

Exo2EgoPolicy: Geometry-Aware Policy Transfer from Exocentric Human Demonstrations

Hritam Basak^{1,2}, Hadi Tabatabaei¹, Xin Yang¹, Shreekanth Gayaka¹, Nan Qiao¹, Yuyin Sun¹, Cheng-Hao Kuo¹, Zhaozheng Yin², Min Sun¹

¹Amazon.com; ²Stony Brook University, USA

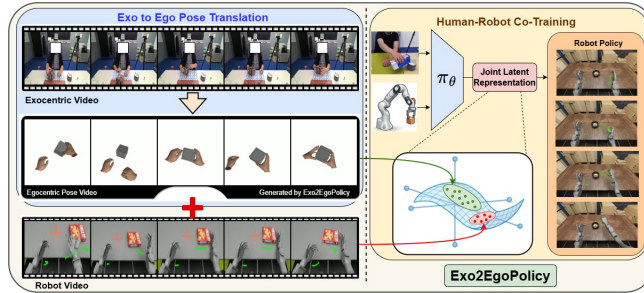


Fig. 1: We propose Exo2EgoPolicy that learns robot policies by translating exocentric (third-person) human video into egocentric (first-person) hand-object pose, enabling egocentric human-robot co-training for diverse manipulation tasks.

Abstract. Learning robot manipulation policies from large-scale human videos is challenging due to the viewpoint mismatch between third-person observations and egocentric robot control. To this end, we introduce Exo2EgoPolicy, a geometry-aware framework that learns egocentric manipulation policies from monocular exocentric video via viewpoint-aligned pose translation. Unlike prior Exo→Ego approaches that rely on diffusion-based pixel synthesis and often produce temporally inconsistent predictions, our method operates directly on a structured pose manifold and explicitly models the geometric ambiguity inherent in monocular observation. We formulate exocentric-to-egocentric translation as a viewpoint-conditioned latent variable model on the $SE(3)$ manifold that disentangles pose from camera transformation and captures the ambiguity induced by unknown camera extrinsics. Under bounded viewpoint motion, the formulation yields pose trajectories that are identifiable up to a global rigid transformation and supports stable sequential inference through temporal regularization. Beyond representation alignment, we show that for quasi-static manipulation tasks whose rewards depend primarily on relative hand-object geometry, pose captures the key task-relevant information required for policy learning. This enables policy co-training on pose-translated human demonstrations alongside limited robot teleoperation data within a unified human-robot state space. Empirically, Exo2EgoPolicy improves temporal consistency, cross-view alignment, and out-of-distribution policy transfer compared to pixel-based translation and other cross-view co-training baselines, achieving

20–30% higher task success on manipulation benchmarks. These results suggest that geometry-aligned pose representations provide a scalable foundation for cross-view policy learning from human video. [Website](#)

Keywords: Exo2Ego · Diffusion Model · Co-training

1 Introduction

From early developmental stages, humans observe external actions and seamlessly internalize them into their own motor repertoire. Driven by mirror-neuron mechanisms, we can effortlessly infer the spatial relationships and hand–object interactions that would be perceived from a first-person perspective simply by watching someone else act. This actor–observer translation ability underpins human imitation learning [22] and has profound implications for robotics [15], augmented/virtual reality [23], and embodied intelligence [55]. In robot learning, projecting a human instructor’s hand–object interactions into the robot’s own viewpoint enables direct policy alignment with demonstrations [19, 80].

Despite the well-established utility of egocentric (first person) data for manipulation learning [20, 79], the overwhelming majority of human demonstrations are recorded in exocentric (third-person) formats. Bridging this viewpoint gap is fundamentally challenging. Unlike standard novel view synthesis, exocentric-to-egocentric (Exo-to-Ego) translation involves extreme viewpoint shifts where critical regions are fully occluded [28]. For example, the inner pages of a book or the fine-grained contact geometry between fingers and objects may be invisible in the exocentric recording but must be inferred in the egocentric perspective. The problem is therefore inherently underdetermined. Existing approaches attempt to resolve this challenge through either geometry-aware regression or generative modeling. Geometry-based methods such as NeRF-style approaches [65, 72] are fundamentally regressive: they require dense multi-view supervision and struggle under single-view input with heavy occlusion. Conversely, generative diffusion models [11, 34, 44] can hallucinate unseen regions, but they operate in pixel space and often exhibit geometric drift and temporal instability. While such generative models may achieve strong perceptual metrics (e.g., FID or LPIPS), they lack the structural guarantees required for downstream robotic control. Frame-wise pixel hallucination further introduces high-frequency jitter and spatial inconsistencies that corrupt continuous kinematic trajectories, rendering them unsuitable for derivative-based manipulation policies.

We argue that this reflects a deeper misalignment in problem formulation. Prior works treat Exo-to-Ego translation purely as a visual synthesis task. However, for robotic manipulation, the objective is not photorealistic rendering, but recovery of action-relevant geometric structure. Manipulation rewards depend primarily on relative hand–object pose, rather than on background texture fidelity. Pixel-level hallucination, therefore, introduces unnecessary variance while failing to enforce temporal and geometric consistency.

To address this, we propose a paradigm shift from pixel synthesis to structured geometric alignment. We introduce **Exo2EgoPolicy** (shown in [Figure 1](#)),

a framework that operates entirely in a structured pose manifold and bypasses raw pixel ambiguities. Instead of generating egocentric images, we obtain explicit 3D hand-object pose from monocular exocentric input and formulate Exo-to-Ego translation as a viewpoint-conditioned latent variable model. This model disentangles intrinsic pose from camera transformation and explicitly characterizes the ambiguity induced by unknown extrinsics. By constraining view translation to the $SE(3)$ manifold, we reduce the underdetermined appearance problem to a geometry-consistent inference problem. We further introduce a temporal regularization that enforces cross-frame smoothness directly in pose space. Under bounded viewpoint motion, the formulation yields trajectories that are identifiable up to a global rigid transformation. We further derive a stability bound indicating that pose prediction errors remain bounded over time, promoting temporally consistent pose trajectories. By operating directly in pose space rather than pixel space, the approach also helps mitigate the trajectory jitter that can arise when diffusion-based view synthesis generates frames independently, while preserving task-relevant geometric structure.

Crucially, this perspective suggests that pose-level abstraction is not merely a practical simplification but a representation aligned with the structure of many manipulation tasks. Modeling manipulation as a Markov Decision Process (MDP), we show that for quasi-static tasks whose reward depends primarily on relative hand-object geometry, pose captures the key task-relevant information required for policy learning. This provides a principled motivation for avoiding pixel reconstruction in favor of geometry-aligned representations. By co-training policies on pose-translated human demonstrations alongside limited real-robot teleoperation data, we help bridge the embodiment gap and enable leveraging third-person human videos as a scalable supervision source for robot learning. By reframing Exo-to-Ego translation as structured geometric inference rather than pixel synthesis, Exo2EgoPolicy shifts cross-view generation from visual appearance modeling toward recovering task-relevant geometric structure for embodied policy learning. Our contributions are as follows:

- **Exo-to-Ego Policy Co-Training Framework.** We introduce a framework that leverages viewpoint translation for embodied control, converting third-person human videos into egocentric pose demonstrations that can be co-trained with robot teleoperation data for manipulation policy learning.
- **Geometry-Aware Pose Translation.** We formalize Exo-to-Ego translation as a viewpoint-conditioned latent variable model on the $SE(3)$ manifold, resulting in trajectories that are identifiable up to a global rigid transformation under bounded viewpoint motion.
- **Temporal Stability.** We introduce temporal regularization in pose space and derive a stability bound indicating that pose prediction errors remain bounded over time, promoting temporally consistent pose trajectories.
- **Pose Sufficiency for Policy Learning.** We provide an MDP-based analysis showing that for quasi-static manipulation tasks whose rewards depend primarily on relative hand-object geometry, pose captures the task-relevant information required for policy learning. Empirically, this pose-centric for-

mulation improves out-of-distribution task success by 20–30% compared to pixel-based translation baselines.

2 Related Works

2.1 Exo-to-Ego View Generation

Exo-to-ego view generation addresses the substantial geometric and appearance gap between third- and first-person perspectives. Early approaches [34, 44] condition egocentric synthesis via channel-wise concatenation of exocentric features. However, the lack of pixel-level correspondence across viewpoints introduces significant spatial misalignment, limiting geometric reasoning and often degrading synthesis fidelity. 4Diff [11] improves cross-view fusion through cross-attention, yet departs from fully pretrained diffusion backbones, which may limit generalization and scalability. Subsequent methods reduce ambiguity through additional supervision or multi-view constraints. EgoExo-Gen [64] conditions generation on the first egocentric frame, effectively narrowing the viewpoint gap but limiting applicability when egocentric initialization is unavailable. Exo2Ego-V [29] performs full video translation using multiple exocentric inputs and decoupled spatial-temporal modules, which constrains scalability and weakens unified spatio-temporal modeling. More broadly, recent video diffusion models [9, 18, 57, 71] have enabled high-quality conditional synthesis, including depth-, image-, and camera-controlled generation [35, 66, 73]. Camera-control approaches typically condition directly on camera extrinsics or leverage intermediate 3D representations [3, 4, 37, 49]. However, these techniques are primarily designed for moderate viewpoint perturbations and struggle under the extreme pose shifts inherent in exocentric-to-egocentric translation.

In contrast, we move beyond pixel-level synthesis and formulate Exo-to-Ego generation as geometry-aware pose inference on the $SE(3)$ manifold. By explicitly disentangling viewpoint from intrinsic hand-object geometry and enforcing temporal stability, our framework directly produces control-ready pose trajectories rather than appearance-only video synthesis.

2.2 Imitation Learning and Learning from Human Videos

Robot imitation learning has demonstrated substantial improvements in dexterity and robustness using large-scale robot-collected datasets [12, 33, 60, 74, 78], facilitated by advances in scalable data collection [1, 13, 69]. Despite these advances, robotics still lacks internet-scale embodied datasets comparable to vision [47] and language [41], limiting open-world generalization across tasks and environments. To leverage abundant human videos [8, 39, 51, 70], prior work typically aligns either observations or actions. Observation-alignment methods rely on intermediate representations such as affordances [2], object keypoints [6, 14, 63], or structured object embeddings [36, 38]. While effective, these modular pipelines can introduce compounding perceptual errors and complicate end-to-end credit assignment.

Action-alignment approaches address embodiment mismatch via trajectory or flow matching [48], object-centric abstractions [21], masking strategies [25], hardware-based hand tracking [59], or teleoperation-based pose remapping such as HumanPlus [16]. In contrast, our method does not require hardware instrumentation and explicit retargeting, directly co-training robot policies on geometry-consistent poses obtained through Exo-to-Ego translation.

2.3 Cross-Embodiment

Cross-embodiment pre-training improves adaptability and generalization across heterogeneous robotic platforms [10, 17, 53, 67, 76]. Recent end-to-end approaches using cross-robot data [31, 56] reduce compounding perceptual errors and exhibit emergent corrective behaviors such as automatic retrying [7]. Other methods enforce viewpoint constraints between human demonstrations and deployment environments [27], trading scalability for a reduced domain gap.

Concurrent work explores learning directly from egocentric human demonstrations [58]. DexCap [59] employs glove-based 3D tracking with RGBD capture, while EgoMimic [25] relies on closely matched egocentric camera viewpoints. In contrast, our approach scales across diverse camera configurations without alignment constraints or specialized hardware. By coupling geometry-aware Exo-to-Ego pose translation with policy co-training, our approach enables scalable cross-embodiment learning from heterogeneous human data.

3 Proposed Method

We present *Exo2EgoPolicy*, a geometry-aware framework for transferring manipulation skills from third-person (exocentric) human demonstrations to egocentric robot control. Rather than synthesizing novel-view pixels, which are computationally expensive and may introduce task-irrelevant visual artifacts, our key insight is to translate only the hand-object kinematic state. The framework consists of four tightly coupled components. First, we formulate Exo-to-Ego translation as a viewpoint-conditioned latent diffusion model over the $SE(3)$ manifold, resolving the geometric ambiguity inherent to single-view projection (subsection 3.1). Second, to prevent the frame-to-frame jitter that arises from independent per-frame generation, we augment the model with a temporal regularization using a motion prior which ensures bounded cumulative trajectory error (subsection 3.2). Third, we provide a formal Markov Decision Process (MDP) argument showing that this viewpoint-aligned pose is a *sufficient statistic* for optimal policy learning in manipulation tasks (subsection 3.3). Finally, we describe a co-training scheme that combines large-scale translated human demonstrations with limited robot teleoperation data in a unified state-action space (subsection 3.4). The overall workflow is shown in Figure 2.

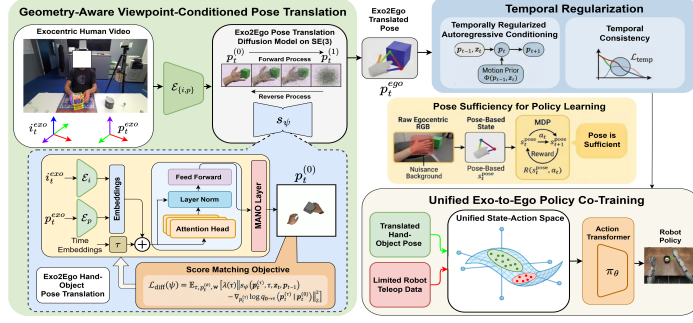


Fig. 2: Overview of the Exo2EgoPolicy framework. (*Top left*) Given an exocentric RGB observation i_t^{exo} and the corresponding exocentric pose p_t^{exo} , the geometric encoders $\mathcal{E}_{\{i,p\}}$ produces a dense latent representation that conditions the Exo2Ego Pose Translation Diffusion Model (subsection 3.1). The score network s_ψ is trained via a continuous-time score-matching objective $\mathcal{L}_{\text{diff}}$ on the $SE(3)$ manifold, running a forward noising process $p_t^{(0)} \rightarrow p_t^{(1)}$ and learning to reverse it to synthesize the egocentric pose p_t^{ego} , as shown in (*bottom left*). (*Top right*) The translated pose p_t^{ego} is passed to the Temporal Regularization module (subsection 3.2), where a motion prior $\Phi(p_{t-1}, z_t)$ enables temporally regularized autoregressive conditioning across frames, supervised by the temporal consistency loss $\mathcal{L}_{\text{temp}}$ (Eq. 4). We formally establish that the generated pose is sufficient for cross-domain policy learning (subsection 3.3). (*Bottom right*) The resulting temporally smooth egocentric pose, together with limited teleoperation data, are projected into a unified state-action space and co-train the policy network π_θ (subsection 3.4) via the joint behavioral cloning objective $\mathcal{L}_{\text{policy}}$ (Eq. 7).

3.1 Geometry-Aware Viewpoint-Conditioned Pose Translation

Problem setup. Let $\mathcal{I}_{1:T}^{\text{exo}} = \{i_1^{\text{exo}}, \dots, i_T^{\text{exo}}\}$ be a sequence of T exocentric RGB frames captured from an uncalibrated, potentially moving camera. Our goal is to recover the corresponding egocentric hand-object pose trajectory $\mathcal{P}_{1:T}^{\text{ego}} = \{p_1, \dots, p_T\}$, where each state $p_t = (p_t^{\text{hand}}, p_t^{\text{obj}})$ lives on the joint hand-object kinematic manifold $\mathcal{M}_{\text{pose}} \triangleq \mathcal{M}_{\text{hand}} \times \mathcal{M}_{\text{obj}}$. Here, $\mathcal{M}_{\text{hand}} = SE(3) \times \mathbb{R}^K$ captures the 6-DoF wrist pose and K -dimensional finger articulation, while $\mathcal{M}_{\text{obj}} = SE(3)$ captures the 6-DoF pose of the manipulated object. We drop “ego” superscript on p_t for brevity; poses are egocentric unless otherwise stated.

Why not pixel-space translation. Recovering an egocentric image from a monocular exocentric view is an ill-posed inverse problem: a single 2D projection is geometrically consistent with infinitely many 3D scenes. Prior pixel-space approaches [34, 44] resolve this ambiguity by hallucinating unobservable scene regions, inevitably producing high-frequency spatial artifacts and structural inconsistencies [11]. These background nuisance variables are irrelevant and possibly detrimental to embodied control. We therefore bypass pixel space entirely and formulate exo-to-ego view translation directly on $\mathcal{M}_{\text{pose}}$.

Geometric latent encoding. We first compress each exocentric frame and pose into a dense, geometry-focused latent: $z_t = \{\mathcal{E}_i(i_t^{\text{exo}}); \mathcal{E}_p(p_t^{\text{exo}})\} \in \mathbb{R}^d$, where

$\mathcal{E}_{\{i,p\}}$ are image and pose encoders, respectively, and p_t^{exo} is corresponding exocentric hand-object pose (extracted from the frame using [45], if not provided). Because the exocentric camera extrinsics could be uncalibrated, multiple valid mappings $z_t \mapsto p_t$ may be consistent with the same observation. We alleviate this with a diffusion model below.

Diffusion model on $\mathcal{M}_{pose}$. Let $\tau \in [0, 1]$ denote the diffusion pseudo-time. We define a forward noising process that smoothly perturbs the true egocentric pose $p_t^{(\tau=0)} \equiv p_t$ into a tractable Gaussian prior on the tangent space of $\mathcal{M}_{pose}$:

$$dp_t^{(\tau)} = f(p_t^{(\tau)}, \tau) d\tau + g(\tau) d\mathbf{w}, \quad (1)$$

where $\mathbf{w}$ is a standard Wiener process, $f(\cdot, \cdot)$ is a linear drift function, and $g(\cdot)$ is the scalar diffusion coefficient. Since the relative camera transform is unknown at both training and inference, we condition the score network only on the observable geometric latent z_t and the previously inferred pose p_{t-1} . The reverse-time stochastic differential equation for pose synthesis is:

$$dp_t^{(\tau)} = \left[f(p_t^{(\tau)}, \tau) - g(\tau)^2 \nabla_{p_t} \log q_\tau(p_t^{(\tau)} | z_t) \right] d\tau + g(\tau) d\bar{\mathbf{w}}, \quad (2)$$

where $\bar{\mathbf{w}}$ is a reverse-time Wiener process. The conditional score function $\nabla_{p_t} \log q_\tau(p_t^{(\tau)} | z_t)$ guides the reverse process toward poses that are both kinematically plausible and consistent with the observed frame. We approximate it with a neural network $s_\psi(p_t^{(\tau)}, \tau, z_t)$, trained via the continuous-time conditional score-matching:

$$\mathcal{L}_{\text{diff}}(\psi) = \mathbb{E}_{\tau, p_t^{(0)}, \mathbf{w}} \left[\lambda(\tau) \left\| s_\psi(p_t^{(\tau)}, \tau, z_t) - \nabla_{p_t^{(\tau)}} \log q_{0 \rightarrow \tau}(p_t^{(\tau)} | p_t^{(0)}) \right\|_2^2 \right], \quad (3)$$

where $\lambda(\tau) > 0$ is a positive weighting function and the expectation is over $\tau \sim \mathcal{U}(0, 1)$, the ground-truth pose $p_t^{(0)}$, and the Wiener noise.

Pose Identifiability up to Rigid Transformation Although the exocentric camera pose c_t is unknown, the generated trajectory is not arbitrary. Rather than predicting absolute coordinates, we set the generative target as the hand pose *relative* to the manipulated object, i.e., $(p_t^{\text{obj}})^{-1} \cdot p_t^{\text{hand}}$. Because the object provides a locally static reference frame, this formulation naturally marginalizes out the dynamic camera. We formalize the pose identifiability as follows.

Proposition 1 (Identifiability up to $SE(3)$). *Suppose s_ψ converges to the true conditional score. Let the relative inter-frame camera motion satisfy $\left\| \log_{SE(3)}(\Delta c_{t,t+1}) \right\| \leq \epsilon$ for sufficiently small $\epsilon > 0$, and assume the manipulated object defines a locally static reference frame. Then the support of the generated egocentric trajectory $\hat{\mathcal{P}}_{1:T}^{\text{ego}}$ collapses to a single manifold, identifiable up to a time-invariant rigid transformation $T_{\text{global}} \in SE(3)$.*

Intuition. Under slow camera motion ($\epsilon \rightarrow 0$), the temporal variation of c_t is negligible relative to intrinsic hand-object kinematics. Conditioning on the object reference frame reduces the conditional distribution to a Dirac measure at the true relative pose sequence, leaving only the arbitrary initial global alignment T_{global} as an unknown. This proposition ensures spatial validity for each frame independently, thus generating frames without temporal coupling might produce high-frequency kinematic jitter across the sequence, which is addressed below.

3.2 Temporally Regularized Pose Generation

We observe that frame-wise independent diffusion produces trajectories that are individually plausible but collectively discontinuous, also supported by findings in [11, 34, 44]. In novel-view synthesis this manifests as flickering; in manipulation, it produces kinematic jitter that corrupts velocity and acceleration signals required by robot controllers. We therefore augment the generative model with a temporally regularized dynamics prior.

Instead of sampling p_t from z_t in isolation, we condition the score network on the previously inferred state p_{t-1} : $s_\psi(p_t^{(\tau)}, \tau, z_t, p_{t-1})$. This makes the generation autoregressive, propagating temporal context across frames.

Temporal consistency loss. We introduce a forward kinematic motion prior $\Phi : \mathcal{M}_{\text{pose}} \times \mathbb{R}^d \rightarrow \mathcal{M}_{\text{pose}}$ that predicts the expected next state given the current state and visual features. Denoting $a \boxminus b \triangleq \log_{SE(3)}(a \cdot b^{-1})$ as the Lie group difference operator following [5, 54], the temporal consistency loss is:

$$\mathcal{L}_{\text{temp}} = \sum_{t=2}^T \left\| \hat{p}_t^{\text{hand}} \boxminus \Phi_{\text{hand}}(p_{t-1}, z_t) \right\|^2 + \beta \left\| \hat{p}_t^{\text{obj}} \boxminus \Phi_{\text{obj}}(p_{t-1}, z_t) \right\|^2. \quad (4)$$

where $\hat{p}_t = (\hat{p}_t^{\text{hand}}, \hat{p}_t^{\text{obj}})$ is the pose output by the diffusion model, $\log_{SE(3)}(\cdot)$ maps $SE(3)$ elements to their tangent-space Lie algebra vectors, and $\beta > 0$ balances hand versus object regularization. This loss acts as a geodesic low-pass filter, suppressing physically implausible jumps while preserving genuine interaction dynamics. The total training objective for the view translation module is $\mathcal{L}_{\text{diff}} + \alpha \mathcal{L}_{\text{temp}}$, where $\alpha > 0$ is a fixed weight.

Proposition 2 (Non-divergence bound). *Let $d_{\mathcal{M}}(\hat{p}_t, p_t^*)$ denote the geodesic distance between the generated pose $\hat{p}_t$ and the ground-truth p_t^* on $\mathcal{M}_{\text{pose}}$. Suppose the single-step prediction error satisfies $d_{\mathcal{M}}(\hat{p}_t, p_t^*) \leq \epsilon$, and the transition operator Φ is L -Lipschitz with $L < 1$. Then the cumulative trajectory error is uniformly bounded:*

$$\sup_{t \in [1, T]} \mathbb{E}[d_{\mathcal{M}}(\hat{p}_t, p_t^*)] \leq \frac{\epsilon}{1 - L}. \quad (5)$$

Intuition. The $L < 1$ (contraction) condition ensures that errors decay rather than compound over time, yielding a finite and controllable error ceiling.

3.3 Pose Sufficiency for Cross-Domain Policy Learning

The preceding modules produce a temporally smooth, viewpoint-aligned pose sequence $\hat{\mathcal{P}}_{1:T}^{\text{ego}}$. We now argue that this sequence contains *all* the information needed for optimal manipulation policy learning, without requiring any pixel-level information.

MDP formulation. We model the manipulation task as a Markov Decision Process (MDP) $\{\mathcal{ST}, \mathcal{A}, \mathcal{T}, R, \gamma\}$, where $\mathcal{ST}$ is the state space of joint hand-object poses, $\mathcal{A}$ is the continuous action space of end-effector displacements and finger articulations, $\mathcal{T} : \mathcal{ST} \times \mathcal{A} \rightarrow \Delta(\mathcal{ST})$ is the stochastic transition operator, $R : \mathcal{ST} \times \mathcal{A} \rightarrow \mathbb{R}$ is the task reward function, and $\gamma \in [0, 1)$ is the discount factor. We define the pose-based state as $st_t^{\text{pose}} = \hat{p}_t = (\hat{p}_t^{\text{hand}}, \hat{p}_t^{\text{obj}}) \in \mathcal{M}_{\text{pose}}$, and optimize the policy $\pi_{\text{pose}}(a_t | st_t^{\text{pose}})$ to maximize expected cumulative reward:

$$J(\pi_{\text{pose}}) = \mathbb{E}_{\pi_{\text{pose}}} \left[\sum_{t=0}^T \gamma^t R(st_t^{\text{pose}}, a_t) \right]. \quad (6)$$

Sufficiency argument. For quasi-static manipulation tasks, the reward R depends only on relative hand-object geometry. The relative transformation $\delta_t = (\hat{p}_t^{\text{obj}})^{-1} \cdot \hat{p}_t^{\text{hand}} \in SE(3) \times \mathbb{R}^K$ deterministically encodes this geometry, rendering background appearance entirely redundant. A policy conditioned on st_t^{pose} thus achieves the same optimal value as one conditioned on raw egocentric RGB, while being fully invariant to domain-specific visual factors (lighting, scene clutter, camera intrinsics). This invariance is key for cross-domain generalization.

3.4 Unified Exo-to-Ego Policy Co-Training

The scarcity of robot teleoperation data is a fundamental bottleneck for scaling manipulation policies. Our viewpoint-aligned pose representation unlocks a simple solution: treat the human hand as a distinct embodiment operating in the *same* kinematic space as the robot end-effector, and co-train on both.

Unified state-action space. We project two datasets into a shared representation:

- **Robot data:** $\mathcal{D}_{\text{robot}} = \{(p_{1:T}^{\text{robot}}, a_{1:T}^{\text{robot}})\}$, collected via teleoperation, where a_t^{robot} is the target 6-DoF end-effector displacement and finger articulation command.
- **Translated human data:** $\mathcal{D}_{\text{human}}^{\text{trans}} = \{(\hat{p}_{1:T}^{\text{human}}, \hat{a}_{1:T}^{\text{human}})\}$, produced by our view translation pipeline, with $\hat{a}_t^{\text{human}}$ defined analogously as the hand displacement in the egocentric frame.

Policy architecture and co-training objective. We parameterize π_θ as an Action Transformer [46] that predicts a future action trajectory given a short history of H joint poses: $\pi_\theta(a_t | p_{t-H:t})$. The policy is trained via joint behavioral cloning:

$$\mathcal{L}_{\text{policy}}(\theta) = \mathbb{E}_{(p,a) \sim \mathcal{D}_{\text{robot}}} [\ell(\pi_\theta(p), a)] + \mu \mathbb{E}_{(\hat{p}, \hat{a}) \sim \mathcal{D}_{\text{human}}^{\text{trans}}} [\ell(\pi_\theta(\hat{p}), \hat{a})], \quad (7)$$

where ℓ is the action regression loss and $\mu > 0$ is a hyperparameter that balances the relative dataset sizes.

4 Experiments and Results

4.1 Dataset and Implementation Details

The encoder $\mathcal{E}_i$ is a pretrained DINOv2-ViT-L/14 [42] projecting to $d=512$. The pose encoder $\mathcal{E}_p$ is a three-layer MLP with hidden dimension 256 and ReLU activations. Hand and object poses are extracted via HAMER [45] and FoundationPose [62], respectively, yielding $z_t \in \mathbb{R}^{563}$. The score network s_ψ is a 12-layer Transformer (8 heads, hidden dim 512) taking as input the embedded noisy pose, a sinusoidal diffusion-time encoding, the flattened context c_t , and the previous pose p_{t-1} , with a linear head. We adopt a VP-SDE with cosine noise schedule ($N=1000$ training steps, $N=50$ DDIM steps at inference) and weighting $\lambda(\tau) = 1 - \cos(\pi\tau/2)$. The motion prior Φ is a two-layer GRU (hidden size 256) pretrained for 20 epochs on ground-truth sequences, with spectral normalization enforcing $L=0.95$. The full translation module is trained on training splits of OakInk V2 [75], H2O [26], AssemblyHands [40], and Assembly101 [52] for 200 epochs, batch size 128, AdamW (lr= 3×10^{-4} , cosine annealing, 5K warmup, weight decay 10^{-4}) on 8 A100 GPUs, with temporal consistency weight $\beta=0.5$. For policy co-training, we follow the dataset composition and training protocol of [68], including the robot teleoperation data, task suite, action representation, and all associated hyperparameters. The only modification is the addition of $\mathcal{D}_{\text{human}}^{\text{trans}}$, consisting of exo-to-ego translated clips from TACO [32] and Mimicplay [58] spanning the same task categories. These samples are incorporated through the co-training objective with balance weight $\mu=0.3$, selected by matching the gradient norms from the two data sources at the start of training.

To provide dense geometric grounding, we augment our pose representation by attaching a lightweight mesh reconstruction head that recovers 3D hand and object meshes from the predicted hand joint parameters via the differentiable MANO layer [50] and object-specific shape priors. For comparability with prior work, we follow standard evaluation protocols [43, 45] and report both 3D joint and mesh accuracy: PA-MPJPE and AUC_J for joint accuracy, and PA-MPVPE, AUC_V , F@5mm, and F@15mm for mesh accuracy. Policy evaluation is conducted on the *Ego Humanoid Manipulation Benchmark* [68], built in NVIDIA Isaac Lab with a Unitree H1 humanoid robot, following the exact evaluation protocol, task definitions, and success criteria defined therein. Additional details are in the supplementary file.

4.2 Evaluation of Pose Translation

Since no prior work directly addresses exocentric-to-egocentric pose translation, we compare against the closest available baselines: pixel-space exo-to-ego view synthesis methods (Pix2PixHD [61], Exo2Ego-V [30], EgoWorld [44], 4Diff [11], EgoX [24]), from whose generated egocentric videos we extract poses using HAMER [45] and FoundationPose [62] to obtain pose estimates for evaluation in a unified pose space. Following the four generalization splits defined in EgoWorld [44], Table 1 reports pose and mesh accuracy across unseen objects,

Table 1: Exo-to-Ego hand pose translation across four generalization splits on H2O [26]. Best in **bold**, second best underlined.

Unseen Objects							Unseen Actions						
Method	PA-MPJPE↓	AUC _J ↑	PA-MPVPE↓	AUC _V ↑	F@5mm↑	F@15mm↑	Method	PA-MPJPE↓	AUC _J ↑	PA-MPVPE↓	AUC _V ↑	F@5mm↑	F@15mm↑
Pix2PixHD [61]	42.6	0.312	48.4	0.284	0.142	0.412	Pix2PixHD [61]	46.3	0.274	52.1	0.251	0.118	0.374
Exo2Ego-V [30]	24.1	0.518	26.3	0.474	0.312	0.645	Exo2Ego-V [30]	28.4	0.471	31.2	0.432	0.274	0.601
EgoWorld [44]	19.4	0.612	22.5	0.581	0.421	0.762	EgoWorld [44]	23.1	0.564	26.8	0.531	0.378	0.714
4Diff [11]	18.2	0.635	20.1	0.592	0.448	0.782	4Diff [11]	21.7	0.588	24.3	0.551	0.401	0.738
EgoX [24]	15.7	0.684	17.5	0.645	0.521	0.844	EgoX [24]	18.9	0.631	21.2	0.597	0.468	0.798
Ours	11.4	0.772	12.8	0.741	0.685	0.912	Ours	14.2	0.718	16.1	0.683	0.624	0.874

Unseen Scenes							Unseen Subjects						
Method	PA-MPJPE↓	AUC _J ↑	PA-MPVPE↓	AUC _V ↑	F@5mm↑	F@15mm↑	Method	PA-MPJPE↓	AUC _J ↑	PA-MPVPE↓	AUC _V ↑	F@5mm↑	F@15mm↑
Pix2PixHD [61]	44.1	0.298	50.2	0.271	0.131	0.394	Pix2PixHD [61]	43.8	0.304	49.6	0.278	0.137	0.403
Exo2Ego-V [30]	25.8	0.497	28.4	0.458	0.291	0.624	Exo2Ego-V [30]	25.2	0.508	27.4	0.466	0.301	0.634
EgoWorld [44]	20.6	0.594	23.9	0.562	0.398	0.741	EgoWorld [44]	20.1	0.603	23.2	0.571	0.410	0.752
4Diff [11]	19.3	0.618	21.8	0.574	0.424	0.762	4Diff [11]	18.8	0.627	21.1	0.583	0.436	0.772
EgoX [24]	16.8	0.661	18.9	0.624	0.498	0.824	EgoX [24]	16.2	0.673	18.2	0.635	0.511	0.834
Ours	12.3	0.751	13.9	0.718	0.658	0.891	Ours	11.8	0.762	13.3	0.729	0.671	0.902

Table 2: Exo-to-Ego hand pose translation on Assembly101 [52].

Method	PA-MPJPE↓	AUC _J ↑	PA-MPVPE↓	AUC _V ↑	F@5mm↑	F@15mm↑
Pix2PixHD [61]	44.8	0.298	51.3	0.271	0.128	0.398
Exo2Ego-V [30]	26.4	0.504	29.1	0.461	0.284	0.618
EgoWorld [44]	21.2	0.598	24.4	0.554	0.392	0.731
4Diff [11]	19.8	0.621	22.6	0.578	0.418	0.754
EgoX [24]	<u>17.1</u>	<u>0.668</u>	<u>19.3</u>	<u>0.629</u>	<u>0.504</u>	<u>0.831</u>
Ours	12.6	0.758	14.2	0.724	0.661	0.897

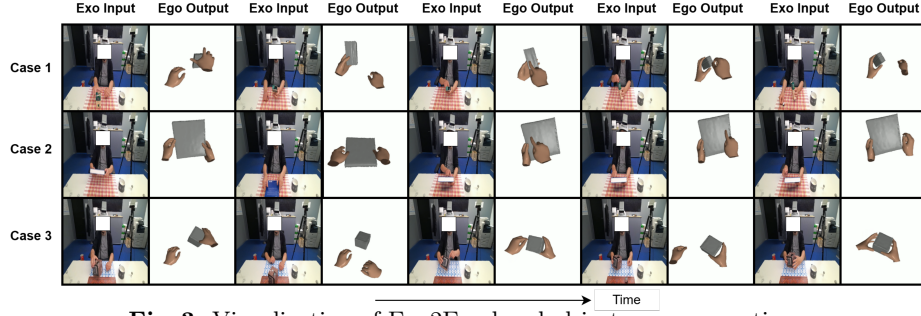


Fig. 3: Visualization of Exo2Ego hand-object pose generation.

actions, scenes, and subjects on H2O [26] benchmark. Our method consistently outperforms all baselines across every split and metric, reducing PA-MPJPE by 27% over the strongest baseline EgoX and improving F@5mm by over 30% in the most challenging Unseen Actions split. The comparatively weaker performance of pixel-space methods like [30, 44, 61] supports our central argument: synthesizing unobserved background regions introduces structural artifacts that propagate into extracted poses, whereas our pose formulation avoids this. We further validate generalization on Assembly101 [52]. As shown in Table 2, our method again achieves the best performance across all metrics, reducing PA-MPJPE from 17.1 (EgoX) to 12.6 and improving F@15mm by 6.6 points. Qualitative examples of exo-to-ego pose translation are shown in Figure 3.

4.3 Policy Co-Training Evaluation

Simulation setup. We evaluate policy performance under co-training using the *Ego Humanoid Manipulation Benchmark* [68], built on NVIDIA Isaac Lab with a Unitree H1 humanoid robot equipped with two Inspire dexterous robotic hands. The benchmark comprises 12 manipulation tasks categorized as either *Short-*

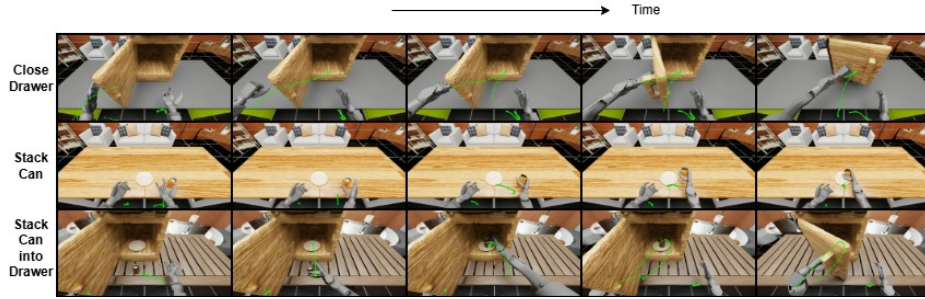


Fig. 4: Visualization of policy rollout after co-training.

Horizon atomic actions: Stack-Can, Push-Box, Open-Drawer, Close-Drawer, Flip-Mug, Pour-Balls, and Open-Laptop; or *Long-Horizon* multi-stage skills: Sort-Cans, Insert-Cans, Unload-Cans, Insert-And-Unload-Cans, and Stack-Can-Into-Drawer. Following [68], we evaluate under two visual settings: *Seen*, where backgrounds match training, and *Unseen*, where all 22 background combinations are previously unseen, with object positions randomized at the beginning of each rollout. We report *Success Rate* (SR) and *Progress Rate* (PSR), following [68]. Policy rollout visualizations in simulation setting are shown in Figure 4.

Baselines. We compare against four categories of methods, each probing a different aspect of the proposed pipeline. *Robot-only* (ACT [77] and EgoVLA-NoPretrain [68]) train solely on $\mathcal{D}_{\text{robot}}$ without any human data, establishing the data-scarce baseline. *Exo-only co-training* injects exocentric poses directly into co-training without any viewpoint translation, isolating whether the benefit comes from more data alone or from the viewpoint translation step itself. *Pixel-space co-training* methods (EgoWorld [44], 4Diff [11], EgoX [24]) first synthesize egocentric videos from exocentric input, extract poses via HAMER [45] and FoundationPose [62], and co-train the same policy, directly measuring whether pose *quality* of the translated poses affect policy learning. *Ego VLA* [68] pretrains on egocentric human videos, representing the strongest prior method that uses human data but avoids the exo-to-ego problem entirely. *Ours (Exo2EgoPolicy)* uses the full pipeline with all contributions.

Analysis. The row groupings in Table 3 and Table 4 reveal a clear and consistent performance hierarchy. First, Exo-only co-training provides negligible improvement over robot-only baselines (+1.1% mean SR on short-horizon Seen), confirming that raw exocentric poses are too geometrically misaligned with the egocentric robot frame to provide useful training signal; additional exocentric data alone does not solve the domain gap. Second, pixel-space view translation methods (EgoWorld, 4Diff, EgoX) improve substantially over robot-only, indicating that view translation is necessary; however, their gains are bottlenecked by pose extraction errors from newly synthesized video regions, with EgoX achieving 75.5% and 66.6% mean SR on short-horizon Seen and Unseen, respectively. Third, EgoVLA demonstrates that egocentric human pretraining is

Table 3: Evaluation on Short-Horizon Tasks (Seen / Unseen visual configurations). SR: Success Rate (%). PSR: Progress Rate (%). Best in **bold**, second best underlined.

Method	Stack-Can		Push-Box		Open-Drawer		Close-Drawer		Flip-Mug		Pour-Balls		Open-Laptop		Mean	
	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑
Seen Configuration																
<i>Without human data</i>																
ACT [77]	22.2	22.2	11.1	85.2	18.5	66.7	48.1	96.3	7.4	7.4	3.7	77.8	63.0	63.0	24.9	59.8
EgoVLA-NoPretrain [68]	55.6	55.6	51.9	75.9	59.3	77.8	100.0	100.0	3.7	3.7	85.2	93.8	96.3	96.3	64.6	71.9
<i>With human data — no view translation</i>																
Exo-only	57.4	57.4	53.7	77.8	61.1	79.6	100.0	100.0	5.6	5.6	86.1	94.4	96.3	96.3	65.7	72.9
<i>With human data — pixel-space view translation</i>																
EgoWorld [44]	63.0	63.0	59.3	80.2	64.8	81.5	100.0	100.0	22.2	22.2	87.0	94.9	97.2	97.2	70.5	77.0
4Diff [11]	66.7	66.7	62.0	81.5	66.7	82.4	100.0	100.0	29.6	29.6	87.9	95.4	98.1	98.1	72.9	79.1
EgoX [24]	70.4	70.4	66.7	82.4	66.7	83.3	100.0	100.0	37.0	37.0	88.9	95.8	99.1	99.1	75.5	81.1
<i>With human data — egocentric pretraining</i>																
EgoVLA [68]	<u>77.8</u>	<u>77.8</u>	<u>70.4</u>	<u>83.3</u>	<u>59.3</u>	<u>81.5</u>	<u>100.0</u>	<u>100.0</u>	<u>59.3</u>	<u>59.3</u>	<u>77.8</u>	<u>92.6</u>	<u>100.0</u>	<u>100.0</u>	<u>77.8</u>	<u>84.9</u>
<i>With human data — pose-space exo2ego view translation</i>																
Ours	82.4	82.4	75.9	87.0	66.7	85.2	100.0	100.0	63.0	63.0	85.2	94.4	100.0	100.0	81.9	87.4
Unseen Configuration																
<i>Without human data</i>																
ACT [77]	9.1	10.6	18.2	87.9	24.2	72.0	63.6	97.0	0.0	0.0	6.1	59.1	53.0	53.0	24.9	54.2
EgoVLA-NoPretrain [68]	57.6	57.6	69.7	82.6	46.2	71.3	86.4	89.9	4.7	4.7	46.0	79.4	48.5	53.0	51.3	62.6
<i>With human data — no view translation</i>																
Exo-only	59.1	59.1	71.2	83.8	47.4	72.6	87.9	91.4	5.8	5.8	47.4	80.6	50.0	54.5	52.7	63.9
<i>With human data — pixel-space view translation</i>																
EgoWorld [44]	62.1	62.1	73.1	84.9	50.5	74.8	90.4	93.2	16.7	16.7	56.1	83.4	60.6	65.2	58.5	68.6
4Diff [11]	63.6	63.6	74.2	85.6	51.3	75.8	93.2	95.5	21.8	21.8	65.2	87.9	68.2	72.7	62.5	71.8
EgoX [24]	65.2	65.2	75.8	86.4	52.6	77.3	95.5	97.0	27.4	27.4	74.2	91.7	75.8	80.3	66.6	75.0
<i>With human data — egocentric pretraining</i>																
EgoVLA [68]	<u>62.1</u>	<u>62.1</u>	<u>75.8</u>	<u>84.9</u>	<u>50.0</u>	<u>75.3</u>	<u>98.5</u>	<u>98.5</u>	<u>30.8</u>	<u>30.8</u>	<u>83.3</u>	<u>94.4</u>	<u>83.3</u>	<u>87.9</u>	<u>69.1</u>	<u>76.3</u>
<i>With human data — pose-space exo2ego view translation</i>																
Ours	68.2	68.2	80.3	88.6	56.4	79.5	100.0	100.0	35.9	35.9	87.5	96.5	87.9	91.7	73.7	80.1

a strong source of diverse priors, outperforming all pixel-space methods on most tasks. Crucially, our method surpasses EgoVLA on both short- and long-horizon tasks under both visual settings, demonstrating that exo-to-ego pose translation extracts higher-quality and more spatially diverse training signal than egocentric video alone. The advantage is largest on long-horizon Seen tasks (+7.4% SR), where temporally consistent pose sequences from our dynamics prior directly reduce compounding errors across subtask boundaries, and on Unseen configurations (+7.3% SR), confirming that our geometry-aligned pose representation generalizes more robustly. Additional findings are in the supplementary material.

4.4 Ablation Study

To isolate the contribution of each design choice in Exo2EgoPolicy, we ablate four design decisions on both pose translation quality (PA-MPJPE↓, F@15mm↑) and downstream policy performance (mean SR↑ on short- and long-horizon tasks under the Unseen configuration), as shown in Table 5.

Object-relative formulation: Removing the object-relative coordinate frame (row a) and predicting poses in absolute world coordinates causes a substantial drop in both PA-MPJPE (11.4→18.4) and short-horizon SR (73.7→58.3%). Without the object as a locally stable reference, the model must implicitly disentangle camera motion from hand-object kinematics, which is ill-posed under uncalibrated exocentric cameras and leads to systematically misaligned poses.

Geometric latent encoding: Replacing the geometry-aware latent representation z_t with direct regression (row b) produces the lowest pose accuracy overall (PA-MPJPE: 21.6), confirming that the explicit geometric encoding is essential for resolving viewpoint ambiguity rather than memorizing appearance-to-pose mappings.

Autoregressive conditioning: Removing p_{t-1} from the score network (row c) eliminates temporal context entirely, degrading both pose quality and policy performance. This confirms that conditioning on the previous state is

Table 4: Evaluation on Long-Horizon Tasks (Seen / Unseen visual configurations). SR: Success Rate (%). PSR: Progress Rate (%). Best in **bold**, second best underlined.

Method	Insert-&-Unload		Stack-Into-Drawer		Sort-Cans		Unload-Cans		Insert-Cans		Mean	
	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑	SR↑	PSR↑
Seen Configuration												
<i>Without human data</i>												
ACT [77]	0.0	11.1	11.1	37.0	0.0	28.6	0.0	35.8	0.0	19.8	2.2	26.5
EgoVLA-NoPretrain [68]	7.4	45.9	0.0	18.5	51.9	79.6	63.0	75.0	11.1	55.6	26.7	54.9
<i>With human data - no view translation</i>												
Exo-only	8.6	47.2	0.0	20.4	53.1	80.2	64.2	76.4	12.3	57.4	27.6	56.3
<i>With human data - pixel-space exo2ego view translation</i>												
EgoWorld [44]	18.5	54.3	14.8	48.1	55.6	82.4	66.7	78.2	14.8	62.3	34.1	65.1
4Diff [11]	22.2	58.6	22.2	55.6	57.4	84.7	68.5	80.1	16.7	65.8	37.4	68.9
EgoX [24]	29.6	64.2	29.6	62.4	59.3	86.8	70.4	81.9	18.5	70.2	41.5	73.1
<i>With human data - egocentric pretraining</i>												
EgoVLA [68]	<u>44.4</u>	<u>77.0</u>	<u>40.7</u>	<u>75.9</u>	<u>55.6</u>	<u>88.9</u>	<u>66.7</u>	<u>83.3</u>	<u>22.2</u>	<u>78.7</u>	<u>45.9</u>	<u>80.8</u>
<i>With human data - pose-space exo2ego view translation (ours)</i>												
Ours	51.9	81.5	48.1	79.6	63.0	91.2	74.1	87.0	29.6	82.4	53.3	84.3
Unseen Configuration												
<i>Without human data</i>												
ACT [77]	0.0	7.9	1.5	33.8	0.0	20.7	1.5	33.8	0.0	21.2	0.6	23.5
EgoVLA-NoPretrain [68]	0.0	29.1	0.0	20.1	15.2	45.8	34.9	55.7	6.1	30.3	11.2	36.2
<i>With human data - no view translation</i>												
Exo-only	0.0	30.3	0.0	21.6	16.7	47.2	36.4	57.2	7.6	31.8	12.1	37.6
<i>With human data - pixel-space exo2ego view translation</i>												
EgoWorld [44]	6.1	38.6	6.1	38.6	18.2	55.3	40.9	62.1	9.1	40.2	16.1	46.9
4Diff [11]	12.1	48.5	12.1	45.5	19.7	60.6	43.9	66.7	10.6	48.5	19.7	53.9
EgoX [24]	21.2	62.1	18.2	54.5	21.2	65.9	47.0	71.2	12.1	56.1	23.9	61.9
<i>With human data - egocentric pretraining</i>												
EgoVLA [68]	<u>31.8</u>	<u>77.0</u>	<u>28.8</u>	<u>61.0</u>	<u>18.2</u>	<u>68.9</u>	<u>50.0</u>	<u>75.8</u>	<u>15.2</u>	<u>62.9</u>	<u>28.8</u>	<u>69.1</u>
<i>With human data - pose-space exo2ego view translation (ours)</i>												
Ours	39.4	82.6	36.4	67.4	25.8	74.2	57.6	80.3	21.2	68.2	36.1	74.5

Table 5: Ablation study on H2O [26] (pose translation, unseen objects) and the *Ego Humanoid Manipulation Benchmark* [68] (policy, Unseen).

Method	Pose Translation		Policy (Unseen)	
	PA-MPJPE↓	F@15mm↑	Short SR↑	Long SR↑
<i>Representation ablations</i>				
(a) w/o object-relative formulation	18.4	0.712	58.3	20.1
(b) w/o geometric latent z_t (direct regression)	21.6	0.671	54.7	17.4
<i>Temporal ablations</i>				
(c) w/o autoregressive conditioning (p_{t-1} removed from s_ψ)	17.1	0.783	62.4	24.8
(d) w/o $\mathcal{L}_{\text{temp}}$ (full temporal module)	19.3	0.748	57.9	19.8
Ours (full)	11.4	0.912	73.7	36.1

necessary even before any explicit smoothness supervision is applied. **Temporal consistency loss:** Removing the full temporal module (row e) collapses performance to near the representation-only baseline, confirming that $\mathcal{L}_{\text{temp}}$ suppresses high-frequency jitter, which is necessary for stable long-horizon execution.

5 Conclusion

We present *Exo2EgoPolicy*, a framework that translates exocentric human demonstrations into egocentric robot actions directly in pose space, avoiding artifacts introduced by pixel-space view synthesis. A viewpoint-conditioned diffusion model on $SE(3)$, coupled with a temporally regularized dynamics prior, produces smooth, geometrically consistent pose trajectories that serve as a sufficient representation for policy learning. Experiments on pose translation benchmarks and the Ego Humanoid Manipulation Benchmark demonstrate that exocentric human video is a scalable and effective source of manipulation knowledge, with consistent gains over all baselines especially on long-horizon tasks.

Acknowledgments

Zhaozheng Yin was supported by NSF grants IIS-2331769, CMMI-2246673, and ECCS-2025929.

References

1. Arunachalam, S.P., Silwal, S., Evans, B., Pinto, L.: Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 5954–5961. IEEE (2023)
2. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13778–13790 (2023)
3. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22875–22889 (2025)
4. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
5. Barfoot, T.D.: State estimation for robotics. Cambridge University Press (2024)
6. Bharadhwaj, H., Mottaghi, R., Gupta, A., Tulsiani, S.: Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. arXiv preprint arXiv:2405.01527 **2** (2024)
7. 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. arXiv preprint arXiv:2410.24164 (2024)
8. Chen, A.S., Nair, S., Finn, C.: Learning generalizable robotic reward functions from "in-the-wild" human videos. arXiv preprint arXiv:2103.16817 (2021)
9. Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., et al.: Sana-video: Efficient video generation with block linear diffusion transformer. arXiv preprint arXiv:2509.24695 (2025)
10. Chen, L.Y., Hari, K., Dharmarajan, K., Xu, C., Vuong, Q., Goldberg, K.: Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. arXiv preprint arXiv:2402.19249 (2024)
11. Cheng, F., Luo, M., Wang, H., Dimakis, A., Torresani, L., Bertasius, G., Grauman, K.: 4diff: 3d-aware diffusion model for third-to-first viewpoint translation. In: European Conference on Computer Vision. pp. 409–427. Springer (2024)
12. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025)
13. Chi, C., Xu, Z., Pan, C., Cousineau, E., Burchfiel, B., Feng, S., Tedrake, R., Song, S.: Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. arXiv preprint arXiv:2402.10329 (2024)
14. Das, N., Bechtle, S., Davchev, T., Jayaraman, D., Rai, A., Meier, F.: Model-based inverse reinforcement learning from visual demonstrations. In: Conference on Robot Learning. pp. 1930–1942. PMLR (2021)
15. Fang, B., Jia, S., Guo, D., Xu, M., Wen, S., Sun, F.: Survey of imitation learning for robotic manipulation: B. fang et al. International Journal of Intelligent Robotics and Applications **3**(4), 362–369 (2019)
16. Fu, Z., Zhao, Q., Wu, Q., Wetzstein, G., Finn, C.: Humanplus: Humanoid shadowing and imitation from humans. arXiv preprint arXiv:2406.10454 (2024)

17. Ghadirzadeh, A., Chen, X., Poklukar, P., Finn, C., Björkman, M., Kragic, D.: Bayesian meta-learning for few-shot policy adaptation across robotic platforms. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1274–1280. IEEE (2021)
18. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
19. Handa, A., Van Wyk, K., Yang, W., Liang, J., Chao, Y.W., Wan, Q., Birchfield, S., Ratliff, N., Fox, D.: Dexpivot: Vision-based teleoperation of dexterous robotic hand-arm system. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9164–9170. IEEE (2020)
20. Hoque, R., Huang, P., Yoon, D.J., Sivapurapu, M., Zhang, J.: Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709 (2025)
21. Hsu, C.C., Wen, B., Xu, J., Narang, Y., Wang, X., Zhu, Y., Biswas, J., Birchfield, S.: Spot: Se (3) pose trajectory diffusion for object-centric manipulation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 4853–4860. IEEE (2025)
22. Hussein, A., Gaber, M.M., Elyan, E., Jayne, C.: Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* **50**(2), 1–35 (2017)
23. Jung, B., Amor, H.B., Heumer, G., Weber, M.: From motion capture to action capture: A review of imitation learning techniques and their application to vr-based character animation. In: Proceedings of the ACM symposium on Virtual reality software and technology. pp. 145–154 (2006)
24. Kang, T., Kim, K., Kim, D., Park, M., Hyung, J., Choo, J.: Egox: Egocentric video generation from a single exocentric video. arXiv preprint arXiv:2512.08269 (2025)
25. Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., Hoffman, J., Xu, D.: Egomimic: Scaling imitation learning via egocentric video. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 13226–13233. IEEE (2025)
26. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10138–10148 (2021)
27. Li, J., Zhu, Y., Xie, Y., Jiang, Z., Seo, M., Pavlakos, G., Zhu, Y.: Okami: Teaching humanoid robots manipulation skills through single video imitation. arXiv preprint arXiv:2410.11792 (2024)
28. Liu, J.W., Mao, W., Xu, Z., Keppo, J., Shou, M.Z.: Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems* **37**, 136149–136172 (2024)
29. Liu, J.W., Mao, W., Xu, Z., Keppo, J., Shou, M.Z.: Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems* **37**, 136149–136172 (2024)
30. Liu, J.W., Mao, W., Xu, Z., Keppo, J., Shou, M.Z.: Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems* **37**, 136149–136172 (2024)
31. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: Rdt-1b: a diffusion foundation model for bimanual manipulation. arXiv preprint arXiv:2410.07864 (2024)
32. Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., Yi, L.: Taco: Benchmarking generalizable bimanual tool-action-object understanding. In: Proceedings

- of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21740–21751 (2024)
33. Lu, C., Cheng, X., Li, J., Yang, S., Ji, M., Yuan, C., Yang, G., Yi, S., Wang, X.: Mobile-television: Predictive motion priors for humanoid whole-body control. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 5364–5371. IEEE (2025)
 34. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In: European Conference on Computer Vision. pp. 407–425. Springer (2024)
 35. Luo, Y., Shi, X., Bai, J., Xia, M., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Camclonemaster: Enabling reference-based camera control for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–10 (2025)
 36. Ma, Y.J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., Zhang, A.: Vip: Towards universal visual reward and representation via value-implicit pre-training. arXiv preprint arXiv:2210.00030 (2022)
 37. Mao, X., Lin, S., Li, Z., Li, C., Peng, W., He, T., Pang, J., Chi, M., Qiao, Y., Zhang, K.: Yume: An interactive world generation model. arXiv preprint arXiv:2507.17744 (2025)
 38. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3m: A universal visual representation for robot manipulation. arXiv preprint arXiv:2203.12601 (2022)
 39. Nguyen, A., Kanoulas, D., Muratore, L., Caldwell, D.G., Tsagarakis, N.G.: Translating videos to commands for robotic manipulation with deep recurrent neural networks. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3782–3788. IEEE (2018)
 40. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12999–13008 (2023)
 41. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article **2**(5), 1 (2023)
 42. 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)
 43. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: Handocnet: Occlusion-robust 3d hand mesh estimation network. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 1486–1495. IEEE (2022)
 44. Park, J., Ye, A.S., Kwon, T.: Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. arXiv preprint arXiv:2506.17896 (2025)
 45. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: CVPR (2024)
 46. Qiu, R.Z., Yang, S., Cheng, X., Chawla, C., Li, J., He, T., Yan, G., Yoon, D.J., Hoque, R., Paulsen, L., et al.: Humanoid policy~ human policy. arXiv preprint arXiv:2503.13441 (2025)
 47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
 48. Ren, J., Sundaresan, P., Sadigh, D., Choudhury, S., Bohg, J.: Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8802–8810. IEEE (2025)

49. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6121–6132 (2025)
50. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **36**(6) (Nov 2017)
51. Rothfuss, J., Ferreira, F., Aksoy, E.E., Zhou, Y., Asfour, T.: Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters* **3**(4), 4007–4014 (2018)
52. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21096–21106 (2022)
53. Shankar, T., Lin, Y., Rajeswaran, A., Kumar, V., Anderson, S., Oh, J.: Translating robot skills: Learning unsupervised skill correspondences across robots. In: *International Conference on Machine Learning*. pp. 19626–19644. PMLR (2022)
54. Sola, J., Deray, J., Atchuthan, D.: A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537* (2018)
55. Sullivan, J.V.: Learning and embodied cognition: A review and proposal. *Psychology Learning & Teaching* **17**(2), 128–143 (2018)
56. 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)
57. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025)
58. Wang, C., Fan, L., Sun, J., Zhang, R., Fei-Fei, L., Xu, D., Zhu, Y., Anandkumar, A.: Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422* (2023)
59. Wang, C., Shi, H., Wang, W., Zhang, R., Fei-Fei, L., Liu, C.K.: Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788* (2024)
60. Wang, L., Chen, X., Zhao, J., He, K.: Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in neural information processing systems* **37**, 124420–124450 (2024)
61. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 8798–8807 (2018)
62. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. p. 17868–17879. IEEE (2024)
63. Xiong, H., Li, Q., Chen, Y.C., Bharadhwaj, H., Sinha, S., Garg, A.: Learning by watching: Physical imitation of manipulation skills from human videos. In: *2021 IEEE/RSJ international conference on intelligent robots and systems (iros)*. pp. 7827–7834. IEEE (2021)
64. Xu, J., Huang, Y., Pei, B., Hou, J., Li, Q., Chen, G., Zhang, Y., Feng, R., Xie, W.: Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732* (2025)

65. Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U.: Point-nerf: Point-based neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5438–5448 (2022)
66. Xue, B., Duan, Z.P., Yan, Q., Wang, W., Liu, H., Guo, C.L., Li, C., Li, C., Lyu, J.: Stand-in: A lightweight and plug-and-play identity control for video generation. arXiv preprint arXiv:2508.07901 (2025)
67. Yang, J., Glossop, C., Bhorkar, A., Shah, D., Vuong, Q., Finn, C., Sadigh, D., Levine, S.: Pushing the limits of cross-embodiment learning for manipulation and navigation. arXiv preprint arXiv:2402.19432 (2024)
68. Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.C., Zou, X., Fang, Y., Cheng, X., Qiu, R.Z., et al.: Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440 (2025)
69. Yang, S., Liu, M., Qin, Y., Ding, R., Li, J., Cheng, X., Yang, R., Yi, S., Wang, X.: Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. arXiv preprint arXiv:2408.11805 (2024)
70. Yang, Y., Li, Y., Fermuller, C., Aloimonos, Y.: Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29 (2015)
71. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
72. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4578–4587 (2021)
73. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 100–111 (2025)
74. Ze, Y., Chen, Z., Wang, W., Chen, T., He, X., Yuan, Y., Peng, X.B., Wu, J.: Generalizable humanoid manipulation with 3d diffusion policies. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2873–2880. IEEE (2025)
75. Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., Lu, C.: Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 445–456 (2024)
76. Zhang, G., Zhong, L., Lee, Y., Lim, J.J.: Policy transfer across visual and dynamics domain gaps via iterative grounding. arXiv preprint arXiv:2107.00339 (2021)
77. Zhao, T.Z., Kumar, V., Levine, S., Finn, C.: Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705 (2023)
78. Zhao, T.Z., Tompson, J., Driess, D., Florence, P., Ghasemipour, K., Finn, C., Wahid, A.: Aloha unleashed: A simple recipe for robot dexterity. arXiv preprint arXiv:2410.13126 (2024)
79. Zhu, Y., Feng, F.: Let me show you: Learning by retrieving from egocentric video for robotic manipulation. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9824–9829. IEEE (2025)
80. Zhu, Y., Joshi, A., Stone, P., Zhu, Y.: Viola: Imitation learning for vision-based manipulation with object proposal priors. In: Conference on Robot Learning. pp. 1199–1210. PMLR (2023)