REFERENCE-CONDITIONED CROSS-EMBODIMENT VIDEO GENERATION WITHOUT 3D ROBOT ASSETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human and robot demonstration videos provide a growing source of manipulation experience, yet each observation couples the interaction with the demonstrator's body. Reusing this experience across embodiments requires the target robot's configuration and occlusions to adapt to its morphology and the surrounding scene. This work presents EGO2ANY, a reference-conditioned video diffusion model that generates corresponding realizations of egocentric manipulation interactions across robot embodiments. It represents the demonstrated interaction separately from target morphology, with morphology-conditioned temporal transport to infer target-specific body layouts throughout the interaction. These layouts guide trajectory-conditioned video synthesis to preserve the source interaction and scene context while remaining aligned with the source actions. EGO2ANY supports embodiment changes within a shared model without requiring target meshes, kinematic models, or a separately prepared target initial frame at inference. We further construct EGOXPAIR, an action-paired dataset containing 1.0M retained videos across 10 target embodiments. Evaluations show that the generated demonstrations improve downstream policy learning, supporting cross-embodiment video generation as reusable manipulation experience. Our code and dataset will be released.

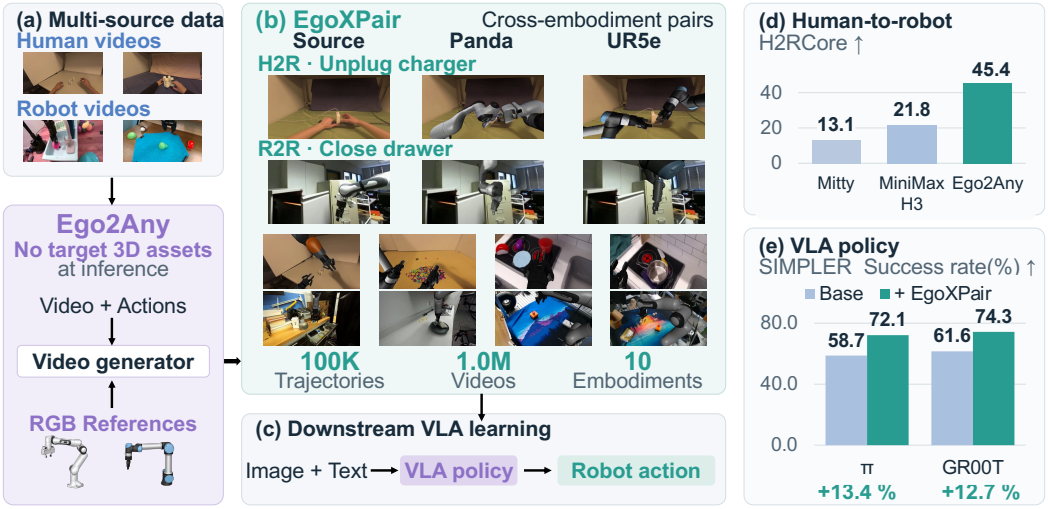

Figure 1: **EGO2ANY** transfers human and robot demonstrations using RGB references without target 3D assets at inference. The resulting EGOXPAIR improves VLA learning.

1 INTRODUCTION

Egocentric manipulation experience from humans and robots provides a growing source of supervision for learning robotic skills (Grauman et al., 2022; Nair et al., 2023). However, each demonstration couples task-relevant interaction with the demonstrator's morphology and motion, requiring similar experience to be collected or reconstructed separately for different robot embodiments. This coupling limits the reuse of growing demonstration collections across robotic platforms. Cross-embodiment

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

transfer offers a path toward scaling such experience by realizing the same underlying interaction through different robot embodiments (Brohan et al., 2023; Walke et al., 2023).

However, scaling cross-embodiment transfer faces two challenges. ❶ **Dependence on target-specific 3D assets.** Geometry-based retargeting pipelines often rely on robot meshes, kinematic models, or rendering procedures (Li et al., 2025a; Wang et al., 2026) to realize the target embodiment. Preparing the assets and configuring associated pipelines introduces embodiment-specific overhead when extending existing demonstrations to additional robotic platforms. ❷ **General embodiment adaptation.** Scalable demonstration reuse calls for flexible adaptation of the same source interaction to arbitrary target robots. Separating task and embodiment representations (Li et al., 2026; Chen et al., 2026) supports such recombination. Yet the same end-effector trajectory can correspond to different full-body configurations across target morphologies. Realistic transfer requires jointly inferring the body layouts, spatial occupancy, and occlusion patterns of target robots throughout the interaction.

To address these challenges, we present EGO2ANY, a reference-conditioned video diffusion framework for transferring egocentric manipulation interactions across robot embodiments. To support flexible source-target composition, we construct *separate representations* for source interactions and target morphology. This includes a source encoder to capture effector–object relations and dense scene evidence, and a target encoder to aggregate multi-view references into local morphology descriptors. The same source interaction can thus be paired with different target embodiments by switching references within the shared model (§3.1). Building on the decoupled representations, *morphology-conditioned temporal transport* combines source actions, interaction context, and target morphology to predict target-body layouts over time. It reuses local appearance descriptors from the target references across frames, while predicting where the corresponding body regions appear and how their shape and visibility change as the robot moves. The resulting layouts capture the target body configuration, spatial occupancy, and occlusion patterns within the source scene (§3.2). Finally, *trajectory-guided video generation* directly modulates the shared video backbone with source actions, while separate conditioning pathways introduce the predicted layouts, source scene features, and target appearance (§3.3). Together, EGO2ANY is able to generate target-robot videos that preserve the demonstrated interaction and scene context while remaining aligned with the source action trajectory.

EGO2ANY provides a flexible interface for cross-embodiment data generation. The morphology-conditioned temporal transport directly predicts target-body layouts from source interactions and target references, which requires no target 3D assets during inference. Moreover, by conditioning the generation with predicted layouts, EGO2ANY eliminates the need to prepare an initial frame depicting target robots for each source demonstration (Chen et al., 2026). Simply switching target references within the shared model allows the same source interaction to be synthesized across embodiments. On this basis, we construct EGOXPAIR, a paired cross-embodiment interaction dataset built from $\mathbf{100}$K unique source trajectories, yielding $\mathbf{1.0}$M retained videos for $\mathbf{10}$ target embodiments after quality screening. Each video sample is paired with end-effector actions, gripper states, language instructions, embodiment identifiers, and provenance, with all embodiment variants grouped by their source demonstration. EGOXPAIR broadens embodiment coverage for VLA pretraining, visual generalization studies, and downstream policy learning.

We evaluate EGO2ANY in terms of both cross-embodiment generation quality and downstream learning utility. Across $\mathbf{100}$ human and $\mathbf{90}$ robot source demonstrations, each paired with $\mathbf{4}$ target embodiments, EGO2ANY achieves an H2RCore of $\mathbf{45.4}$ versus $\mathbf{21.8}$ for the strongest baseline, and an R2RCore of $\mathbf{52.4}$ versus $\mathbf{20.0}$. Its target-embodiment scores reach $\mathbf{43.1}$ and $\mathbf{50.0}$, exceeding the strongest baseline in each direction by $\mathbf{26.4}$ and $\mathbf{40.6}$ points, respectively. In closed-loop policy learning evaluations, augmenting StarVLA-π (StarVLA Community, 2026; Black et al., 2025) and StarVLA-GR00T (NVIDIA et al., 2025) with EGOXPAIR improves average success on SIMPLER (Li et al., 2025c) by $\mathbf{13.4}\%$ and $\mathbf{12.7}\%$, respectively. Our code and dataset will be released.

2 RELATED WORK

Cross-embodiment transfer. Cross-embodiment learning seeks to reuse task-relevant experience across agents with different appearances, kinematics, and action spaces. One line of work learns embodiment-agnostic visual or task-progress representations, allowing demonstrations to be related beyond the identity of the acting agent (Zhou et al., 2021; Zakka et al., 2022). Other approaches discover reusable skills from unaligned human and robot videos (Xu et al., 2023; Kim et al., 2025a), or

represent behavior through transferable latent and geometric actions (Ye et al., 2025; Bu et al., 2025; Tong et al., 2026). In parallel, large-scale policy learning shares experience across heterogeneous robots and action interfaces within generalist models (Open X-Embodiment Collaboration, 2024; Ghosh et al., 2024). These approaches primarily organize transfer in a learned representation or policy space. EGO2ANY provides a complementary form of experience reuse at the observation level. It synthesizes how the same demonstrated interaction would appear when carried out by a reference-specified robot, while retaining the source scene, object evolution, viewpoint, and timeline.

Human-to-robot observation translation. Human videos offer a scalable source of supervision for robot learning (Jeong et al., 2026), but the visual mismatch with robot observations hinders their direct reuse. Retargeting bridges this gap by combining hand-pose estimation, inpainting, and robot rendering to replace the human actor while retaining task-relevant scene content (Lepert et al., 2025; Li et al., 2025a; Lepert et al., 2026; Wang et al., 2026). RoboEdit combines 3D retargeting with video editing (Guo et al., 2026b). In these pipelines, target-body realizations depend on target kinematics, rendering assets, or prescribed motion. On the other hand, video diffusion methods learn the observation transformation more directly, drawing on synchronized human–robot demonstrations (Xie et al., 2026), shared pose conditions (Ci et al., 2025), synthetic video pairs (Song et al., 2025), or joint visual and action alignment (Li et al., 2025b). Related work separates task from embodiment in video editing (Li et al., 2026) and adapts motion transfer to new humanoids using unpaired video (Song et al., 2026). EGO2ANY instead predicts target-body layouts over time from multi-view references and source actions, then conditions video synthesis on them.

Controllable robot world models. Robot video models predict action-conditioned observations (Finn et al., 2016; Ebert et al., 2018; Zhu et al., 2025; Guo et al., 2026a). Cross-embodiment extensions incorporate kinematic 4D conditions (Xu et al., 2026), rendered skeletons (Wu & Gao, 2026), explicit robot geometry (Kim et al., 2026), multiple action interfaces (Liu & Shorinwa, 2026), or embodiment-agnostic contact trajectories (Azirar et al., 2026). AnyWorld recomposes egocentric interactions using a skeleton-control video, camera trajectory, edited target initial frame, and embodiment tag (Chen et al., 2026). EGO2ANY instead infers target-body layouts from source actions and multi-view RGB references to guide video synthesis, requiring no target meshes, kinematic models, online rendering, or a separately prepared target initial frame for each source demonstration at inference.

3 METHOD

Problem Formulation. Cross-embodiment interaction transfer aims to realize a demonstrated interaction with a different robot embodiment while preserving its manipulation behavior. Given a source egocentric RGB sequence $\mathcal{X} = \{\boldsymbol{x}_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$ with resolution $H \times W$ over T frames, its associated action trajectory $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_t \in \mathbb{R}^{2\times 10}\}_{t=1}^{T}$, camera intrinsics $\boldsymbol{K} \in \mathbb{R}^{3\times 3}$, and V action-independent references $\mathcal{R} = \{\boldsymbol{r}_v \in \mathbb{R}^{H\times W\times 3}\}_{v=1}^{V}$ of a target robot, the goal is to synthesize a target sequence $\mathcal{Y} = \{\boldsymbol{y}_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$ that realizes the interaction with the target embodiment while preserving scene context, viewpoint, and temporal progression. Each $\boldsymbol{\alpha}_t$ contains two effector slots, each encoding a 3D tool-center-point (TCP) position, a 6D orientation, and grasp openness.

Overview. As shown in Fig. 2, EGO2ANY first encodes the source interaction and scene context together with multi-view target references to obtain complementary representations of the demonstrated interaction and target embodiment (§3.1). It then performs morphology-conditioned temporal transport to infer the target-specific spatial realization of the interaction over time (§3.2). Finally, the resulting spatial and motion conditions, together with source scene context and target appearance, guide a video diffusion model to synthesize the interaction with the target embodiment (§3.3).

3.1 SOURCE AND TARGET REPRESENTATIONS

Cross-embodiment generation requires preserving what happens in the source interaction while adapting how it is realized by the target robot. Accordingly, the source branch represents interaction dynamics, action trajectories, and scene context, while the target branch extracts morphology and appearance from action-independent multi-view references. Together, they provide source dynamics and target structure for the subsequent morphology-conditioned temporal transport stage.

Source Interaction and Scene. Given the source sequence $\mathcal{X}$, a scene encoder $\mathcal{F}^S$, comprising a shared RGB patch encoder and a temporal transformer, produces scene tokens $\boldsymbol{S} \in \mathbb{R}^{T\times N\times D}$, with

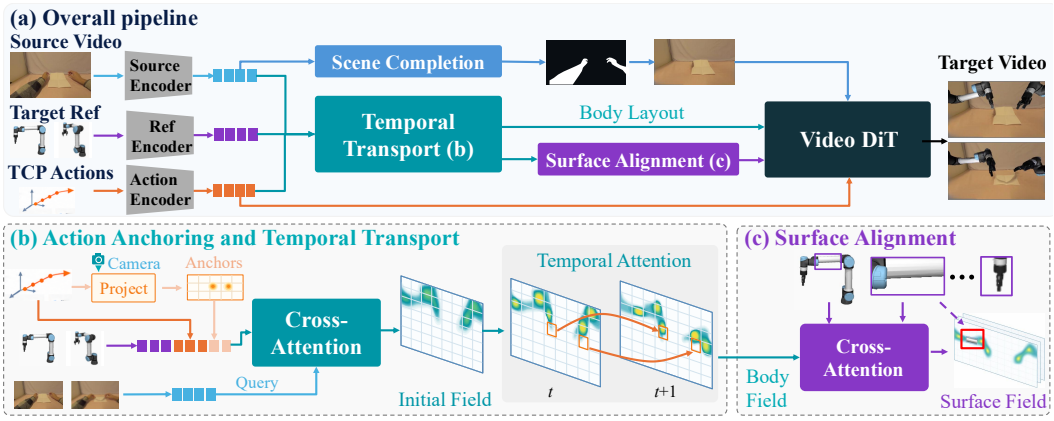

Figure 2: **Schematic overview of EGO2ANY. (a)** Source video, actions, and target references guide video diffusion. **(b)** Action anchoring, placement, and temporal transport predict body layouts. **(c)** Cross-attention aligns reference patches with the body field to produce surface features.

N tokens of dimension D per frame. To extract interaction-centric information, an interaction encoder $\mathcal{F}^A$ employs three learned queries $\boldsymbol{Q}^A \in \mathbb{R}^{3\times D}$ for the two effectors and the manipulated object. The queries localize these entities and aggregate their local appearance, object-relative geometry, temporal motion, and visibility into interaction tokens $\boldsymbol{A} \in \mathbb{R}^{T\times D}$ summarizing the interaction:

$$\boldsymbol{S} = \mathcal{F}^S(\mathcal{X}), \qquad \boldsymbol{A} = \mathcal{F}^A(\boldsymbol{S}, \boldsymbol{Q}^A). \tag{1}$$

In parallel, the action trajectory $\boldsymbol{\alpha}$ is encoded into command tokens $\boldsymbol{C} \in \mathbb{R}^{T\times D}$ that represent endpoint motion and grasp states. To preserve scene evolution while removing the source embodiment, a body-mask predictor $\mathcal{F}^M$ estimates a soft mask sequence $\mathcal{M} = \{\boldsymbol{m}_t \in [0,1]^{H\times W}\}_{t=1}^T$ from $\mathcal{X}$. A completion network $\mathcal{F}^B$ fills the masked regions using the corresponding scene tokens $\boldsymbol{S}_t \in \mathbb{R}^{N\times D}$, yielding the complete body-removed scene carrier $\mathcal{B} = \{\boldsymbol{b}_t \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^T$:

$$\boldsymbol{m}_t = \mathcal{F}^M(\boldsymbol{x}_t), \qquad \boldsymbol{b}_t = \mathcal{F}^B\big((1-\boldsymbol{m}_t)\odot \boldsymbol{x}_t, \boldsymbol{m}_t, \boldsymbol{S}_t\big), \tag{2}$$

where $\odot$ denotes element-wise multiplication. The resulting carrier preserves the manipulated object and surrounding scene while eliminating source-body appearance from subsequent generation.

Target Morphology and Appearance. The target branch characterizes robot morphology and appearance from the action-independent references $\mathcal{R}$. A reference encoder $\mathcal{F}^R$ extracts patch-level features from each reference image and augments them with spatial coordinates, acquisition-slot embeddings, and foreground cues, producing reference tokens $\boldsymbol{P}_0 \in \mathbb{R}^{J\times D}$ with J tokens in total. To summarize target morphology, a morphology resampler $\mathcal{F}^G$ employs L learned queries $\boldsymbol{Q}^G \in \mathbb{R}^{L\times D}$ to aggregate $\boldsymbol{P}_0$ into morphology tokens $\boldsymbol{G} \in \mathbb{R}^{L\times D}$. A residual cross-attention projection $\mathcal{F}^P$ further refines the reference tokens into appearance-preserving patch memory $\boldsymbol{P} \in \mathbb{R}^{J\times D}$:

$$\boldsymbol{G} = \mathcal{F}^G(\boldsymbol{Q}^G, \boldsymbol{P}_0), \qquad \boldsymbol{P} = \boldsymbol{P}_0 + \mathcal{F}^P(\boldsymbol{P}_0, \boldsymbol{G}). \tag{3}$$

The morphology tokens $\boldsymbol{G}$ encode target structure for subsequent temporal transport, while the patch memory $\boldsymbol{P}$ retains fine-grained local appearance cues for faithful target video generation.

3.2 MORPHOLOGY-CONDITIONED TEMPORAL TRANSPORT

The same endpoint motion can induce substantially different whole-body realizations across robot embodiments, including body configuration, projected support, and occlusion. EGO2ANY therefore treats the source action as a sparse interaction constraint and uses target morphology to infer its embodiment-specific spatial realization over time. The transport module first anchors the commanded motion in the image plane, then predicts and temporally refines the target-body layout using the source interaction, scene context, and target morphology throughout the complete interaction sequence.

Action Anchoring and Placement. The action trajectory $\boldsymbol{\alpha}$ specifies endpoint motion in the source-camera coordinate system. For effector $i \in \{1,2\}$, its 3D TCP position $\boldsymbol{e}_{t,i} \in \mathbb{R}^3$ at frame t is projected by the camera intrinsics $\boldsymbol{K}$ to an image coordinate $\boldsymbol{u}_{t,i} \in \mathbb{R}^2$. The projected endpoint is rasterized into a Gaussian heatmap $\boldsymbol{E}_{t,i} \in [0,1]^{h\times w}$ on an $h \times w$ spatial grid:

$$\boldsymbol{u}_{t,i} = \Pi(\boldsymbol{K}\boldsymbol{e}_{t,i}), \qquad \boldsymbol{E}_{t,i} = \mathcal{E}_\sigma(\boldsymbol{u}_{t,i}), \tag{4}$$

where Π denotes perspective projection and $\mathcal{E}_\sigma$ denotes Gaussian rasterization with spread σ. Stacking the two endpoint heatmaps gives the spatial action anchors $\boldsymbol{E}_t \in [0,1]^{h\times w\times 2}$. Using source scene tokens $\boldsymbol{S}$ as queries, cross-attention integrates the endpoint anchors, full command sequence $\boldsymbol{C}$, and target reference memory $\boldsymbol{P}$ to produce an initial spatial field $\boldsymbol{\Phi}^0 \in \mathbb{R}^{h\times w\times D}$. The complete action trajectory defines the commanded workspace, while the target references determine embodiment-dependent placement and scale within the surrounding source scene.

Temporal Body Layout. The initial spatial field $\boldsymbol{\Phi}^0$ establishes target placement but does not capture body evolution over time. Stacking the frame-wise endpoint anchors gives $\boldsymbol{E} \in [0,1]^{T\times h\times w\times 2}$. A temporal transport network $\mathcal{F}^\Phi$ expands $\boldsymbol{\Phi}^0$ over T frames and jointly integrates the spatial anchors $\boldsymbol{E}$, command tokens $\boldsymbol{C}$, interaction tokens $\boldsymbol{A}$, scene tokens $\boldsymbol{S}$, and morphology tokens $\boldsymbol{G}$:

$$\boldsymbol{\Phi} = \mathcal{F}^\Phi\big(\boldsymbol{\Phi}^0, \boldsymbol{E}, \boldsymbol{C}, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{G}\big), \qquad \boldsymbol{\Phi} \in \mathbb{R}^{T\times h\times w\times D}. \tag{5}$$

Within $\mathcal{F}^\Phi$, frame-aligned action and source features refine each frame's spatial layout, while temporal attention across corresponding grid locations propagates body configuration through time. Conditioning on $\boldsymbol{G}$ adapts this evolution to the target morphology, producing a coherent target-body layout aligned with the demonstrated interaction while preserving temporal consistency across frames.

Reference-conditioned Surface Support. The temporal body field $\boldsymbol{\Phi}$ captures the evolving target-body layout but does not explicitly preserve fine-grained surface cues from the target references across the interaction sequence. A surface support module $\mathcal{F}^Z$ therefore associates $\boldsymbol{\Phi}$ with the morphology tokens $\boldsymbol{G}$ and patch memory $\boldsymbol{P}$, producing a reference-conditioned surface field $\boldsymbol{Z} \in \mathbb{R}^{T\times h\times w\times D}$:

$$\boldsymbol{Z} = \mathcal{F}^Z\big(\boldsymbol{\Phi}, \boldsymbol{G}, \boldsymbol{P}\big). \tag{6}$$

During training, part-level supervision encourages consistent correspondence between reference regions and target-body regions across time. Auxiliary heads further predict soft robot occupancy, normalized depth and validity, and base and endpoint heatmaps. Together with $\boldsymbol{\Phi}$ and $\boldsymbol{Z}$, these cues provide structured spatial and appearance guidance for subsequent video generation.

3.3 Conditioned Video Generation

The temporal transport stage provides target-specific spatial and surface guidance, while the source representations preserve scene context and demonstrated motion. These complementary conditions are integrated by the video generator to synthesize the target-embodiment realization of the interaction.

Structured Conditioning. The video generator is built on LingBot-Video-Dense 1.3B (Ma et al., 2026). The scene carrier $\mathcal{B}$ and body mask $\mathcal{M}$ anchor background and object evolution, while the command tokens $\boldsymbol{C}$ provide motion guidance. The temporal body field $\boldsymbol{\Phi}$ and surface field $\boldsymbol{Z}$ specify target-body structure, with the scene tokens $\boldsymbol{S}$ and patch memory $\boldsymbol{P}$ supplying dense source context and target appearance through residual conditioning paths. The text encoder remains empty so that generation is driven by the interaction and embodiment conditions throughout the denoising process.

Conditional Diffusion Generation. Given the structured conditions above, the video diffusion generator $\mathcal{F}^Y$ synthesizes the target sequence using scene, motion, structure, and appearance conditioning:

$$\mathcal{Y} = \mathcal{F}^Y\big(\mathcal{B}, \mathcal{M}, \boldsymbol{C}, \boldsymbol{S}, \boldsymbol{\Phi}, \boldsymbol{Z}, \boldsymbol{P}\big). \tag{7}$$

The scene carrier and mask enter the generative stream directly, while the remaining representations provide motion, spatial, source-context, and target-appearance conditioning throughout generation.

3.4 Implementation Details

Training. To construct aligned training pairs, we retarget source actions to robot assets and composite their renderings into source scenes (Appendix C). The resulting target videos are paired with source videos, actions, and action-independent target references. Rendered geometry provides spatial supervision, while offline tracklets and source/reference masks supervise source localization, motion, scene completion, and reference parsing. Full-video supervision is restricted to pairs passing source-object preservation screening; other samples contribute only valid auxiliary labels:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{spatial}} + \mathcal{L}_{\text{source}} + \mathcal{L}_{\text{reference}}. \tag{8}$$

Here, $\mathcal{L}_{\text{spatial}}$ supervises body occupancy, anchors, visible surfaces, and composite depth, while $\mathcal{L}_{\text{source}}$ and $\mathcal{L}_{\text{reference}}$ supervise source-side reconstruction and reference parsing. Training progresses

from source perception and action-conditioned placement to full temporal generation. Source actions remain available throughout training, and the generator receives predicted layouts, with teacher body geometry used only for supervision. See additional loss and optimization details in Appendix B.

Inference. EGO2ANY requires only the source video, action trajectory, camera intrinsics, and action-independent RGB references of the target robot. All target layouts and surface conditions are predicted from these inputs for video generation. No target meshes, kinematic models, future target masks, rendered future-body guidance, or auxiliary annotations are required.

4 EGOXPAIR DATASET

Paired cross-embodiment data provide aligned supervision for learning transferable interaction patterns. Repeating interactions across physical robots under matched scenes and viewpoints limits scalable collection. To address this bottleneck, EGOXPAIR uses the trained EGO2ANY model to expand source demonstrations into one-to-many aligned interaction groups for policy learning.

Dataset Construction. We pair each source demonstration with multiple target embodiments and synthesize their videos through offline EGO2ANY inference. Each generation uses the source video, action trajectory, camera intrinsics, and action-independent RGB references of the target robot, without target meshes, kinematic models, or rendered future-body guidance. Generated videos undergo automatic screening for action consistency, temporal stability, embodiment integrity, semantic alignment, and visual quality. Failed generation attempts are discarded and may be retried for the same source–target route. Retained videos are paired with end-effector actions, gripper states, language instructions, target-embodiment identifiers, and source provenance to support policy training. We assign splits by source identity before expansion, keeping all variants of a demonstration in the same split to prevent source-level leakage across the training, validation, and test sets.

Dataset Scale and Statistics. After quality screening, EGOXPAIR contains $\mathbf{1.0}\boldsymbol{M}$ retained videos derived from $\mathbf{100}\boldsymbol{K}$ unique source trajectories across $\mathbf{10}$ target robot embodiments. Each source–target route contributes one accepted video; rejected attempts and retries do not increase the retained count. All variants of a source share its actions, scene, viewpoint, and timeline.

5 EXPERIMENTS

5.1 IMPLEMENTATION DETAILS

We build EGO2ANY on LingBot-Video-Dense 1.3B (Ma et al., 2026), using 32-frame clips at 480×832 resolution and 9 reference images per target robot. Conditioning features have dimension 256, with 32 morphology slots. Training proceeds through source-perception pretraining, static morphology learning, and joint dynamic transfer. We use AdamW with learning rates of 10^{-4} for conditioning and perception modules and 10^{-5} for rank-16 backbone LoRA, with weight decay 10^{-2}. The VAE is frozen. Appendix B provides further implementation details.

5.2 CROSS-EMBODIMENT VIDEO GENERATION

Evaluation setup. We evaluate whether a robot specified by reference images can reproduce a source interaction in human-to-robot (H2R) and robot-to-robot (R2R) transfer. The H2R test set comprises 100 human demonstrations from EgoDex (Hoque et al., 2026). The R2R test set comprises 90 robot demonstrations, with 30 each from BridgeData V2, DROID, and RH20T (Walke et al., 2023; Khazatsky et al., 2024; Fang et al., 2024). Both test sets use the same four target embodiments: Flexiv Rizon 4, Franka Panda, UR5e, and Unitree G1. Together, these targets cover robotic arms equipped with grippers and humanoid robots with dexterous hands across both transfer directions.

Evaluation metrics. Following H2R-Bench Rong et al. (2026), we report target embodiment, scene preservation, task action, goal transition, and temporal continuity on a 0–100 scale. Embodiment scoring requires the requested robot to perform the task, while the other domains characterize preservation of the interaction. H2RCore and R2RCore summarize the two transfer directions using a weighted average. Appendix A details the rubric, domain definitions, and evaluation protocol.

Baselines. We compare against Native LingBot Video 1.3B (Ma et al., 2026), Mitty 14B (Song et al., 2025), VACE 1.3B (Jiang et al., 2025), and MiniMax H3 33B (MiniMax, 2026) using their

Table 1: **Quantitative results on human-to-robot transfer (§5.2).** Size denotes backbone parameters; ± denotes standard deviation across random seeds.

Method	Size	Semantic domains ↑					Overall ↑
		Embodiment	Scene	Action	Goal	Temporal	H2RCore
Mitty	14B	$2.9_{\pm 1.2}$	$71.5_{\pm 3.7}$	$49.9_{\pm 1.7}$	$50.0_{\pm 2.8}$	$27.1_{\pm 2.6}$	$13.1_{\pm 1.0}$
VACE	1.3B	$3.8_{\pm 0.9}$	$\mathbf{92.9}_{\pm 3.8}$	$\mathbf{95.0}_{\pm 4.3}$	$\mathbf{86.8}_{\pm 4.3}$	$\mathbf{81.2}_{\pm 2.8}$	$21.3_{\pm 0.8}$
MiniMax H3	33B	$16.7_{\pm 1.1}$	$50.1_{\pm 2.1}$	$42.0_{\pm 4.0}$	$27.8_{\pm 1.8}$	$50.7_{\pm 2.6}$	$21.8_{\pm 0.9}$
Native LingBot	1.3B	$1.4_{\pm 0.2}$	$78.1_{\pm 4.3}$	$35.5_{\pm 4.2}$	$3.5_{\pm 2.8}$	$45.8_{\pm 1.9}$	$9.7_{\pm 0.5}$
LingBot-Control	1.3B	$15.2_{\pm 2.9}$	$64.3_{\pm 1.7}$	$38.7_{\pm 2.9}$	$31.3_{\pm 3.0}$	$52.4_{\pm 3.0}$	$21.4_{\pm 2.3}$
EGO2ANY (ours)	1.3B	$\mathbf{43.1}_{\pm 2.1}$	$71.8_{\pm 4.4}$	$46.7_{\pm 2.8}$	$44.4_{\pm 4.5}$	$58.9_{\pm 2.3}$	$\mathbf{45.4}_{\pm 1.7}$

Table 2: **Quantitative results on robot-to-robot transfer (§5.2).** Size denotes backbone parameters; ± denotes standard deviation across random seeds.

Method	Size	Semantic domains ↑					Overall ↑
		Embodiment	Scene	Action	Goal	Temporal	R2RCore
Mitty	14B	$0.0_{\pm 0.0}$	$46.6_{\pm 2.3}$	$38.2_{\pm 2.0}$	$33.3_{\pm 3.9}$	$25.0_{\pm 3.2}$	$7.7_{\pm 0.3}$
VACE	1.3B	$1.9_{\pm 0.5}$	$86.3_{\pm 3.7}$	$\mathbf{96.4}_{\pm 3.0}$	$\mathbf{94.2}_{\pm 3.9}$	$80.1_{\pm 4.3}$	$19.8_{\pm 0.5}$
MiniMax H3	33B	$2.3_{\pm 1.3}$	$\mathbf{96.3}_{\pm 1.9}$	$81.0_{\pm 1.5}$	$83.3_{\pm 2.7}$	$\mathbf{95.8}_{\pm 2.6}$	$19.3_{\pm 1.1}$
Native LingBot	1.3B	$0.0_{\pm 0.0}$	$65.1_{\pm 3.2}$	$17.4_{\pm 3.1}$	$7.9_{\pm 3.5}$	$19.4_{\pm 3.3}$	$6.0_{\pm 0.3}$
LingBot-Control	1.3B	$9.4_{\pm 3.4}$	$80.7_{\pm 4.2}$	$45.1_{\pm 3.0}$	$84.3_{\pm 4.0}$	$32.8_{\pm 3.9}$	$20.0_{\pm 2.7}$
EGO2ANY (ours)	1.3B	$\mathbf{50.0}_{\pm 3.4}$	$72.6_{\pm 3.7}$	$52.9_{\pm 4.2}$	$71.8_{\pm 2.6}$	$47.7_{\pm 1.6}$	$\mathbf{52.4}_{\pm 2.8}$

native generation interfaces. LingBot-Control provides a closer comparison: it adapts the LingBot backbone with ControlNet-style conditioning (Zhang et al., 2023) on the source video and target references, without action-trajectory input, using the same training setup as EGO2ANY.

Human-to-robot transfer. EGO2ANY achieves the highest H2RCore (**45.4**) and embodiment score (**43.1**), surpassing MiniMax H3 by **23.6** and **26.4** points, respectively (Table 1). VACE leads in the other four domains but has a low embodiment score of 3.8, showing that interaction preservation alone does not ensure transfer to the requested robot. Compared with LingBot-Control, EGO2ANY improves embodiment (15.2 to **43.1**), scene preservation (64.3 to **71.8**), action (38.7 to **46.7**), goal transition (31.3 to **44.4**), and temporal continuity (52.4 to **58.9**). These gains indicate better target-robot realization while preserving scene context, task execution, and temporal coherence.

Robot-to-robot transfer. EGO2ANY also achieves the highest R2RCore, reaching **52.4** versus 20.0 for LingBot-Control, a **32.4**-point gain (Table 2). Its embodiment score of **50.0** exceeds the strongest baseline by **40.6** points. VACE 1.3B leads in action and goal transition (96.4 and 94.2), and MiniMax H3 leads in scene preservation and temporal continuity (96.3 and 95.8), but their embodiment scores remain low at 1.9 and 2.3, respectively, yielding R2RCore scores of 19.8 and 19.3. Compared with LingBot-Control, EGO2ANY improves action (45.1 to **52.9**) and temporal continuity (32.8 to **47.7**), while scoring lower in scene preservation (80.7 to **72.6**) and goal transition (84.3 to **71.8**). Across both directions, the results support stronger target-robot realization and higher embodiment-weighted overall scores, with remaining trade-offs in interaction preservation. Figure 3 provides qualitative comparisons. We further demonstrate morphology generalization with Unitree G1 qualitative comparisons in Appendix G.1.

5.3 DOWNSTREAM VLA POLICY EVALUATION

Training data and policy backbones. We evaluate StarVLA-π (StarVLA Community, 2026; Black et al., 2025) and StarVLA-GR00T (NVIDIA et al., 2025), both using the Cosmos-Predict2-2B visual backbone (NVIDIA, 2025). For each policy variant, we compare benchmark-only training, augmentation with matched EGOXPAIR, and augmentation with the full EGOXPAIR corpus. Matched EGOXPAIR is generated from the benchmark's own training demonstrations: the selected BridgeData V2 (Walke et al., 2023) split for SIMPLER and the training split of each LIBERO suite.

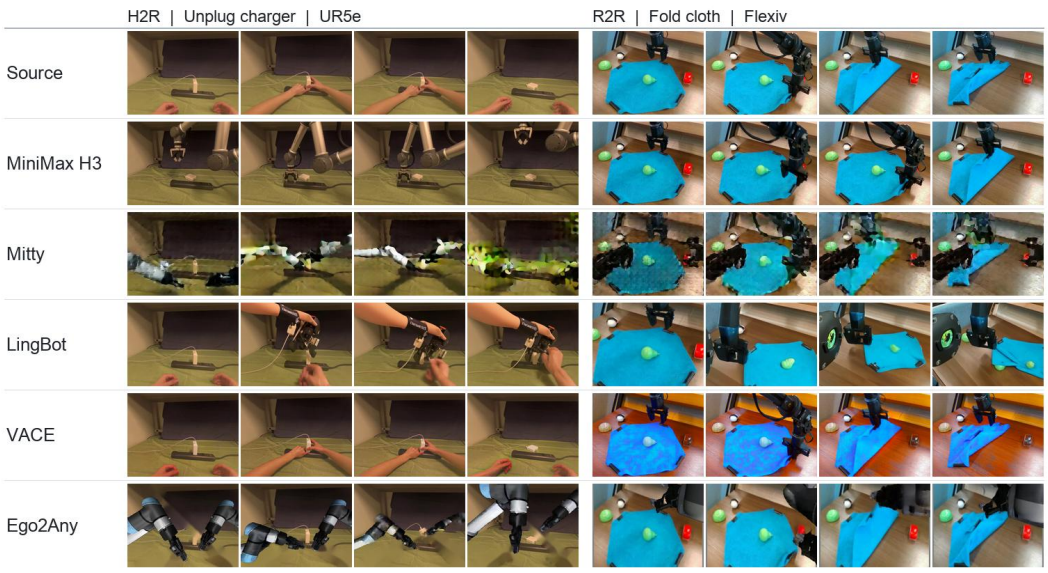

Figure 3: **Qualitative H2R and R2R comparisons.** Selected frames from a human-to-robot charger-unplugging example with UR5e (left) and a robot-to-robot cloth-folding example with Flexiv (right).

Table 3: **SIMPLER-WidowX visual matching success rates (% ↑) (§5.3).** ± denotes standard deviation across random seeds.

Method	Size	Spoon–Towel	Carrot–Plate	Block Stack	Eggplant–Basket	Avg.
Representative methods						
π_0	3B	29.1	0.0	16.6	62.5	27.1
SpatialVLA	4B	16.7	25.0	29.2	**100.0**	42.7
Our 2B baselines						
StarVLA-π	2B	$81.4_{\pm 2.4}$	$55.2_{\pm 3.1}$	$25.1_{\pm 2.8}$	$73.0_{\pm 2.6}$	$58.7_{\pm 1.4}$
+ *matched* EGOXPAIR	2B	$85.2_{\pm 2.3}$	$60.4_{\pm 3.2}$	$33.7_{\pm 3.4}$	$78.7_{\pm 2.5}$	$64.5_{\pm 1.6}$
+ EGOXPAIR	2B	$\mathbf{90.1}_{\pm 2.1}$	$68.0_{\pm 3.3}$	$43.7_{\pm 3.5}$	$86.6_{\pm 2.2}$	$72.1_{\pm 1.7}$
StarVLA-GR00T	2B	$80.4_{\pm 2.7}$	$65.4_{\pm 2.9}$	$20.0_{\pm 2.5}$	$80.6_{\pm 2.3}$	$61.6_{\pm 1.3}$
+ *matched* EGOXPAIR	2B	$84.3_{\pm 2.5}$	$68.1_{\pm 2.9}$	$32.5_{\pm 3.3}$	$84.7_{\pm 2.2}$	$67.4_{\pm 1.5}$
+ EGOXPAIR	2B	$88.7_{\pm 2.3}$	$\mathbf{72.2}_{\pm 3.0}$	$\mathbf{46.9}_{\pm 3.6}$	$89.5_{\pm 2.0}$	$\mathbf{74.3}_{\pm 1.8}$

Evaluation benchmarks. We evaluate closed-loop task success on SIMPLER-WidowX (Li et al., 2025c) and LIBERO (Liu et al., 2023). SIMPLER adapts on BridgeData V2 and evaluates four visual-matching tasks: Spoon–Towel, Carrot–Plate, Block Stack, and Eggplant–Basket. LIBERO uses its training demonstrations and evaluates Spatial, Object, Goal, and Long. We also report results for π_0 (Black et al., 2025), SpatialVLA (Qu et al., 2025), and OpenVLA (Kim et al., 2025b) policies.

Comparisons. On SIMPLER (Table 3), EGOXPAIR augmentation increases the reported average success rate from 58.7% to **72.1%** for StarVLA-π and from 61.6% to **74.3%** for StarVLA-GR00T, corresponding to gains of **13.4** and **12.7** percentage points. Improvements occur on all four tasks for both backbones. Block Stack shows the largest gains, increasing from 25.1% to **43.7%** and from 20.0% to **46.9%**, respectively. The improvement across both policy variants suggests that the utility of the generated demonstrations extends beyond a single policy formulation.

On LIBERO (Table 4), both backbones improve by **1.4** percentage points on average, reaching **96.9%** and **96.6%**. Gains are positive across all four suites and are largest on LIBERO-Goal, at **3.1** and **2.5** percentage points. The smaller absolute gains coincide with higher baseline success rates on this benchmark. Together, these results support the utility of cross-embodiment video generation as a source of supervision for downstream manipulation policies.

Table 4: **LIBERO success rates (% ↑) (§5.3).** ± denotes standard deviation across random seeds.

Method	Size	Spatial	Object	Goal	Long	Avg.
Representative methods						
OpenVLA	7B	84.7	88.4	79.2	53.7	76.5
π_0	3B	96.8	98.8	95.8	85.2	94.1
Our 2B baselines						
StarVLA-π	2B	$98.9_{\pm 0.4}$	$98.3_{\pm 0.5}$	$94.4_{\pm 0.7}$	$90.4_{\pm 0.9}$	$95.5_{\pm 0.3}$
+ *matched* EGOXPAIR	2B	$99.1_{\pm 0.4}$	$98.5_{\pm 0.5}$	$95.8_{\pm 0.7}$	$91.0_{\pm 0.8}$	$96.1_{\pm 0.3}$
+ EGOXPAIR	2B	$\mathbf{99.4}_{\pm 0.3}$	$\mathbf{98.9}_{\pm 0.4}$	$97.5_{\pm 0.7}$	$\mathbf{91.8}_{\pm 0.7}$	$\mathbf{96.9}_{\pm 0.4}$
StarVLA-GR00T	2B	$97.4_{\pm 0.6}$	$98.0_{\pm 0.5}$	$95.1_{\pm 0.8}$	$90.4_{\pm 0.7}$	$95.2_{\pm 0.2}$
+ *matched* EGOXPAIR	2B	$98.0_{\pm 0.5}$	$98.3_{\pm 0.5}$	$96.2_{\pm 0.7}$	$90.7_{\pm 0.7}$	$95.8_{\pm 0.3}$
+ EGOXPAIR	2B	$98.7_{\pm 0.4}$	$98.8_{\pm 0.5}$	$\mathbf{97.6}_{\pm 0.6}$	$91.4_{\pm 0.6}$	$96.6_{\pm 0.3}$

Table 5: **Component ablations (§5.4).** ± denotes standard deviation across random seeds.

Configuration	Semantic domains ↑					Overall ↑
	Embodiment	Scene	Action	Goal	Temporal	H2RCore
LingBot-Control	$15.2_{\pm 2.9}$	$64.3_{\pm 1.7}$	$38.7_{\pm 2.9}$	$31.3_{\pm 3.0}$	$52.4_{\pm 3.0}$	$21.4_{\pm 2.3}$
+ Action conditioning	$17.3_{\pm 2.7}$	$64.8_{\pm 2.0}$	$42.1_{\pm 3.1}$	$36.8_{\pm 3.4}$	$53.0_{\pm 2.8}$	$23.6_{\pm 2.2}$
+ Source-body removal	$22.6_{\pm 2.5}$	$69.5_{\pm 3.6}$	$43.0_{\pm 2.9}$	$38.4_{\pm 3.8}$	$53.8_{\pm 2.7}$	$28.3_{\pm 2.0}$
+ Frame-wise layout	$34.2_{\pm 2.4}$	$70.2_{\pm 3.9}$	$44.1_{\pm 3.0}$	$40.3_{\pm 4.1}$	$54.6_{\pm 2.6}$	$37.8_{\pm 2.0}$
+ Temporal transport	$39.4_{\pm 2.3}$	$70.6_{\pm 4.2}$	$45.5_{\pm 2.7}$	$42.5_{\pm 4.3}$	$57.3_{\pm 2.5}$	$42.2_{\pm 1.9}$
EGO2ANY (full)	$\mathbf{43.1}_{\pm 2.1}$	$\mathbf{71.8}_{\pm 4.4}$	$\mathbf{46.7}_{\pm 2.8}$	$\mathbf{44.4}_{\pm 4.5}$	$\mathbf{58.9}_{\pm 2.3}$	$\mathbf{45.4}_{\pm 1.7}$

5.4 ABLATIONS OF CROSS-EMBODIMENT GENERATION

Evaluation setup. Table 5 progressively adds action conditioning, source-body removal, frame-wise layout, temporal transport, and surface support to LingBot-Control. The layout stage jointly introduces interaction and morphology encoders, action-anchored layout prediction, and geometric conditioning and supervision; temporal transport enables cross-frame attention. The full model further adds reference-conditioned surface support with part-correspondence supervision. We evaluate all configurations using the five semantic domain scores and H2RCore.

Ablation analysis. Action conditioning improves action and goal scores by 3.4 and 5.5 points, but embodiment by only 2.1, suggesting that motion commands alone provide limited guidance for target-body realization. Source-body removal raises embodiment and scene scores by 5.3 and 4.7 points, consistent with reduced source-body interference and better scene preservation. Frame-wise layout yields the largest H2RCore gain (28.3 to 37.8), with an 11.6-point embodiment improvement but only a 0.8-point temporal gain. Temporal transport further improves embodiment by 5.2 points and temporal continuity by 2.7, supporting the complementary contributions of spatial layout prediction and cross-frame modeling. Surface support adds 3.7 embodiment and 3.2 H2RCore points, bringing the full model to 43.1 and 45.4, respectively. Together, these components consistently improve all five semantic domains and raise H2RCore by 24.0 points over LingBot-Control.

6 CONCLUSION

We presented EGO2ANY, a framework for transferring egocentric manipulation interactions across robot embodiments through video generation. By separating source interaction and scene context from target morphology, EGO2ANY uses morphology-conditioned temporal transport to guide a shared video generator toward target-specific realizations of the demonstrated interaction. We further introduced EGOXPAIR, which pairs generated cross-embodiment videos with action annotations for downstream policy learning. Results on human-to-robot and robot-to-robot generation, together with policy evaluations on SIMPLER and LIBERO, suggest that this formulation can support both interaction transfer and the reuse of demonstrations for robot learning.

AI USE STATEMENT

In this work, generative AI models were used to synthesize cross-embodiment videos for the EGOX-PAIR dataset and to complete masked background regions during training pseudo-label construction. These uses are described in § 4 and Appendix C and D. We also used multimodal models to score generated videos under the evaluation protocol described in Appendix A and A.4. Additionally, generative AI tools assisted with improving the grammar, phrasing, clarity, and readability of the manuscript. The authors take responsibility for the final manuscript, methodology, data, evaluation procedures, and reported conclusions.

REFERENCES

Sami Azirar, Enrico Pallotta, Jan Nogga, Jürgen Gall, Sven Behnke, and Hermann Blum. ContactFlow: A Video Action Conditioning That Transfers Across Embodiments. *arXiv preprint arXiv:2607.26579*, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *RSS*, 2025.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *RSS*, 2023.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to Act Anywhere with Task-centric Latent Actions. In *RSS*, 2025.

Cheng Chen, Jerry Bai, Jiacheng Wei, Boyu Chen, Xiaoji Zheng, Fan Wu, Minghao Yang, Tianrun Chen, Ruibo Li, Xiaoyu Yue, Xiaoyang Guo, Yixiao Ge, Guosheng Lin, and Fayao Liu. AnyWorld: Factorized Egocentric World Models for Cross-Embodiment Generalization. *arXiv preprint arXiv:2608.29242*, 2026.

Hai Ci, Xiaokang Liu, Pei Yang, Yiren Song, and Mike Zheng Shou. H2R-Grounder: A Paired-Data-Free Paradigm for Translating Human Interaction Videos into Physically Grounded Robot Videos. *arXiv preprint arXiv:2512.09406*, 2025.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control. *arXiv preprint arXiv:1812.00568*, 2018.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot. In *ICRA*, 2024.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised Learning for Physical Interaction through Video Prediction. In *NeurIPS*, 2016.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. In *RSS*, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*, 2022.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A Controllable Generative World Model for Robot Manipulation. In *ICLR*, 2026a.

Yaowei Guo, Zeng Tao, Yuxin Jiang, Yunuo Chen, Zhiyang Dou, Yuxiang Ma, Yin Yang, Demetri Terzopoulos, Ying Jiang, and Chenfanfu Jiang. RoboEdit: Turning Human Manipulation Videos into Scalable Robot Experience. *arXiv preprint arXiv:2608.18948*, 2026b.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. In *ICLR*, 2026.

Jinho Jeong, Se June Joo, Jaehyun Kang, Dongyun Kim, Yena Kim, Hanjung Kim, and Seon Joo Kim. HuRo: Robotizing Human Videos for Scalable VLA Pretraining. *arXiv preprint arXiv:2609.10706*, 2026.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *ICCV*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *RSS*, 2024.

Byungjun Kim, Taeksoo Kim, Hyunsoo Cha, and Hanbyul Joo. Robot-Factored World Models via Robot Rendering. *arXiv preprint arXiv:2607.22535*, 2026.

Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. UniSkill: Imitating Human Videos via Cross-Embodiment Skill Representations. In *CoRL*, 2025a.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel,

Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *CoRL*, 2025b.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training Robots Without Robots Using Only Human Videos. In *CoRL*, 2025.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from In-the-wild Human Videos using Data-Editing. In *ICRA*, 2026.

Guangrun Li, Yaoxu Lyu, Zhuoyang Liu, Chengkai Hou, Jieyu Zhang, and Shanghang Zhang. H2R: A Human-to-Robot Data Augmentation for Robot Pre-training from Videos. *arXiv preprint arXiv:2505.11920*, 2025a.

Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, Xinze Chen, Yun Ye, Guan Huang, Zhenbo Song, and Xingang Wang. MimicDreamer: Aligning Human and Robot Demonstrations for Scalable VLA Training. *arXiv preprint arXiv:2509.22199*, 2025b.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating Real-World Robot Manipulation Policies in Simulation. In *CoRL*, 2025c.

Zhiyuan Li, Wenyan Yang, Wenshuai Zhao, Yue Ma, Yuanpeng Tu, Pekka Marttinen, and Joni Pajarinen. Bridging the Embodiment Gap: Disentangled Cross-Embodiment Video Editing. *arXiv preprint arXiv:2605.03637*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *NeurIPS*, 2023.

Kechen Liu and Ola Shorinwa. CLAP: Cross-Embodiment Video World Models Are Zero-Shot Physical Simulators. *arXiv preprint arXiv:2608.27406*, 2026.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, Yanhong Zeng, Qin Zhao, Zifan Shi, Wei Wu, Hao Ouyang, Qiuyu Wang, Shangzhan Zhang, Jiahao Shao, Yipengjing Sun, Liangxiao Hu, Lunke Pan, Nan Xue, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Ka Leong Cheng. Scaling Mixture-of-Experts Video Pretraining for Embodied Intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

MiniMax. MiniMax H3. GitHub repository, `https://github.com/MiniMax-AI/MiniMax-H3`, 2026.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A Universal Visual Representation for Robot Manipulation. In *CoRL*, volume 205, 2023.

NVIDIA. Cosmos-Predict2: World Simulation Model for Physical AI. GitHub repository, 2025.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734*, 2025.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *ICRA*, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. In *RSS*, 2025.

Dingyi Rong, Yue Shi, Chaofan Ma, Jiezhang Cao, Zongrui Wang, Zeyu Zhang, Yao Mu, Guangtao Zhai, and Ning Liu. H2R-Bench: Benchmarking Human-to-Robot Manipulation Video Generation in World Models. *arXiv preprint arXiv:2608.13049*, 2026.

Yiren Song, Cheng Liu, Weijia Mao, and Mike Zheng Shou. Mitty: Diffusion-based Human-to-Robot Video Generation. *arXiv preprint arXiv:2512.17253*, 2025.

Yiren Song, Xiyao Deng, Pei Yang, Yihan Wang, and Mike Zheng Shou. OmniHumanoid: Streaming Cross-Embodiment Video Generation with Paired-Free Adaptation. *arXiv preprint arXiv:2605.12038*, 2026.

StarVLA Community. StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing. *arXiv preprint arXiv:2604.05014*, 2026.

Mutian Tong, Han Jiang, Qiao Feng, Lingjie Liu, and Jiatao Gu. PointAction: 3D Points as Universal Action Representations for Robot Control. *arXiv preprint arXiv:2606.03943*, 2026.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A Dataset for Robot Learning at Scale. In *CoRL*, 2023.

Ye Wang, Pei Lin, Xiong-Hui Chen, Haoqi Yuan, Zhixuan Liang, Yiyang Huang, Anzhe Chen, Zixing Lei, Jie Zhang, Tao Zhang, Haoyang Li, Tong Zhang, Chenxi Xiao, Ziyuan Jiao, and Qin Jin. Ego2Robot: Scalable Robot Data Synthesis from Egocentric Human Data. *arXiv preprint arXiv:2608.02580*, 2026.

Zhuoyuan Wu and Jun Gao. OSCAR: Omni-Embodiment Action-Conditioned World Model for Robotics. *arXiv preprint arXiv:2606.04463*, 2026.

Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Haoran Chen, Shiwei Shen, Jiaqi Leng, Zuxuan Wu, and Yu-Gang Jiang. Human2Robot: Learning Robot Actions from Paired Human-Robot Videos. In *AAAI*, 2026.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. XSkill: Cross Embodiment Skill Discovery. In *CoRL*, 2023.

Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. Kinema4D: Kinematic 4D World Modeling for Spatiotemporal Embodied Simulation. In *CVPRW*, 2026.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent Action Pretraining from Videos. In *ICLR*, 2025.

Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. XIRL: Cross-embodiment Inverse Reinforcement Learning. In *CoRL*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023.

Yuxiang Zhou, Yusuf Aytar, and Konstantinos Bousmalis. Manipulator-Independent Representations for Visual Imitation. In *RSS*, 2021.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A Fine-Grained World Model for Robot Manipulation. In *ICCV*, 2025.

APPENDIX CONTENTS

A Cross-Embodiment Generation Evaluation **15**

A.1 Evaluation scope and review procedure 15

A.2 Semantic domains and embodiment constraints 15

A.3 Overall transfer scores 15

A.4 Multi-Model Evaluation 16

B Additional Implementation Details **16**

C Deterministic Rendering for Cross-Embodiment Pseudo-Labels **16**

C.1 Source Normalization and Embodiment Expansion 16

C.2 Retargeting and Native Rendering 17

C.3 Scene Carrier and Depth-Aware Composition 17

C.4 Automated Execution, Reproducibility, and Label Scope 18

D EGOXPAIR Construction Protocol **19**

E Limitations **19**

F Broader Impact **20**

G Additional Qualitative Results **21**

G.1 Morphology Generalization to Dexterous Hands 23

A Cross-Embodiment Generation Evaluation

A.1 Evaluation scope and review procedure

The H2R test set contains 100 EgoDex human-source demonstrations, and the R2R test set contains 90 robot-source demonstrations: 30 each from BridgeData V2, DROID, and RH20T. Target embodiments are Flexiv Rizon 4, Franka Panda, UR5e, and Unitree G1. We compare complete systems through their native conditioning interfaces; differences therefore do not isolate individual components.

We use independent multi-model scoring under a shared rubric. Each model judge receives the source video, task contract, and target-robot references, inspects each complete output at its native duration, and locates relevant events in both sequences. Contact is examined at native resolution. Each criterion receives an integer score from 0 (absent or contradicted) to 4 (fully demonstrated), supported by object-specific evidence and source/output time intervals. Clip duration and smooth camera motion alone do not establish manipulation success. Appendix A.4 specifies the model panel and score aggregation.

A.2 Semantic domains and embodiment constraints

We divide the 15 component scores by four and combine them into five domains in $[0, 1]$:

- **Target embodiment (E):** body structure, exact robot identity, effector fidelity, and the target robot's acting role, with weights $(0.25, 0.35, 0.20, 0.20)$.
- **Scene preservation (S):** initial object state, spatial layout, object persistence, and workspace visibility, with weights $(0.30, 0.35, 0.20, 0.15)$.
- **Task action (A):** manipulated object, event order, plausible contact, causal object response, and functional support, with weights $(0.20, 0.20, 0.25, 0.25, 0.10)$.
- **Goal transition (G):** evidence of the required object-state change during the video.
- **Temporal continuity (T):** persistent identities, attachment, and coherent event progression.

The weighted embodiment score is capped by the identity, effector, and acting-role components. It is set to zero if the target robot is absent or another primary actor performs the task. Scene and action scores can consequently be high even when embodiment transfer fails. A desired state that is already present in the opening frame does not demonstrate a goal transition. Functional support may come from an effector, table, fixture, or other feasible strategy, without matching the human hand count.

A.3 Overall transfer scores

We report H2RCore and R2RCore as the overall scores for human-to-robot and robot-to-robot transfer, respectively. For each direction, each model judge's component scores are normalized to $[0, 1]$ and combined into the five domains, including the embodiment constraints above. We take the arithmetic mean of the three judges' domain scores for each source–output pair, average across target robots for each source, and then average across sources with equal weight. Let $\bar{E}_d, \bar{S}_d, \bar{A}_d, \bar{G}_d, \bar{T}_d$ denote these source-balanced domain means for direction $d \in \{\mathrm{H2R}, \mathrm{R2R}\}$. We define

$$\mathrm{Core}_d = 100\left(0.80\bar{E}_d + 0.06\bar{S}_d + 0.08\bar{A}_d + 0.04\bar{G}_d + 0.02\bar{T}_d\right). \tag{9}$$

Here $\mathrm{Core}_{\mathrm{H2R}}$ is H2RCore and $\mathrm{Core}_{\mathrm{R2R}}$ is R2RCore. The weights sum to one and are shared across methods and transfer directions. Embodiment receives 80% of the total weight to prioritize actual cross-embodiment transfer. A model that fails to generate the requested robot may retain the source actor and reproduce its motion, obtaining high scene-preservation, action, and temporal scores despite failing to change the embodiment. An embodiment-dominant weighting reduces the reward for this source-reproduction shortcut and places greater emphasis on whether the requested target robot performs the interaction. With this weighting, an output scoring zero on embodiment can receive at most 20 overall, even with perfect scores in the other domains. The remaining 20% captures scene preservation, task execution, goal transition, and temporal continuity. All domain and overall scores are presented on a 0–100 scale, rounded to one decimal place.

This additive score allows strengths in one domain to compensate for weaknesses in another. It measures overall semantic quality rather than the fraction of successful transfers, so the embodiment

column remains necessary for assessing whether the requested robot performs the interaction. The embodiment constraints described above remain part of the domain rubric; no minimum or additional gate is applied when combining the five domains into H2RCore or R2RCore.

A.4 Multi-Model Evaluation

Gemini 3.5 Flash, Qwen3.7-Plus, and GPT-6 Astra independently score the same outputs using the shared 15-criterion rubric. Our score definitions and weights differ from H2R-Bench (Rong et al., 2026), so the aggregate scores are not directly comparable.

Scores used in the main tables. The generation scores in Tables 1 and 2 use the equally weighted arithmetic mean of the three model judges' scores. For each source–output pair, we first compute each judge's domain scores using the component weights and embodiment constraints above, then average the three judges' domain scores. These means are aggregated across target robots and source demonstrations before computing H2RCore or R2RCore.

B Additional Implementation Details

Architecture and inputs. We use LingBot-Video-Dense 1.3B (Ma et al., 2026) with 32-frame 480×832 clips, a 30×52 spatial field, $D = 256$, and $L = 32$ morphology slots. Each target has nine 256×256 RGB references spanning three views and three configurations, with independent reference dropout of 0.2. To assess sensitivity to the reference inputs, we evaluate embodiments seen during training using a different set of reference images. Actions encode camera-frame TCP positions, 6D rotations, and grasp openness in two effector slots, synchronized to RGB; absent effectors are masked and windows with unresolved poses or calibration are excluded. Intrinsics follow image resizing/cropping, and endpoint heatmaps use a standard deviation of 1.5 grid cells. Layout and source adapters follow generator blocks 3, 9, 15, and 21, and reference adapters follow blocks 3, 11, and 19 (zero-based), with zero-initialized output projections.

Supervision and optimization. Source episodes are split before window extraction. Training progresses through source perception/completion, initial placement with the source encoder frozen, and joint temporal generation. Full-video supervision requires source-object preservation screening; other samples contribute only valid auxiliary labels. Depth supervision requires compatible camera-frame depths, normalized by clip median and median absolute deviation. Surface labels use a temporally shared Hungarian part assignment. In Eq. (8), flow matching has weight 1, occupancy and coordinate-motion losses have weight 0.05, and the remaining spatial, source, and reference losses have weight 0.1. AdamW uses learning rates of 10^{-4} for conditioning/perception and 10^{-5} for rank-16 backbone LoRA, weight decay 10^{-2}, and gradient clipping at norm 1. The VAE and copied patch projection remain frozen. Generation uses predicted scene and layout features; teacher geometry and masks provide supervision only.

C Deterministic Rendering for Cross-Embodiment Pseudo-Labels

We implement an offline pipeline that converts source demonstrations into aligned robot renderings and composite RGB pseudo-labels. It supplies a reproducible geometric teacher for cross-embodiment video learning: source motion determines the target end-effector path, native robot assets determine the rendered body, and source-body removal provides the scene carrier. Camera calibration, robot geometry, and inverse kinematics (IK) are used during preparation. Inference uses source-camera intrinsics to project supplied TCP commands, but requires neither robot geometry nor IK. This rendered teacher corpus is distinct from the EgoXPair expansion in Section 4.

C.1 Source Normalization and Embodiment Expansion

Unified source representation. Dataset-specific adapters normalize EgoDex, DROID, BridgeV2, and RH20T demonstrations into a shared representation containing RGB frames, original frame indices, camera intrinsics, image dimensions, end-effector positions and rotations in the camera frame, normalized gripper openness, validity flags, and source identity. Positions use meters and

rotations use 3×3 matrices in $\mathrm{SO}(3)$. We check for finite values, proper rotations, positive focal lengths, compatible array shapes, strictly increasing frame indices, valid action frames, and openness in $[0, 1]$. Projection uses the supplied focal lengths and principal point; samples with skewed intrinsics are rejected. Frame indices bind all outputs to the sampled source timeline rather than implying a common control frequency across datasets.

For EgoDex, we retain both hands in each sequence, with positions of shape $T \times 2 \times 3$ and explicit left–right ordering; samples containing only a selected hand are excluded. Robot demonstrations retain their source identifiers and corresponding camera calibration. Samples with unresolved calibration or ambiguous source–camera associations are excluded before rendering.

Target library and sample identity. The target library contains ten robot arms: Franka Panda, UR5e, Flexiv Rizon4, Kinova Gen3, KUKA iiwa14, Franka FR3, xArm7, Sawyer, UR10e, and Lite6. Panda uses its native gripper, Rizon4 uses AG95, and the remaining assemblies use Robotiq 2F85 assets. These are synthetic assemblies of native rendering assets, not claims of identical real-world mounting arrangements. A *route* is one source sequence paired with one target assembly; a bimanual route instantiates two arms in a shared scene. Each target is processed independently, so an infeasible route does not discard feasible alternatives for the same source.

Source membership and train/validation/test assignments are frozen before embodiment expansion. All target variants and temporal windows retain their source grouping. We extract 32-frame training windows from source sequences, increasing the number of training examples without adding independent demonstrations.

C.2 Retargeting and Native Rendering

Fixed placement and sequential IK. For each target, we initialize joints from an embodiment-specific configuration and search a bounded, ordered list of base orientations. We test yaw angles $0°, -45°, 45°, -90°, 90°$ and retain the first feasible sequence. Each candidate fixes the base transform for the entire clip. For a single arm, its translation aligns the initial tool-center point (TCP) with the first source endpoint, and its orientation combines the candidate yaw with the initial source and robot tool orientations. Writing this camera-from-base transform as $(\mathbf{R}_b, \mathbf{t}_b)$, the IK targets are

$$\mathbf{p}_t^b = \mathbf{R}_b^\top (\mathbf{p}_t^c - \mathbf{t}_b), \qquad \mathbf{R}_t^b = \mathbf{R}_b^\top \mathbf{R}_t^c. \tag{10}$$

The superscripts c and b denote camera and robot-base coordinates. Each frame is solved from the preceding joint solution, using joint bounds and damped numerical IK. For bimanual inputs, both arms are solved in one collision scene with fixed bases; an episode-level alignment maps each hand's source rotations to its tool frame while preserving their temporal changes.

Geometric acceptance. A candidate must satisfy the position and orientation tolerances in Table 6 throughout the clip. Collision checks cover the sampled configurations and five interior joint-interpolation points per segment, including cross-arm collisions for bimanual sequences. Native assembly exclusions are retained, and permitted gripper closure contacts are handled explicitly. These sampled checks do not establish continuous-time collision freedom or collision avoidance against unmodeled scene objects. Foreground occupancy is also checked in every frame to reject invisible or excessively large renderings. If all candidates fail, we reject the route and record the reason.

Aligned render products. MuJoCo's native EGL renderer produces robot RGB, a segmentation-derived foreground mask, and metric robot depth from the same joint states and camera. The camera supports distinct horizontal and vertical focal lengths and an off-center principal point. For each rendered sequence, we also store the joint sequence, base transforms, requested grasp widths, selected endpoint poses, residuals, and source frame indices, keeping RGB and spatial labels aligned.

Requested gripper width is computed from source openness and the assembly's maximum aperture. Exported joint and gripper fields serve as diagnostics rather than verified action supervision.

C.3 Scene Carrier and Depth-Aware Composition

Removing the source body. Source processing is shared across target embodiments. Existing masks and backgrounds are reused after validation; otherwise, SAM3 segments the human arms/hands

Table 6: Parameters of the rendering pipeline. These thresholds determine sample acceptance but do not guarantee visual or physical correctness.

Component	Setting
IK position / rotation tolerance	0.025 m / 15°
IK iterations / damping	160 / 10^{-3}
IK rotation weight / update limit	0.04 m / 0.12 rad
Interior collision samples per segment	5
Per-frame robot foreground fraction	$[0.002,\ 0.65]$
Source-mask dilation radius	4 pixels
Removal canvas / inference steps / seed	480×832 / 12 / 42
DA3 processing resolution	504
Scene–robot depth comparison margin	0.03 m
TCP depth lower-bound tolerance / pass rate	0.10 m / 0.90
Minimum visible robot fraction for depth-aware composition	0.35

or robot arm/gripper, and MiniMax completes the dilated removal region. Let $\mathbf{E}$ be the binary edit mask and $\widehat{\mathbf{B}}$ the completion output. The stored carrier is

$$\mathbf{B} = (1 - \mathbf{E}) \odot \mathbf{X} + \mathbf{E} \odot \widehat{\mathbf{B}}. \tag{11}$$

An exact pixel check verifies that the carrier equals the source outside $\mathbf{E}$. This constraint preserves observed pixels outside the edit region, but does not establish object preservation where the object overlaps that region. Completed pixels remain pseudo-labels and may contain reconstruction artifacts.

Metric-depth checks and occlusion. The depth-enabled configuration applies the Depth Anything 3 (DA3) Nested metric-depth model to the completed carrier. Predictions must be finite, positive, and marked metric by the provider; the fifth percentile of its confidence values must be at least 1.0. As a geometric sanity check, the median scene depth in a 5×5 neighborhood of each in-frame projected source TCP must be at least its observed camera depth minus 0.10 m, for at least 90% of these observations. This check can reject inconsistencies but does not validate dense depth.

For valid robot and scene depths d_r and d_s, scene pixels occlude the rendered robot when $d_s + \delta < d_r$, with $\delta = 0.03$ m. With $\mathbf{A}$ denoting the robot mask and $\mathbf{O}$ these scene-front pixels, we compose

$$\mathbf{V} = \mathbf{A} \wedge \neg \mathbf{O}, \qquad \mathbf{Y}^\star = \mathbf{V} \odot \mathbf{C}_r + (1 - \mathbf{V}) \odot \mathbf{B}, \tag{12}$$

where $\mathbf{C}_r$ is robot RGB and $\mathbf{V}$ is its visible mask. We retain both the full and visible masks. Depth-aware composition requires at least 35% of the rendered robot pixels, aggregated over the clip, to remain visible. If the depth or visibility check fails, we retain the previously accepted RGB composite with $\mathbf{V} = \mathbf{A}$ and record why depth-based occlusion was omitted. Estimated carrier depth is not treated as ground-truth scene depth.

C.4 Automated Execution, Reproducibility, and Label Scope

Resumable stage execution. The pipeline performs geometry processing, source segmentation, background completion, optional depth estimation, composition, and sample indexing. Geometry is checked first to avoid scene completion for sources with no feasible target. Source-level outputs are computed once and shared across target embodiments, while target geometry and composites are stored separately. Each stage runs with its required models and software environment. Samples are assigned to GPUs using a fixed partition, with limits on retries and execution time.

For reproducibility, we record the source selection, configuration, code, assets, and model versions, together with hashes of each stage's inputs and outputs. Completed results are reused only after their recorded dependencies and outputs have been verified. Geometric rejections are retained, while execution failures may be retried. Writes are protected against interruption and concurrent access; changes to dependencies trigger a new run configuration or a stage-specific compatibility check. Here, *deterministic* refers to fixed input selection, ordered geometric search, seeded completion, and verified reuse within the recorded environment, rather than bitwise-identical neural inference across different hardware or software.

Data selection and supervision scope. Each sample records geometric validity, RGB availability, dataset split, and eligible supervision. Structurally valid training composites may be retained for RGB augmentation despite removal artifacts; this does not establish object preservation, coherent interaction, or temporal consistency. Full-video supervision additionally requires source-object preservation screening (Section 3.4). Rendered outputs are not treated as verified joint-action, contact-force, tactile, object-pose, or task-success labels, and successful rendering does not establish executable manipulation.

D EGOXPAIR CONSTRUCTION PROTOCOL

Appendix C documents the implemented rendering teacher and its RGB pseudo-label scope. The following protocol details concern the learned-generator expansion into EGOXPAIR.

Route identity and action alignment. EGOXPAIR indexes source–target routes before generation. A fixed set of 100K source demonstrations and ten target embodiments defines one million source–target routes. Candidate generations are screened at the attempt level, with seeds and retry indices recorded. One accepted video per route yields the final 1.0M retained videos; rejected attempts, retries, and temporal windows are not counted as additional videos. All routes derived from a source share its train/validation/test assignment.

Eligible routes contain source effector poses, grasp states, synchronized timestamps, camera intrinsics, and explicit tool/frame conversions as specified in Appendix B. Commands are stored separately from their rendered or observed realization. Unexecuted gripper-width requests and visual-only alignment offsets are excluded from action labels. Renderer outputs provide auxiliary geometry or RGB augmentation according to the recorded supervision scope; action-paired generation uses source actions.

Video–action consistency. A frozen visual evaluator extracts target TCP image tracks and visible gripper opening/closing events from generated RGB, independently of the generator's predicted anchors. Projected source actions provide comparison tracks. The criteria include normalized endpoint error, adjacent-frame displacement error, and gripper-event timing over jointly measurable frames. Target-body detection precedes motion comparison so that retained source hands cannot satisfy agreement with the source action. A 2D endpoint match measures visual motion consistency rather than certifying 3D contact or physical execution.

Screening and provenance. Screening additionally checks source-object preservation, target-body integrity, temporal artifacts, and semantic agreement with the demonstration. Evaluator thresholds are fixed using a separate development split with static, reversed-motion, wrong-embodiment, retained-source, and missing-object controls. Occluded or unmeasurable events receive an indeterminate outcome rather than automatic acceptance. Infrastructure failures are retryable, whereas geometric or semantic failures retain their rejection reason. Records include the source identifier, target assembly, reference capture, target actions, instruction, generation configuration, screening scores, and acceptance decision. Candidate routes, attempts, and accepted videos are counted separately.

E LIMITATIONS

Input and supervision requirements. EGO2ANY requires synchronized source actions, camera intrinsics, and informative multi-view target references. It therefore does not directly support arbitrary uncalibrated videos without action annotations. Errors in these inputs, source-body removal, or scene completion can propagate to the generated interaction. Although inference requires no target 3D assets, training uses a geometric teacher built with robot assets and inverse kinematics. Its coverage and pseudo-label artifacts constrain the supervision available to the model.

Physical feasibility. The model predicts visual realizations of supplied actions rather than executable target joint trajectories. Preserving source endpoint motion and object evolution does not establish that a different robot can realize the same contact, grasp, or force. Neither the learned layouts nor the video–action screening certify collision-free motion, dynamic feasibility, or manipulation

success. Generated demonstrations consequently require target-specific validation before use in physical control.

F BROADER IMPACT

Potential benefits. Reusing demonstrations across embodiments could reduce the need for repeated data collection on each robotic platform. By providing a shared reference-conditioned generation interface, EGO2ANY may broaden access to cross-embodiment training data and support research on visual generalization. These benefits depend on the quality of generated demonstrations and their validation for the intended tasks.

G Additional Qualitative Results

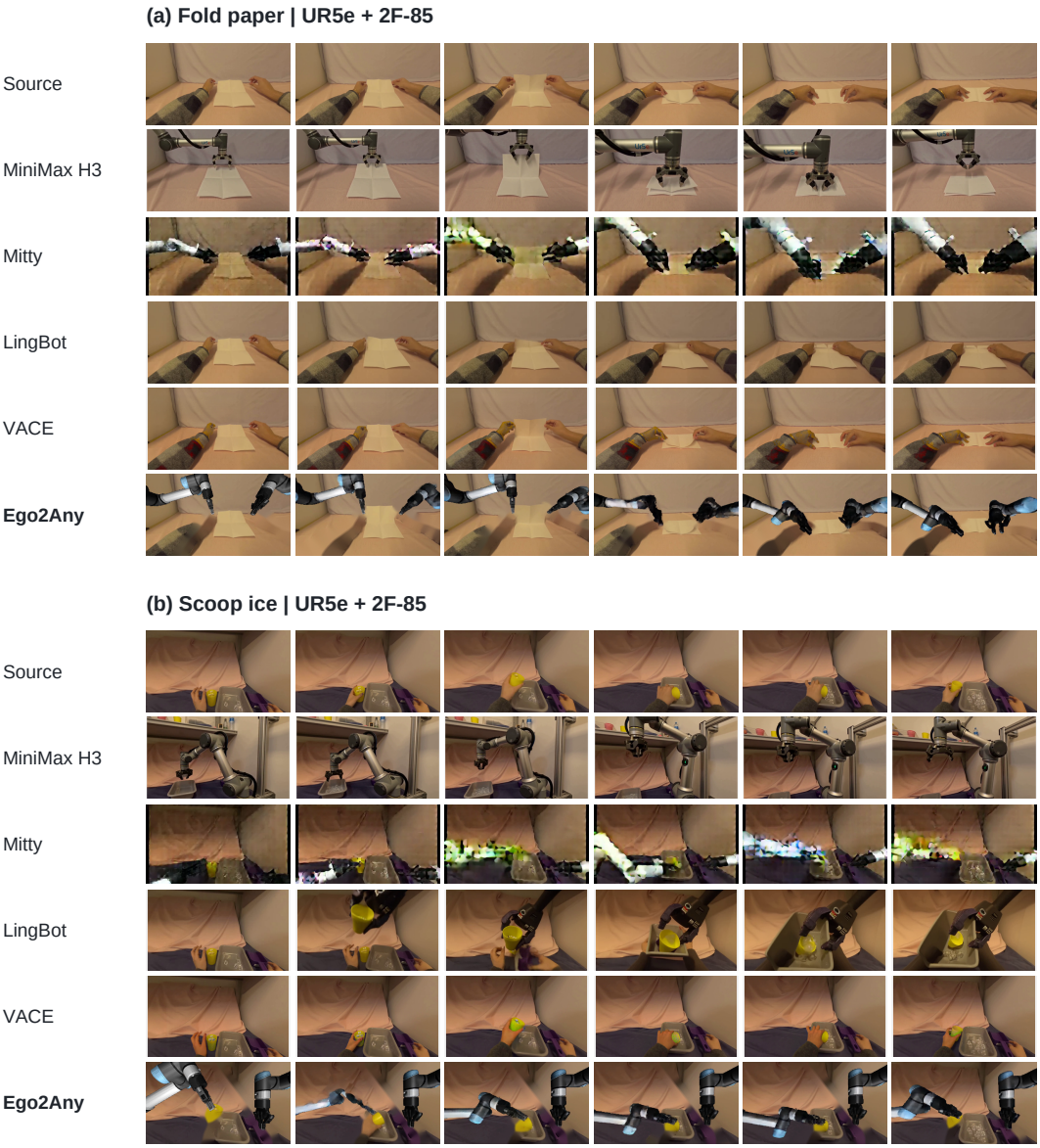

Figure 4: **Additional H2R comparisons.** Human demonstrations of (a) folding paper and (b) scooping ice, with UR5e + 2F-85 as the target embodiment in both cases.

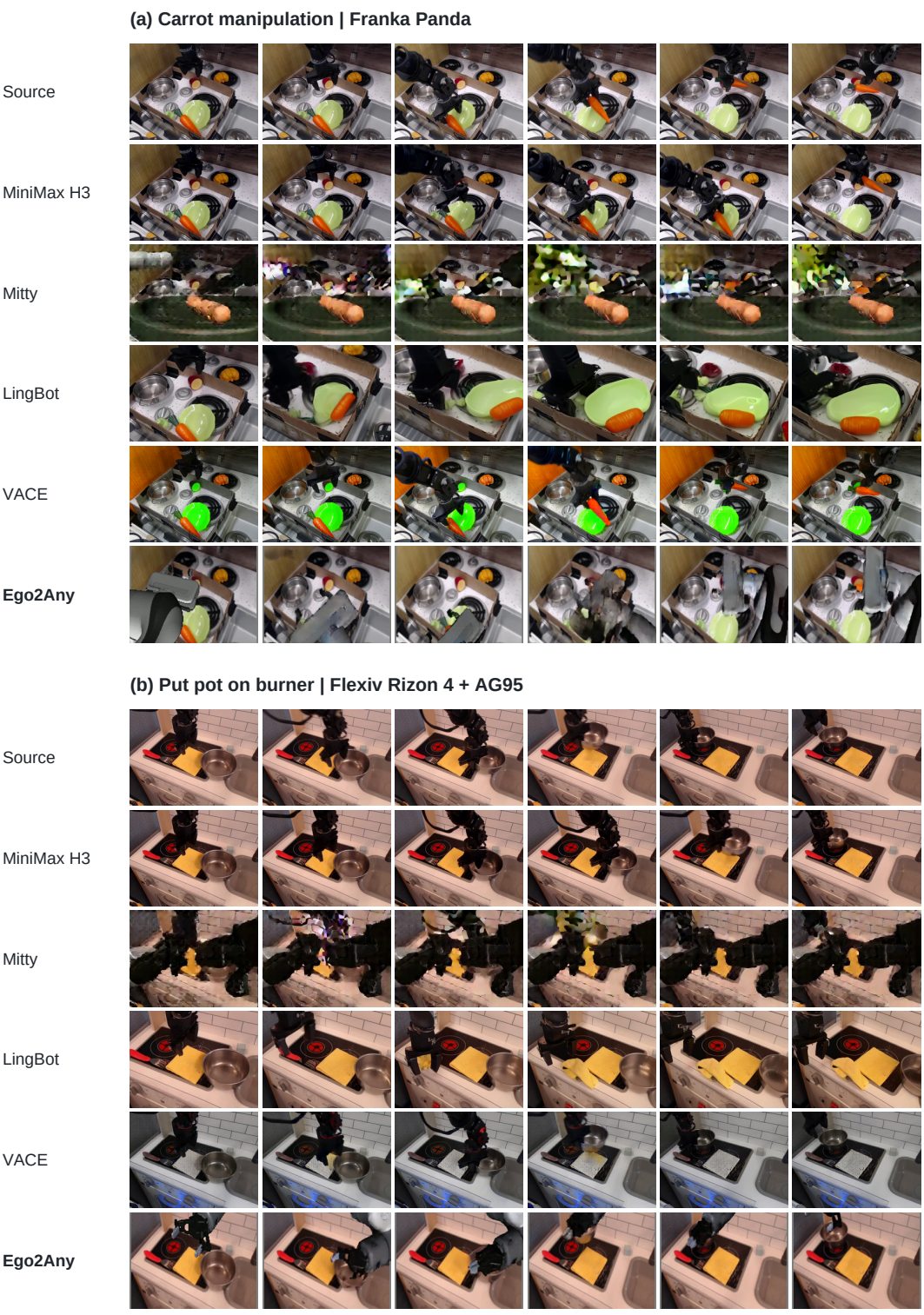

Figure 5: **Additional R2R comparisons.** Bridge examples of (a) carrot manipulation with a Franka Panda target and (b) placing a pot on a burner with a Flexiv Rizon 4 + AG95 target.

G.1 Morphology Generalization to Dexterous Hands

Figure 6 illustrates morphology generalization in H2R transfer to Unitree G1 with three-finger dexterous hands on Jenga manipulation and case-zipping tasks. Together with the arm-and-gripper results, these examples demonstrate support for both robotic arms and dexterous hands.

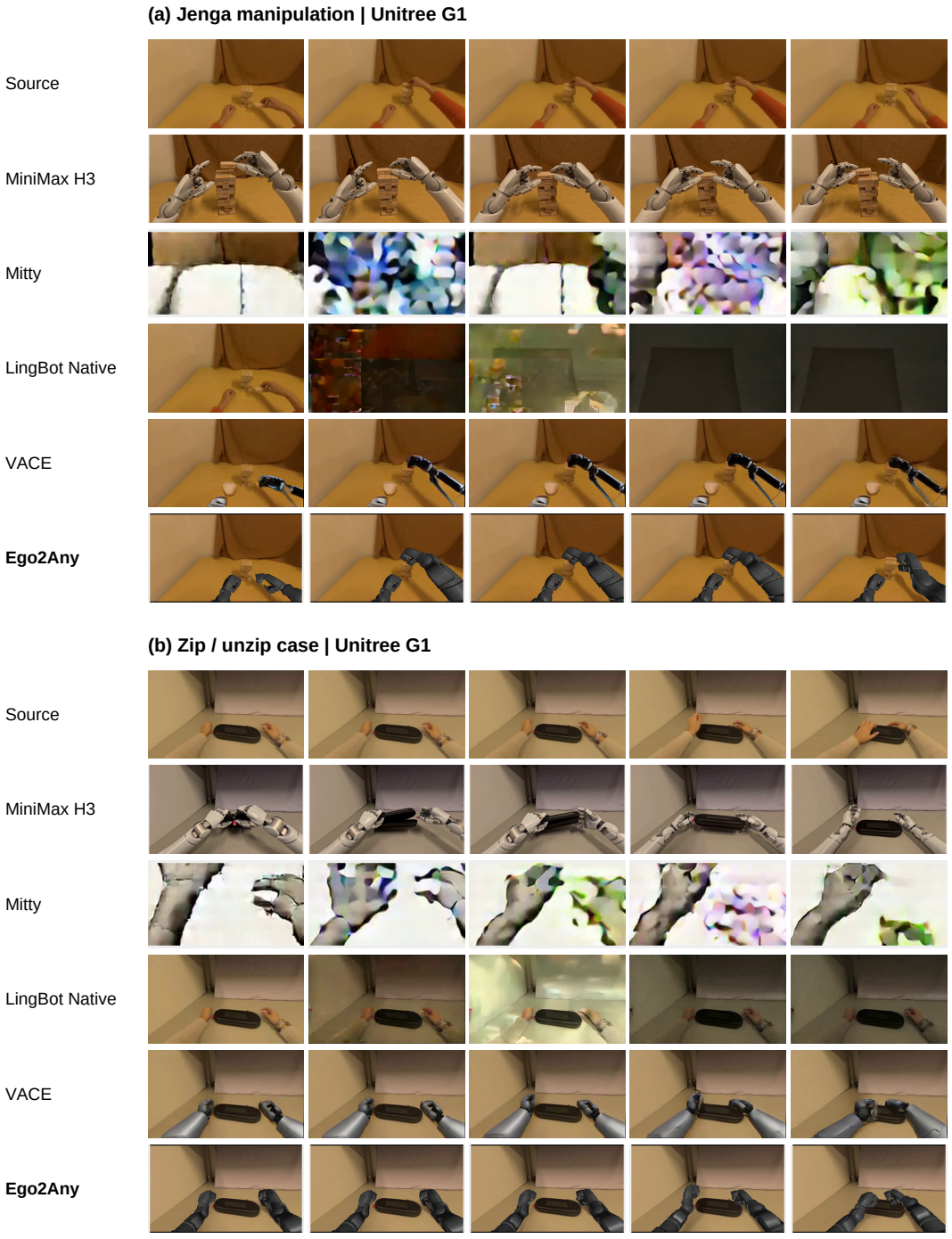

Figure 6: **Morphology generalization to G1 dexterous hands.** H2R comparisons on (a) Jenga manipulation and (b) zipping/unzipping a case.