% from scratch).

Contributions.

- A pixel-aligned action formulation that predicts end-effector targets as dense image-space classifications, inheriting encoder equivariance for spatial and viewpoint robustness.
- Real-robot experiments on three tasks showing 86–95pp improvements over coordinate regression, with zero-shot viewpoint and environment transfer.
- Controlled simulation experiments isolating the action head across six OOD generalization axes.
- A video-to-action recipe where PARA heads on video diffusion features achieve 92% vs. 0% for global regression.
- Cross-embodiment transfer via point-track pretraining, achieving 66% with 10 demos vs. 10% from scratch.

2 Related Work

Visuomotor policy learning. Behavioral cloning from RGB typically predicts actions in robot coordinates. ACT predicts action chunks from a CLS-token representation; Diffusion Policy models the action distribution with denoising. These approaches operate in coordinate space and are sensitive to viewpoint and position shift.

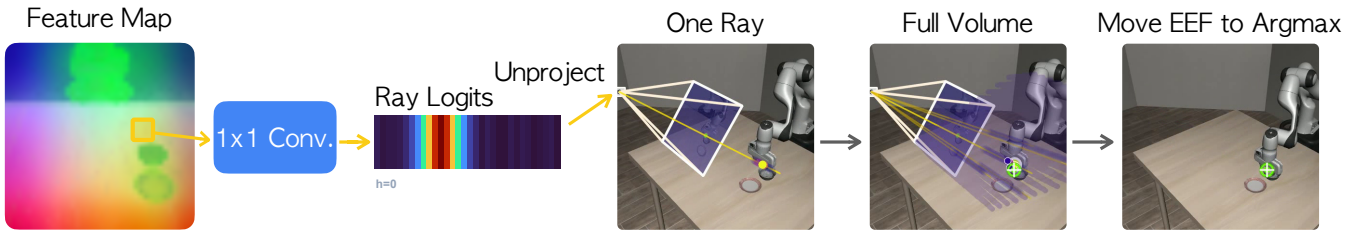

Figure 3: PARA method. A DINOv2 backbone produces spatial features; a 1×1 conv head predicts a 2D heatmap stacked with per-pixel height-bin logits to form a full 3D volume. The 3D argmax yields the end-effector target in camera coordinates, transformed to the robot frame via known camera extrinsics.

Pixel-aligned prediction. Dense, spatially-aligned outputs are standard in 3D vision (depth, flow, correspondence). Transporter Networks and CLIPort use dense pick-and-place predictions but are limited to top-down planar tasks. PARA extends pixel-aligned prediction to 6-DoF manipulation via height-bin classification.

Point tracking for robot control. Recent approaches use predicted 2D point trajectories as intermediate representations but still regress global coordinates from tracked points. We compare against a motion-track baseline and show dense pixel classification provides stronger robustness.

Video models for robot control. UniPi and SuSIE generate goal-conditioned video and extract actions via inverse dynamics. PARA heads read actions directly from video diffusion features without a separate inverse model.

3 Method

3.1 Problem Setup

We consider behavioral cloning from demonstrations $\mathcal{D} = \{(I_t, a_t, K, T_{\text{cam}})\}$, where I_t is an RGB image, a_t is the end-effector action (3D position + gripper state), K is the camera intrinsic matrix, and T_{cam} is the extrinsic transform. We assume a known support surface defining a world-frame height axis. The policy predicts the next N_W actions from a single image.

3.2 Pixel-Aligned Heatmap Volume

PARA decomposes action prediction into three pixel-space classification problems (Figure 3).

2D localization. A vision encoder f_θ produces spatial features $F \in \mathbb{R}^{H' \times W' \times C}$. A 1×1 conv head, bilinearly upsampled to (H, W), produces heatmap logits $Z \in \mathbb{R}^{N_W \times H \times W}$ per timestep. Supervision is cross-entropy over the flattened $H \times W$ grid:

$$\mathcal{L}_{\text{spatial}} = -\frac{1}{N_W} \sum_{k=1}^{N_W} \log \frac{\exp(Z_k[u_k^*, v_k^*])}{\sum_{u,v} \exp(Z_k[u, v])}, \tag{1}$$

where $p_k^* = (u_k^*, v_k^*)$ is the ground-truth pixel obtained by projecting the demonstrated end-effector position. At inference, $\hat{p}_k = \arg\max_{u,v} Z_k[u, v]$.

Height prediction. A second head produces per-pixel logits over N_H height bins: $H_{\text{vol}} \in \mathbb{R}^{N_W \times N_H \times H \times W}$. During training, height loss is evaluated at the ground-truth pixel (teacher forcing):

$$\mathcal{L}_{\text{height}} = -\frac{1}{N_W} \sum_{k=1}^{N_W} \log \frac{\exp(H_{\text{vol},k}[h_k^*, u_k^*, v_k^*])}{\sum_{j=1}^{N_H} \exp(H_{\text{vol},k}[j, u_k^*, v_k^*])}. \tag{2}$$

At inference, height logits are read at the predicted pixel $\hat{p}_k$.

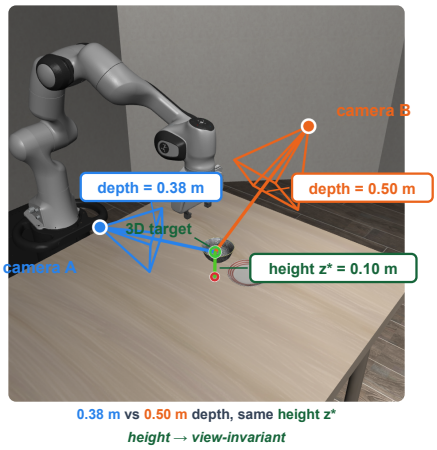

Figure 4: Height is view-invariant. For the same physical target, depth changes with camera position (0.38 m vs. 0.50 m), but height above the table is constant. PARA predicts height, making the lifting step camera-invariant.

Gripper prediction. A third head predicts per-pixel gripper-state logits over N_G bins: $G \in \mathbb{R}^{N_W \times N_G \times H \times W}$, trained identically.

3.3 3D Recovery via Height-Plane Intersection

Given predicted pixel $\hat{p}_k = (u_k, v_k)$ and height $\hat{h}_k$, we recover the 3D target by intersecting the camera ray through (u_k, v_k) with the plane $z = \hat{h}_k$ in world coordinates. Predicting height rather than depth is key: height is defined in the world frame (distance above the table) and is invariant to camera position, while depth changes with viewpoint for the same physical point (Figure 4).

3.4 Start-Keypoint Conditioning

The current end-effector position is projected into the image and a learnable embedding is added to the corresponding patch token, providing spatial grounding without explicit robot state input.

3.5 Training Details

For backbone experiments, we use DINOv2 ViT-S/16, producing 28×28 patch features for 448×448 inputs. Heads are 1×1 convolutions upsampled bilinearly. Total loss: $\mathcal{L} = \mathcal{L}_{\text{spatial}} + \mathcal{L}_{\text{height}} + \mathcal{L}_{\text{gripper}}$. Hyperparameters: $N_W = 12$ timesteps, $N_H = 32$ height bins, $N_G = 32$ gripper bins.

4 Experiments

We evaluate PARA on four fronts: real-robot manipulation (Section 4.1), controlled simulation OOD analysis (Section 4.2), video diffusion as policy backbone (Section 4.3), and cross-embodiment transfer via point-track pretraining (Section 4.4). In all experiments, PARA and baselines share the same vision backbone and training data, isolating action parameterization.

Baselines. ACT (Action Chunking with Transformers): CLS token from the shared DINOv2 backbone fed to an MLP regressing (x, y, z) + gripper directly—standard coordinate regression. Motion Tracks: predicts 2D point tracks across frames and regresses end-effector coordinates from tracked positions.

4.1 Real Robot Experiments

Setup. We evaluate on an SO-100 robot arm with a single wrist-mounted RGB camera. All methods use DINOv2 ViT-S/16, trained on 20 kinesthetic demonstrations per task from a single viewpoint, with no data augmentation. Three tasks: pick and place (teacup on saucer), wipe table, and fold towel.

In-distribution performance. Figure 2a reports task completion rates. PARA achieves 95–97% across all three tasks. ACT achieves at most 11%, failing to reach correct locations despite identical visual features. Motion Tracks achieves 61% on wipe table (a coarse sweeping task) but only 5% on precise pick-and-place, indicating that sparse point tracking helps with gross motion but lacks precision for fine manipulation.

Out-of-distribution transfer. Figure 2b tests generalization on pick-and-place. For zero-shot viewpoint transfer (camera repositioned, no additional data): PARA 52%, both baselines 0%. With 5 fine-tuning demonstrations at the new viewpoint: PARA 87%, ACT 4%, Motion Tracks 0%. For new environment (different table, background, lighting): PARA 94%, ACT 0%, Motion Tracks 6%. PARA transfers because its predictions depend on local object appearance, not global scene features.

4.2 OOD Analysis in Simulation

The real-robot results show practical impact but confounds make it hard to isolate why PARA helps. We use LIBERO to run controlled experiments where PARA and ACT share identical backbones, data, and evaluation, differing only in the action head.

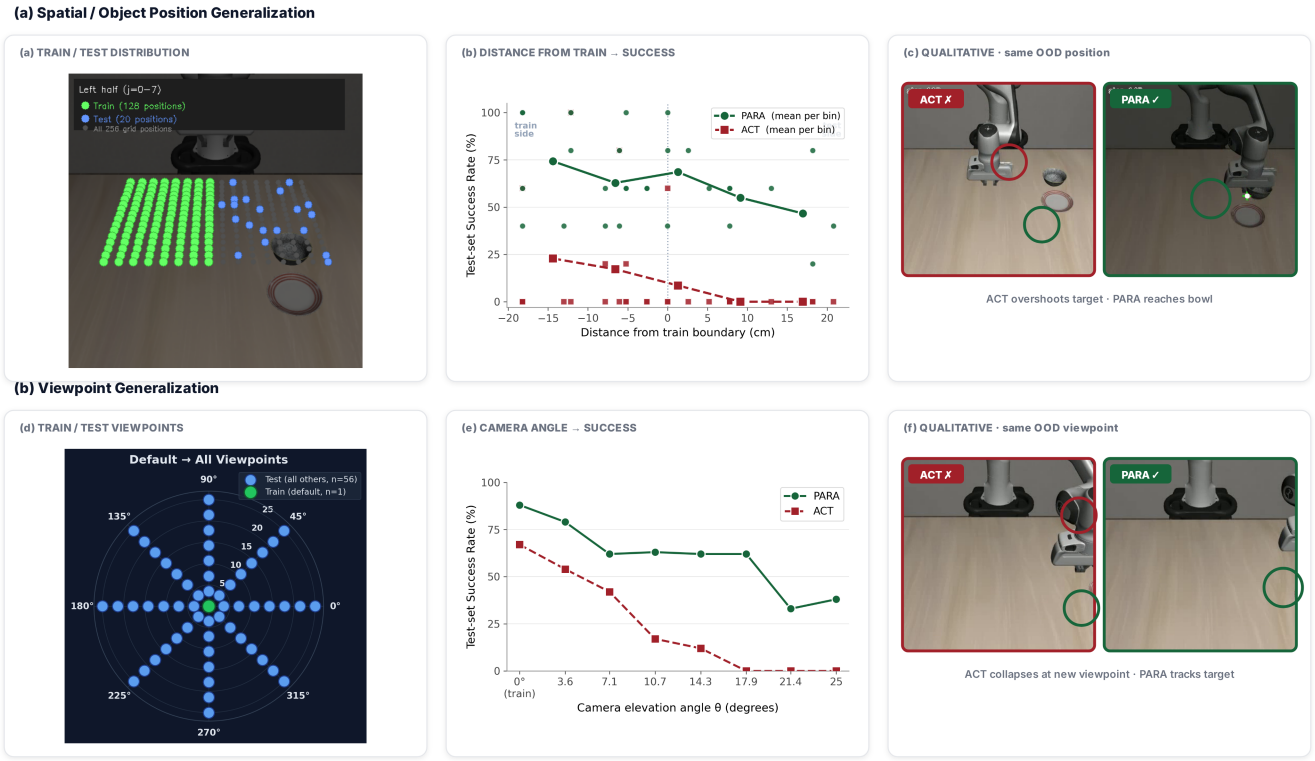

Figure 5: Controlled OOD analysis (LIBERO simulation). (a) Spatial generalization: (i) train/test position distribution, (ii) success vs. distance from training boundary—PARA degrades gracefully while ACT collapses, (iii) qualitative comparison at the same OOD position. (b) Viewpoint generalization: (i) polar plot of train (green) vs. test (blue) viewpoints, (ii) per-θ success—PARA holds $\sim$62% through 17.9° while ACT drops to 0% beyond 14.3°, (iii) qualitative comparison at an OOD viewpoint.

Setup. Task: pick-and-place (bowl on plate), LIBERO spatial task 0. Teleport servo execution isolates action prediction from controller dynamics. Object-position dataset: 16×16 grid, 39×60 cm, 256 demos. Viewpoint dataset: 8×8 grid ($\theta \in [0°, 25°]$, $\phi \in [0°, 315°]$), 10 demos per viewpoint, 640 total.

OOD object position. Train on one half of the position grid, test on the other (Figure 5a). Left-to-right extrapolation: PARA 54%, ACT 1%. Near-to-far: PARA 46%, ACT 7%. ACT reaches toward memorized training positions; PARA's heatmap tracks the object. Figure 5a(ii) shows success as a function of distance from the training boundary: PARA degrades gradually while ACT collapses immediately.

OOD camera viewpoint. Both models trained at default viewpoint ($\theta = 0°$), tested across the full grid (Figure 5b). PARA 61% across all viewpoints; ACT 24%. Figure 5b(ii) shows the per-θ breakdown: PARA maintains $\sim$62% through $\theta = 17.9°$; ACT degrades monotonically and collapses to 0% beyond 14.3°. Hemisphere transfer (train left, test right): PARA 40%, ACT 10%.

Data efficiency and distractors. With N=32 corner demonstrations: PARA 54%, ACT 33%. With dense coverage (N=64), ACT catches up (71% vs. 68%), confirming PARA's advantage is specifically OOD. Distractor robustness (clean train, cluttered test): PARA 60%, ACT 40%.

Failure modes. ACT fails by reaching to memorized locations (wrong position). PARA fails on gripper timing (correct reach, drops during transport).

4.3 Video Diffusion as Policy Backbone

Video diffusion models produce spatially-aligned features that PARA can exploit directly. We attach PARA heads to the UNet of Stable Video Diffusion (SVD, 7 frames at 576×320), using concatenated features from decoder up-blocks

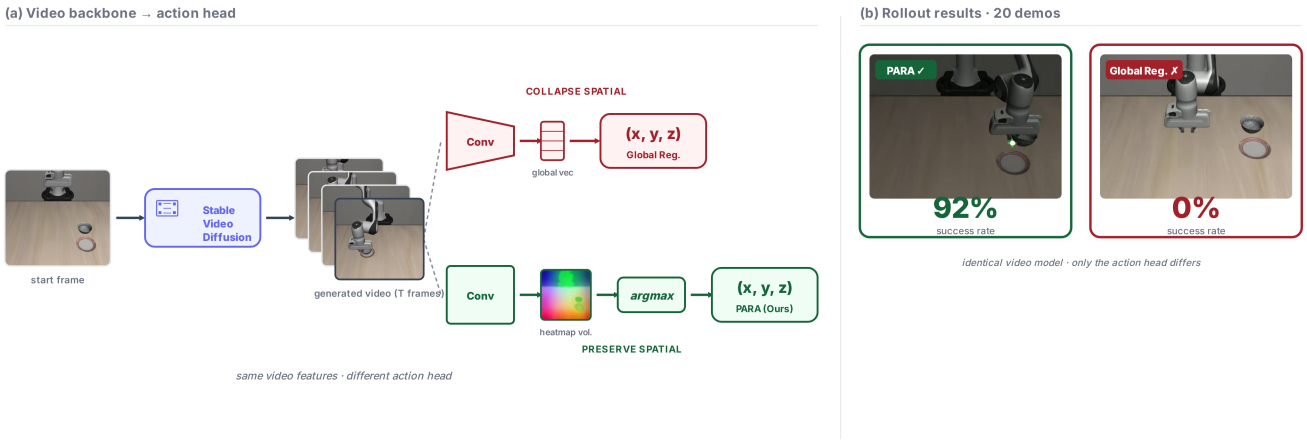

Figure 6: Video diffusion as policy backbone. (a) The same SVD UNet features are fed to two action heads: global regression (average pool → MLP, collapses spatial structure) vs. PARA (conv → argmax, preserves spatial structure). (b) Rollout results with 20 demos: PARA achieves 92% task success; global regression scores 0%. The only difference is the action head.

at 64×64 resolution.

Two-stage training. We pretrain the SVD model for 4K steps (diffusion loss only), then jointly fine-tune with PARA heads for 3K steps using separate learning rates (UNet: 10^{-6}, PARA: 10^{-4}). This achieves 92% task success, outperforming joint-from-scratch (55% at 10K steps) with less total compute.

Co-adaptation is essential. Frozen video backbone + PARA heads: 0%. Video features are spatially informative but not action-relevant without fine-tuning.

PARA vs. global regression. Replacing PARA heads with global average pooling + MLP on the same UNet features with the same two-stage training: 0%. This is the clearest evidence that pixel alignment, not just strong features, enables video-to-action transfer (Figure 6).

4.4 Cross-Embodiment Transfer via Point-Track Pretraining

PARA's pixel-aligned prediction has a natural connection to point tracking: both reason about where things are in the image. We exploit this by pretraining the PARA backbone on videos with circle overlays—the robot is masked out and replaced with an orange circle at the end-effector position (Figure 7a). This teaches the model to track points across frames without requiring robot-specific action labels, enabling pretraining on diverse embodiments.

Few-shot fine-tuning. We pretrain on circle-overlay videos from one embodiment, then fine-tune with PARA heads on a target robot with limited demonstrations. Figure 7b shows results on LIBERO with varying numbers of fine-tuning demos. At 10 demonstrations, PARA pretrained achieves 66% vs. 10% from scratch—a 6.6× improvement. ACT benefits less from the same pretraining (23% pretrained vs. 10% scratch), confirming that pixel-aligned prediction is better positioned to exploit correspondence-based pretraining.

5 Discussion

Why does pixel-aligned prediction help? Coordinate regression maps a global image representation to a global 3D target—a mapping that changes with every shift in camera, position, or layout. PARA decomposes this into where in the image (inherits encoder equivariance) and at what height (invariant by construction). With only 20 demonstrations, this decomposition makes action prediction tractable where global regression cannot memorize the mapping.

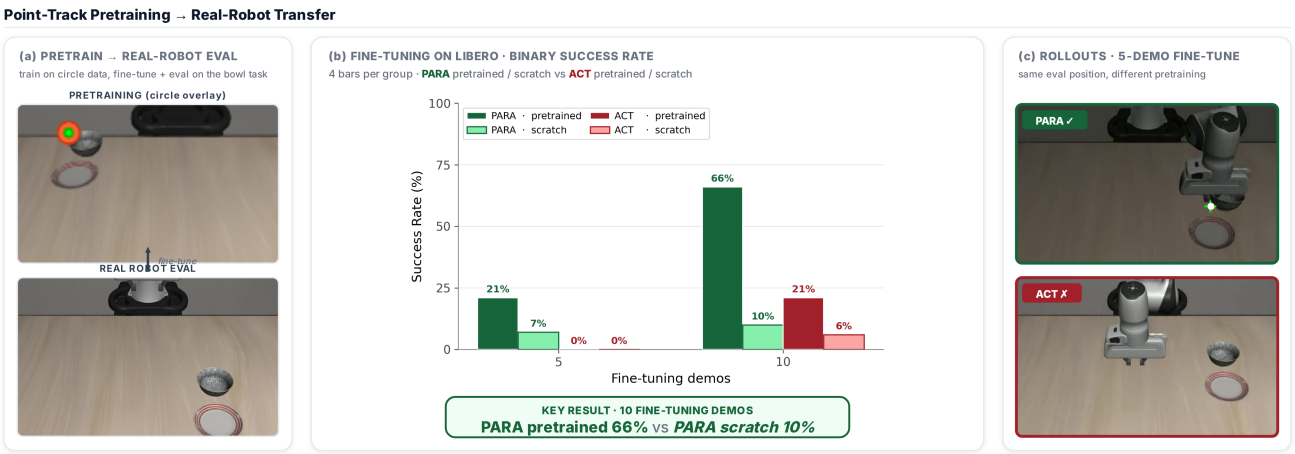

Figure 7: Cross-embodiment transfer via point-track pretraining. (a) Pretraining data: circle overlays mark the end-effector on training videos with the robot masked out, enabling embodiment-agnostic point-tracking supervision. (b) Fine-tuning on LIBERO: PARA pretrained achieves 66% with 10 demos vs. 10% from scratch; ACT benefits less from the same pretraining. (c) Qualitative rollouts after 5-demo fine-tuning on a different embodiment.

When does coordinate regression suffice? With dense coverage (N=64 uniform positions), ACT achieves 71% vs. PARA's 68%. PARA's advantage is specifically in the OOD and low-data regimes—precisely where real-robot learning operates.

Limitations. Simulation experiments use a single LIBERO task; real-robot experiments cover three tasks on one embodiment (SO-100, a low-cost arm). PARA assumes a known support surface for height-based lifting. Teleport servo in simulation bypasses controller dynamics. Per-position evaluation uses 5 episodes (high variance per position; aggregates are reliable).

6 Conclusion

We presented PARA, a pixel-aligned action formulation that predicts robot actions as dense image-space classifications. On a real robot, PARA achieves 95–97% across three tasks with 20 demonstrations and transfers to new viewpoints and environments where baselines collapse. Controlled simulation experiments confirm the advantage stems from action parameterization. Pixel alignment makes video diffusion models effective action backbones (92% vs. 0% for global regression) and enables cross-embodiment transfer via point-track pretraining (66% vs. 10% from scratch). These results suggest that how actions are parameterized matters as much as how images are encoded.

Acknowledgements. Placeholder.