SCIENTIFIC OASIS

Decision Making Advances

Journal homepage: www.dma-journal.org
ISSN: 2956-2384

Beyond Teleoperation: Enhancing VLA Robustness via Explicit Kinematic Retargeting of Human Demonstrations

José António Nunes Andrade[1], Mauro Castelli[1,*]

[1] NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Campus de Campolide, 1070-312, Lisboa, Portugal

ARTICLE INFO

Article history:
Received 17 May 2026
Received in revised form 18 June 2026
Accepted 28 June 2026
Available online 7 August 2026

Keywords:
Robot Learning; Vision-Language-Action Models; Motion Retargeting; Inverse Kinematics; Terminal State Ambiguity

ABSTRACT

Vision-Language-Action (VLA) models have demonstrated remarkable capabilities in generalized robotic control, yet their scalability is fundamentally bottlenecked by the high cost and low diversity of teleoperated data. While abundant, human demonstration videos cannot be directly utilized for policy training due to the severe morphological differences between human anatomy and robotic manipulators. To bridge this embodiment gap, this work proposes a lightweight retargeting pipeline that kinematically retargets human interaction data (DexYCB) onto six-degree-of-freedom manipulator trajectories to fine-tune policies based on the pi0.5 architecture. By prioritizing Cartesian positional alignment via constrained Inverse Kinematics (IK) and introducing an object-based grasping heuristic, smooth geometric priors are generated without relying on computationally heavy visual synthesis. Physical evaluations demonstrate that retargeted models significantly outperform standard teleoperation (40.6% success rate), achieving 65.6% success via co-training and a peak 78.1% success rate via two-stage cross-embodiment co-training. Furthermore, evaluations under extreme visual clutter reveal that explicitly retargeted policies exhibit immunity to semantic visual distractors. Finally, we examined and analysed Terminal State Ambiguity, a temporal failure mode where generative models fail to terminate the task when exposed to scenarios similar to the nature of human video priors.

1. Introduction

Vision-Language-Action (VLA) models represent a fundamental evolution in robotic control. Building upon the semantic reasoning capabilities of Vision-Language-Models (VLMs), VLAs integrate action modalities to unify perception, semantics, and motion generation within end-to-end frameworks [1,2]. However, the scalability of these generalist systems is fundamentally bottlenecked by the scarcity of physical robot data. Traditional teleoperation is expensive, time-consuming, and

* *Corresponding author.*
E-mail address: mcastelli@novaims.unl.pt

https://doi.org/10.31181/dma412026180

limited in diversity, inherently constraining a policy's ability to generalize across novel tasks and hardware [3–5].

Human video demonstrations offer a compelling solution by providing abundant, diverse data from in-the-wild physical interactions. Recent approaches like EgoVLA and Masquerade have demonstrated the viability of bridging the embodiment gap by using Inverse Kinematics (IK) to retarget human motion into synthetic robot actions [6–8]. However, these frameworks exhibit critical limitations in hardware morphological generalization. Existing explicit IK methods primarily target highly dexterous, anthropomorphic end-effectors, leaving a significant gap in how to compress human dexterity into divergent robotics morphologies. Conversely, state-of-the-art generalist VLAs rely on multi-robot embodiment training datasets, trusting that human-to-robot transfer will naturally "emerge" at scale without explicit geometric grounding. While this emergent transfer is demonstrably effective, it relies on scaling to a magnitude that demands immense computational resources - fundamentally underutilizing the quality of dense geometric priors available in human data [1,2].

The central research problem addressed in this paper is the embodiment gap between human video demonstrations and robot action execution, which directly enforces a critical data scarcity bottleneck in imitation learning. While human manipulation videos contain rich spatial and semantic information, they do not directly provide robot-compatible joint trajectories. This creates a practical bottleneck for VLA learning: teleoperation data are expensive and limited, while human demonstrations are abundant but morphologically incompatible with low Degrees of Freedom (DoF) manipulators. The overarching question is whether explicit geometric retargeting can convert human interaction priors into robot-action supervision in a way that improves policy robustness without requiring large-scale teleoperated robot data or computationally expensive visual synthesis.

To systematically evaluate the viability of the framework presented, this study is driven by the following three core research questions (RQs):

- (RQ1) Does the inclusion of retargeted human data improve the policy's execution on the original teleoperated task?
- (RQ2) Does training on diverse human videos enhance the VLA's visual robustness when exposed to localized clutter?
- (RQ3) How do explicitly human retargeted spatial priors perform under conditions of extreme visual noise?

The main contributions of this work are fourfold. First, this work proposes a lightweight explicit kinematic retargeting pipeline that maps human grasping demonstrations from DexYCB into the joint-action space of a 5 DoF SO-101 robotic arm with a 1-DoF parallel gripper [9]. Second, unlike recent approaches that rely on visual synthesis or robotized video generation, the proposed method uses only geometric trajectory conversion, constrained IK, temporal Exponential Moving Average (EMA) smoothing filter, and object-motion-based grasp inference. Third, this work empirically shows that adding a small amount of retargeted human data improves downstream VLA performance over teleoperation-only training, increasing perfect task success from 40.6% to 65.6% in the co-trained setting and to 78.1% in the two-stage cross-embodiment setting. Fourth, the study identifies and analyzes Terminal State Ambiguity (TSA), a failure mode in which generative policies complete the manipulation objective but fail to terminate safely due to a mismatch between human demonstrations and robot teleoperation trajectories.

2. Related Work

The pursuit of generalist VLA policies has driven significant efforts in large-scale dataset aggregation, yielding foundational robotic corpora such as BridgeData [10] and RT-X [11]. However, because these datasets rely predominantly on expensive, time-consuming teleoperation, the literature has increasingly turned toward in-the-wild human video demonstrations to achieve the scale necessary for robust generalization [2,11,12].

Within this paradigm shift, existing research diverges fundamentally on how to bridge the morphological gap between human anatomy and robotic manipulators. A substantial portion of the literature utilizes human videos purely for visual and linguistic grounding, relying on the assumption that massive-scale training allows models to implicitly align human and robotic embodiments [1,2,4,6]. Conversely, to explicitly bridge this gap, a growing body of work seeks to translate human demonstrations into robot trajectories. This explicit approach traces a distinct historical trajectory, from early joint geometric correspondences [13,14], to learning-based methods that introduced more flexible mappings between human and robot motion representations [12,15], to modern generative pipelines capable of synthesizing VLA training data at scale [6,7,16].

The following sections review the core pillars of this progression: the architectural evolution of VLA models, the mechanisms of motion retargeting, and the state-of-the-art frameworks leveraging these techniques for scalable policy training.

2.1 Vision Language Action models

Recent advancements in VLA models integrate visual perception, natural language understanding, and action generation into unified frameworks. Surveys such as Kawaharazuka *et al.* [3], Sapkota *et al.* [4], and Feng *et al.* [8] highlighted progress in their real-world deployment, emphasizing multimodal fusion, architectural innovations for efficiency, and applications in humanoid robotics and autonomous navigation. Xiao *et al.* [5] provided a comprehensive overview of robot learning in the foundation model era, noting that while large-scale pretraining enhances robustness, it simultaneously underscored persistent challenges in data scarcity for embodied tasks.

Prominent examples of state-of-the-art foundation models include GR00T-N1, which was pretrained on a "data pyramid" consisting of internet video at the base, synthetic data in the middle, and real-world robot data at the top, demonstrating emergent generalization in bimanual manipulation [1]. Similarly, Black *et al.* [2] presented pi0.5, a VLA with open-world capabilities also trained on Web data, semantic prediction, and multi-robot trajectories for broad real-world manipulation. Importantly, pi0.5 utilizes a continuous flow-matching architecture for action generation, making it highly adaptable to executing fluid, complex movements and heavily reliant on dense demonstration data to achieve precise kinematic alignment.

2.2 Motion Retargeting in Robotics

In the context of leveraging human videos for robot learning, the two most common approaches to motion retargeting are geometric methods and learning-based methods. The former rely on explicit kinematic mappings and analytical solutions like IK to adapt poses while preserving structural relationships. Geometric methods provide an interpretable and resource-efficient solution but can be limited in handling complex nonlinear differences. By contrast, learning-based methods, which use deep learning architectures to infer flexible mappings from training data, offer an alternative to accommodate complex nonlinear differences in embodiments [17].

Geometric retargeting emphasizes rule-based or optimization-driven transformations that map human joint configurations, or end-effector poses, to robot kinematics, making it efficient for real-time applications and scenarios with similar morphologies. For instance, Darvish *et al.* [13] propose whole-body geometric retargeting for humanoid robots, establishing joint correspondences via IK to maintain geometric consistency across differing DoF. Building on this, Huang *et al.* [18] introduce a one-shot whole-body motion learning framework that retargets human motion clips and directly maps 3D joint rotations to the robot's kinematic structure, using geodesic distances and interpolation to avoid collision poses, enabling scalable policy training from minimal human data as an alternative to teleoperation [17].

Conversely, learning-based retargeting leverages deep learning to infer flexible mappings. Yan *et al.* [19] presented ImitationNet, an unsupervised approach that learns a shared latent space between human and robot motions via contrastive learning, enabling direct retargeting without requiring manually paired demonstrations. Similarly, Park *et al.* [20] focused on transferring human hand skills by learning joint manifolds that map human motions, robot actions, and object dynamics to robotic dexterous execution. Complementing these two methods, Yin *et al.* [14] introduced a method blending geometric principles with learning-based techniques to convert human keypoints to robot keypoints at high speed (1 kHz) while incorporating strict joint constraints.

2.3 Leveraging Human Video Demonstrations for Scalable Policy Training

Despite the advances in foundational VLAs, acquiring sufficiently diverse robot data for dexterous manipulation remains a critical bottleneck requiring precise embodiment alignment [3,4]. To address this, some works sought to leverage abundant human demonstration videos to generate synthetic robot-compatible data with the goal of informing pretraining and fine-tuning [21–23]. These methods collectively mark a shift toward scalable VLA adaptation by translating explicitly human motions into robot trajectories, serving as a viable alternative to costly teleoperation. These approaches can be categorized into foundational pretraining and targeted fine-tuning paradigms.

2.3.1 Foundational approaches

Foundational approaches prioritize end-to-end training pipelines, utilizing human videos as the base for imitation learning. Qiu *et al.* [12] presented HAT, a manipulation policy that models humans as humanoid embodiments in unified state-action spaces, co-training on human data to develop manipulation policies without relying on extensive robot-specific datasets. Lepert *et al.* [24] introduced Phantom, which extracts actions from human videos to train imitation policies, eliminating initial hardware needs. Their following work [6], Masquerade, extends this by editing human videos into "robotized" demonstrations by removing humans from the data and implanting robot overlays to pretrain a vision encoder and fine-tune a diffusion VLA. Similarly, EgoVLA, pretrains VLAs on human videos with hand retargeting via IK, followed by fine-tuning on limited robot data to achieve dexterous manipulation [7].

2.3.2 Fine-tuning approaches

Complementing foundational training, fine-tuning approaches leverage existing pretrained models, augmenting their capabilities with synthesized human data. Li *et al.* [23] presented MimicDreamer, a framework that aligns human and robot demonstrations through viewpoint stabilization, constrained IK for action retargeting, and diffusion vision synthesis to produce synthetic

robot videos and trajectories from human demonstrations. This approach fine-tunes the VLA backbone pi0 (iteration before pi0.5 discussed previously) [25] on blended synthetic robot data and a small quantity of real datasets, demonstrating a 14.7% absolute improvement over robot-only baselines in bimanual tasks.

Likewise, Luo *et al.* [26] introduced Being-H0, which advances dexterous VLA design by first extending a VLM (InternVL3) backbone into VLA capabilities through physical instruction tuning on an extensive human video dataset (UniHand). By refining the policy on sparse real-robot episodes, Being-H0 achieves success rates between 60% and 90% on pick-and-place tasks, exceeding reference models like GR00T-N1.5 by 15% and 30% [1].

Finally, while these methods demonstrate the efficacy of utilizing human data, they often rely on complex visual synthesis pipelines (e.g., MimicDreamer) or are targeted toward highly dexterous, high-DoF hardware (e.g., Being-H0), leaving a gap for lightweight, purely kinematic fine-tuning approaches applicable to non-anthropomorphic manipulators.

2.3.3 Comparative synthesis

While the aforementioned foundational and fine-tuning approaches demonstrate the efficacy of utilizing human demonstration data, a critical comparative analysis reveals a significant gap regarding computational complexity and morphological constraints. Existing frameworks typically fall into two extremes: they either focus on highly dexterous, anthropomorphic hardware that closely mimics human anatomy (e.g., EgoVLA, Being-H0), or they rely on resource-intensive generative visual synthesis and video-altering pipelines to force embodiment alignment (e.g., Masquerade, MimicDreamer) [6,7,23,26]. Table 1 synthesizes the proposed method against these state-of-the-art cross-embodiment methodologies, contrasting their characteristics such as computational overhead and target hardware scope.

The proposed method seeks to minimize computational overhead while maximizing the performance metric (increase of 37.5%) by strictly refining the generated data and training approach. Existing frameworks frequently attempt to solve the embodiment gap holistically, resulting in "high" overhead: Masquerade relies on generative video overlays to pre-train the Vision Encoder, while MimicDreamer synthesizes both visual and kinematic data to fine-tune the full VLA network [6,23]. Furthermore, while EgoVLA also focuses on kinematic data, it computes resource-intensive, multi-joint IK specifically to map human fingers to complex anthropomorphic hands [7]. In contrast, the proposed method completely bypasses visual synthesis and dexterous finger mapping. By utilizing the human wrist as a spatial proxy for the end-effector and inferring gripper actuation via an object-displacement heuristic, the method produces pure kinematic data that allows the training phase to focus exclusively on fine-tuning the policy's action head. This streamlined architecture achieves a 37.5% absolute improvement in perfect task execution over the baseline policy, demonstrating that explicitly retargeted kinematic priors can dramatically scale VLA performance on accessible, platform-agnostic manipulators without the immense rendering overhead demanded by state-of-the-art data augmentation.

3. Methodology

To address the data scarcity inherent in robotic learning, this work proposes a cross-embodiment transfer pipeline that retargets human grasping demonstrations into robotic control policies. The proposed methodology is predicated on the hypothesis that large-scale human interaction data co-

Table 1
Comparative analysis of cross-embodiment VLA frameworks

Method	Generated data	Training approach	Computational overhead	Performance metric	Target embodiment scope
EgoVLA [7]	Kinematic Data (Dexterous multi-joint finger IK)	Pre-trains Action Head	Moderate (Requires extensive egocentric pre-training)	+18% absolute success rate over no-pretrain baselines on unseen visual configurations	Anthropomorphic robotic hands
Masquerade [6]	Visual Data (Robot video overlays)	Pre-trains Vision Encoder	High (Requires generative multi-stage video editing and robot asset overlays)	Enhanced robustness to background visual shifts; high real-world imitation success	General rigid manipulators
MimicDreamer [23]	Visual + Kinematic Data (Diffusion + IK)	Fine-tunes Full VLA	High (Requires diffusion-based video generation with multi-view)	+14.7% absolute success rate improvement over robot-only baselines in bimanual tasks	Bimanual robotic platforms
Being-H0 [26]	Kinematic Data (3D hand trajectories)	Pre-trains Full VLA	High (Requires large-scale physical instruction tuning on extensive video corpora)	+15% to +30% success rate improvement over foundation models like GR00T-N1.5	High-DoF / Dexterous systems
Proposed method	Kinematic Data (IK based on human wrist as end effector)	LoRA Fine-tunes Action Head	Minimal (Pipeline only requires geometric calculation)	+37.5% absolute increase in perfect execution (40.6% to 78.1%) under a low-data teleoperation baseline	Platform-agnostic (Any manipulator via URDF integration)

ntains latent action primitives, specifically approaching vectors and grasp affordances, that can be kinematically retargeted to robotic manipulators without relying on massive compute for emergent alignment. This section details the pipeline for converting the DexYCB dataset into robot trajectories, the specific policy architectures employed, and the training protocol designed to bridge the visual and kinematic gaps.

3.1 System and Tools Overview

All experiments are conducted on an SO-101 robotic manipulator operating with a five-DoF arm equipped with a one-DoF parallel-jaw gripper. The teleoperation configuration utilizes a leader-follower interaction, where a passive SO-101 arm captures spatial actions synchronized with visual

observations. It is important to note that while empirical validation in this study is conducted exclusively on the SO-101 manipulator, the proposed retargeting pipeline is architecturally designed to be platform-agnostic. The core software stack relies on generalized geometric optimization constraints that adapt dynamically to the target robot's Unified Robot Description Format (URDF) file. Consequently, transitioning the framework to alternate robotic configurations primarily requires substituting the target URDF (which natively updates the corresponding joint boundaries) and modifying the extrinsic base frame calibration (T_{calib}) and wrist-to-gripper transformation matrix (T_{offset}), entirely bypassing the need to retrain generative visual models for new embodiments.

The sensory apparatus consists of two light sources and a single exocentric RGB camera ($I_{external}$) capturing the global workspace context shown in Figure 1. Additionally, to minimize arbitrary visual noise and better align the visual distribution in the retargeted DexYCB dataset, the physical workspace was standardized using a dark table cover and dark background curtains.

Fig. 1. Workspace setup on the left, exocentric RGB image on the right

Some of the stages of the method presented, specifically dataset merging, policy training, and inference, are orchestrated using Hugging Face's LeRobot framework. To maintain a stable physical control frequency of 30 Hz despite the high computational latency inherent to large VLA models, we leverage LeRobot's asynchronous server-client functions. The client handles robot state observation and physical actuation, while inference is offloaded to a dedicated server equipped with a single NVIDIA RTX A6000 GPU (48 GB VRAM). Because network and inference latencies are non-negligible, we employed a Receding Horizon Control (RHC) strategy, commonly referred to in continuous generative models as action chunking. Rather than predicting single discrete actions, the pi0.5 policy generates a temporal batch of future continuous actions (T=50 steps) per inference call. The client executes a subset of this trajectory window while asynchronously requesting the subsequent batch before the current buffer hits the threshold (15%, seven steps), ensuring perfectly smooth physical control. Finally, the aforementioned GPU hardware footprint is critical not only for inference speed but for the training phase, where fine-tuning the massive pi0.5 architecture via Low-Rank Adaptation (LoRA) consumes approximately 36 GB of VRAM.

3.2 Retarget Pipeline

The DexYCB dataset provides 3D keypoint annotations of human hands (via the MANO model) interacting with 20 distinct objects across 1000 episodes [9]. This data is processed through the

retargeting pipeline to transform the human demonstrations into a format compatible with the SO-101 kinematic chain.

To map the human workspace to the robot workspace, the human wrist pose $T_H \in SE(3)$ is projected into a target end-effector pose T_{Target}. The target pose is therefore computed as per Eq. (1):

$$T_{Target} = T_{calib} * \mathrm{T_H} * \mathrm{T_{offset}} \tag{1}$$

where T_{calib} represents the static calibration of the robot base relative to the DexYCB world origin, and T_{offset} is a rigid body transformation aligning the human wrist frame with the functional center of the robot's parallel gripper.

Human workspace dimensions derived from raw demonstration data can vary wildly across subjects, introducing severe scale discrepancies if directly retargeted. To ensure the continuous feasibility of robotic policy learning, the Cartesian coordinates of T_{Target} were normalized prior to solving for joint angles. This spatial trajectory is then linearly scaled down so that its maximum Cartesian range maps strictly to the maximum robot reach (0.20 meters) of the SO-101, ensuring that every waypoint resides safely within the manipulator's physical reach.

The corresponding robot joint $q \in \mathbb{R}^5$ is derived from solving the IK optimization problem. The gripper joint is explicitly excluded from this IK pass. The error is minimized between the robot's Forward Kinematics $FK(q)$ and the target pose according to the objective function, expressed in Eq. (2):

$$q^* = argmin_q \left(\left|\left| pos\big(FK(q)\big) - pos\big(T_{target}\big) \right|\right|_2^2 + \lambda \left|\left| rot\big(FK(q)\big) \ominus rot\big(T_{target}\big) \right|\right|^2 \right) \tag{2}$$

where λ acts as a regularization term balancing translational, $pos(\cdot)$ and rotational precision, $rot(\cdot)$. During pipeline development, the synthetic IK trajectories were benchmarked against real teleoperation data by measuring the maximum angular delta per frame (Δq_{max}) to ensure kinematic safety. Initial trials used nonzero λ values, which resulted in high joint jitter, as the solver oscillated erratically to satisfy orientation constraints. Consequently, with the goal of prioritizing positional accuracy, rotational precision was sacrificed by strictly setting λ to 0. This eliminates the rotational penalty, explicitly forcing the IK solver to guarantee Cartesian positional accuracy at the expense of geometric orientation.

Additionally, to resolve the instability and enforce temporal coherence, a sequential seeding strategy is implemented alongside dynamic bounding constraints. The IK solver for frame $t + 1$ is always initialized using the solved joint angles from frame t, and the solver's search space is restricted to a maximum movement of 15^0 per frame, operating strictly within the hardware joint limits defined by the official Hugging Face SO-101 URDF file. Finally, to filter out any residual tracking noise inherited from the MANO estimations, an EMA smoothing filter is applied ($\alpha = 0.35$) to the final 5 DoF trajectory.

Concurrently, the continuous aperture of the human hand is abstracted into a binary gripper state $g \in \{0, 1\}$. Rather than relying on the noisy distance between human fingers, a purely object-centered kinematic heuristic was implemented. The gripper closure is triggered strictly by detecting the initial physical displacement of the target object with respect to a threshold. Let $p_{obj}^{(t)}$ represent the smoothed 3D spatial coordinates of the target object at time step t, and $p_{obj}^{(0)}$ be its initial resting position. A grasp event is defined as occurring at the first step t_{grasp} where the object's Cartesian

displacement exceeds a defined tracking noise threshold r. This condition is formally expressed in Eq. (3):

$$t_{grasp} = \min\left\{t \middle| \left\| p_{obj}^{(t)} - p_{obj}^{(0)} \right\|_2 > r\right\} \tag{3}$$

To determine the threshold r, an empirical analysis was conducted for each episode. This involved graphical analysis of object trajectories across diverse episodes to assess a fundamental engineering trade-off: minimizing the threshold to provide immediate grasp triggering upon object motion, while preventing premature false triggers caused by visual tracking jitter (MANO model noise). Validation was performed using an empirical subset of 88 episodes, comprising a dense sampling of 50 episodes from the black marker condition and a stratified sampling of two episodes from each of the remaining object categories. Based on this analysis, the tracking noise threshold was set to r=0.045 meters.

To ensure trajectory completeness and policy stability in edge cases where the object is occluded or remains strictly stationary (e.g., failed human grasps in the dataset), a failsafe is implemented to trigger closure at 80% of the total episode length (T_{total}). As stated in Eq. (4), the resulting binary command, g_t, remains in an open state (0) until the triggering event t_{grasp} or failsafe occurs, and is held in a closed state (1) for the duration of the trajectory.

$$g_t = \begin{cases} 0\ (Open), & \text{if } t < t_{grasp} \\ 1\ (Closed), & \text{if } t \geq t_{grasp} \end{cases} \tag{4}$$

Because the pi0.5 architecture is a VLA, the final stage of the retarget pipeline required pairing every kinematic trajectory with a corresponding natural language instruction. These prompts were generated by extracting semantic object labels directly from each episode metadata file. Once the object being interacted with ($object\ x$) is retrieved, the string is stored as “Pick up $object\ x$ and hold it.”.

The resulting dataset comprises 1000 episodes across 20 objects. A critical characteristic of this dataset is the presence of a visual domain shift (the action labels correspond to the robot, while the visual observations contain the original images of human hands). This experimental design explicitly tests the capacity of the policy to learn semantic action affordances independent of the agent's visual embodiment.

To synthesize these kinematic constraints into a continuous processing pipeline, the pseudocode presented in Algorithm 1 outlines the complete, frame-by-frame execution loop used to retarget the human DexYCB demonstrations into synchronized SO-101 robotic joint trajectories.

3.3 Training Protocol

Three distinct pi0.5 policies were trained for 100k steps, with checkpoints saved every 20k steps. The training corpus comprises two distinct data sources: a standard teleoperation dataset and the explicitly retargeted DexYCB dataset.

The teleoperation dataset contains 100 episodes collected via a leader-follower configuration, executing the specific task: “Pick up the black marker and place it in the box”. Hence, these demonstrations were recorded in an environment containing only the marker, with zero visual distractors, and using the physical setup presented in Section 3.1. The teleoperation dataset was intentionally limited to 100 episodes to evaluate the proposed approach in a realistic low-data regime. This setting reflects the practical motivation of the study: teleoperation is costly and time-

Algorithm 1
Retarget Pipeline Pseudocode

// 1. Initialization and Parameter Definition
Parameters:
$T_{calib} \in SE(3)$ (Base calibration matrix)
$T_{offset} \in SE(3)$ (Wrist-to-gripper matrix)
λ = 0 (IK rotation weight)
α = 0.35 (EMA smoothing filter)
r = 0.045 (Noise threshold in meters)
q_{max} (Actuator velocity limits)
Δt (Time step per frame)

Initialize variables:
Q = [] (Empty joint trajectory array)
G = [] (Empty gripper state array)
T_H = Load_MANO_Poses()
O = Load_Object_Trajectory()
q_{prev} = Robot_Home_Configuration
obj_{init} = O[0] (Initial resting position of object)

// 2. Execution Loop
FOR t = 1 **TO** length(T_H) **DO:**
// Spatial Alignment
$T_{Target} = T_{calib} * T_H[t] * T_{offset}$
// Gripper Heuristic
Δd = Euclidean Distance (O[t], obj_{init})
IF $\Delta d > r$ **THEN:**
g_t = 1 (Trigger gripper close)
ELSE:
g_t = 0 (Maintain open state)
END IF

// Dynamic Bounding
q_{min} = q_{prev} - (q_{max} * Δt)
q_{max} = q_{prev} + (q_{max} * Δt)

// Kinematic Optimization (Sequentially Seeded)
q_{raw} = IK_Solve(target = T_{Target}, seed = q_{prev}, bounds = [q_{min} ,q_{max}], weight = λ)

// Temporal Smoothing (EMA Filter)
q_t = α * q_{raw} + (1- α) * q_{prev}

// State Update and Storage
APPEND q_t TO Q
APPEND g_t \text{ TO G
q_{prev} = q_t

END FOR

// 3. Metadata and Compilation
L_{inst} = Generate_Language_Prompt (Object_Label)

RETURN Q, G, L_{inst}

consuming, especially for laboratories with limited robotic infrastructure. The goal was therefore not to maximize absolute performance through large-scale robot data collection, but to test whether retargeted human demonstrations can improve policy learning when only a small teleoperation set is available.

In contrast, the retargeted DexYCB dataset contributes an additional 1000 episodes of cross-embodiment demonstrations. The dataset includes 8 perspectives for each episode, and to ensure ease of replication on the physical SO-101 hardware, a single fixed perspective (an exocentric view) was chosen for the scope of this research. These episodes are conditioned on language instructions in the format "Pick up $object\ x$ and hold it.", spanning 20 diverse household objects. Unlike the teleoperation data, every DexYCB episode features dense visual clutter, containing the target object alongside three to five random semantic distractors. Notably, the objects appear as the target in 50 episodes and, in the remaining episodes, they are used as distractors. One of the objects present in the retargeted dataset is a black marker; hence, the respective 50 episodes of a human grasping a black marker amidst clutter provide a relevant semantic bridge to the downstream teleoperated task.

As summarized in Table 2, the three pi0.5 policies employ distinct training paradigms, leveraging different subsets of the data pool.

- *Control group* (teleoperation only) – Fine-tuned exclusively on the 100 teleoperated robot episodes.
- *Co-trained* (single-stage) – Fine-tuned simultaneously on a mixed dataset of 100 real teleoperated episodes and 50 retargeted DexYCB episodes (specifically humans grasping a black marker).
- *Cross-embodiment* (two-stage fine-tuning) – First stage fine-tuning on 1000 retargeted DexYCB episodes spanning all 20 objects, followed by a second-stage fine-tuning utilizing the mixed dataset (100 real plus 50 retargeted marker episodes).

Table 2
Dataset split per policy

Policy	Teleoperation Data	Retargeted DexYCB
Control group	100	0
Co-trained	100	50 (Black marker only)
Cross-embodiment 1st stage	0	1000 (all 20 objects)
Cross-embodiment 2nd stage	100	50 (Black marker only)

To address the visual discrepancy between the human hands observed during DexYCB training and the robot arm observed during inference, a frozen backbone strategy was adopted for the VLA architectures. By freezing the weights of the pre-trained VLM, the policy is forced to rely on its robust, existing internal semantic representations of "grasping". This restricts the training update entirely to the action head, facilitating the learning of the SO-101 kinematics while preventing the visual encoder from overfitting the pixel distribution of human skin. This strategy drastically reduces the trainable parameter space, optimizing computational efficiency.

3.4 Experiments

The resulting policies are evaluated on a standardized pick-and-place task across three distinct physical experiments. The core task consists of grasping a black marker and placing it into a designated box - "Pick up the black marker and place it in the box". The baseline teleoperation recordings consisted exclusively of executing this task on a clean table with no distractors. In contrast,

the DexYCB episodes consist of a human grasping an object and holding it approximately 20 cm above the table, surrounded by various distractor objects.

3.4.1 In-distribution task proficiency (experiment 1)

Experiment 1 evaluates the model's accuracy on the baseline task (clean table, varying marker positions/orientations, example shown in Figure 2). A primary concern when injecting synthetically retargeted data into a continuous VLA architecture is that noisy human priors might degrade the precise kinematic execution learned from real-world teleoperation. Therefore, Experiment 1 is designed to answer a foundational research question (RQ1): **Does the inclusion of retargeted human data improve the policy's execution on the original teleoperated task?**

Fig. 2. Image of experiment 1 setup

By evaluating the models on the baseline, In-Distribution (ID) environment (a clean workspace with no distractors, the same set used during teleoperation recording), it is tested whether the synthetic data induces degrading noise or successfully enhances task proficiency. Furthermore, due to the limited size of the real-world teleoperation dataset, offline validation was impractical. Hence, optimal checkpoints were selected empirically through this experiment and physical validation sweeps across the 40k-100k step range. Checkpoints were mathematically justified by selecting the epoch that yielded the highest physical success rate and lowest execution time prior to the onset of policy collapse.

3.4.2 Single distractor (experiment 2)

While Experiment 1 established baseline proficiency in a sterile environment, real-world deployment requires policies that do not collapse under visual clutter. Experiment 2 introduces a localized visual distractor to evaluate robustness to visual noise, hence posing the following research question (RQ2): **Does training on diverse human videos enhance the VLA's visual robustness when exposed to localized clutter?**

The single distractor object is placed 5 cm away from the target object, the black marker. As illustrated in Figure 3, a banana was used as the ID distractor (present in the DexYCB dataset) and a blue duct tape roll as the Out-of-Distribution (OOD) distractor (absent from DexYCB).

Fig. 3. Images of experiment 2 setup (ID distractor on the left, OOD distractor on the right)

3.4.3 Dense clutter (experiment 3)

Experiment 3 was designed to induce extreme visual noise to stress-test the policies. Experiment 3 introduces dense workspace clutter, placing three distinct objects within a 5 cm radius of the target marker. This configuration addresses the final research question (RQ3): **How do explicitly human retargeted spatial priors perform under conditions of extreme visual noise?**

As presented in Figure 4, the models were evaluated against both a densely packed ID configuration (banana, round can, servo box) and an OOD configuration (blue duct tape, clothespin, pink cup).

Fig. 4. Images of experiment 3 setup (ID distractors on the left, OOD distractors on the right)

3.5 Evaluation Framework

To ensure reproducibility across experiments, the physical testing workspace was divided into four quadrants, as illustrated in Figure 5. For each run, the target marker was placed in a specific quadrant at a predetermined orientation (north, east, west, and south for the policy-training steps checkpoint evaluation and Experiment 1, north and east for Experiments 2 and 3) and a score was assigned. Hence, the policy-training checkpoint used in the experiments was selected after the mentioned initial evaluation, which was done in the same test conditions as Experiment 1. Within experiments, each policy was evaluated using the same spatial quadrant and object-orientation protocol to ensure sample consistency. Experiment 1 used N = 32 trials per policy at the selected checkpoint. Experiments 2 and 3 used N = 16 trials per policy and distractor type (adding up to 32 trials per Experiment). All policies were evaluated under identical workspace geometry, camera position, lighting, prompt, and evaluation criteria.

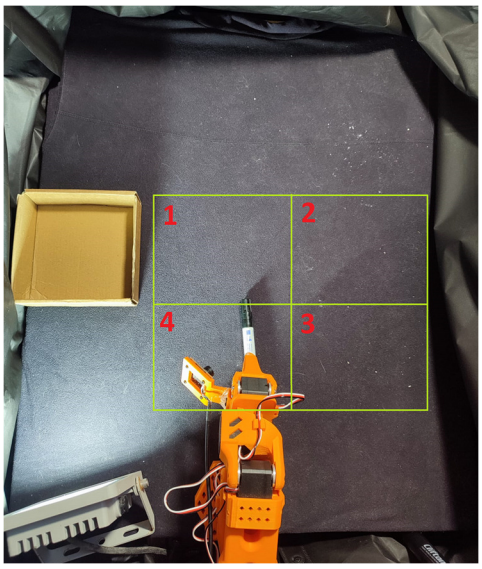

Fig. 5. Evaluation quadrants

Statistical significance between policy performances was evaluated using the non-parametric Mann-Whitney U test. This test was selected due to the metrics used for physical evaluation, specifically the evaluation criteria which include discrete milestones and the inherently right-skewed execution times, both of which violate the assumption of normality required for standard parametric testing. Furthermore, to quantify the magnitude of the observed performance differences, non-parametric effect sizes were calculated using the rank-biserial correlation (r).

Finally, Table 3 presents the evaluation criteria grid applied, which is based on physical milestones used to empirically evaluate the execution of the policies tested.

Table 3
Evaluation grid

Score	Milestone achieved	Description of physical behaviour	What it suggests about the policy
0	Critical Failure	Does not move, moves erratically, or grabs the wrong object (semantic failure).	Vision Encoder failed to identify the target, or Action Head suffered IK collapse.
0.25	Approach	End-effector correctly navigates to the target's XYZ vicinity, but fails to make contact (stops short, closes early).	Spatial/Kinematic awareness is working, but depth perception or action-timing is misaligned.
0.5	Contact	Physically touches the correct object, but fails to secure a grasp (knocks it over, gripper slips, bad wrist rotation).	Accurate semantic targeting, but lacks fine motor control or geometric orientation awareness.
0.75	Lift / Hold	Successfully grasps and lifts the object off the table but drops it mid-air or misses the final drop zone (the box).	Excellent visual grounding and grasping; failure is localized on the long-horizon trajectory.
0.9	Task Success / Rogue	Flawlessly grasps, lifts, and drops the target, but fails to return to the neutral base state (lingers or attempts secondary grasps).	Policy capable of executing the task, but suffers from unsafe continuous motion priors or lack of termination conditions.
1	Perfect Execution	Flawlessly grasps, lifts, and drops the target into the box, returning safely to base.	Perfect end-to-end VLA execution. Executes the task and returns to the initial position.

4. Results

4.1 In-Distribution Task Proficiency (Experiment 1)

The efficacy of explicit kinematic retargeting was evaluated by comparing the Control Group (teleoperation only) policy against the Co-Trained and Cross-Embodiment policies.

To ensure rigorous evaluation, all policies were tested across checkpoints (40k to 100k steps) utilizing a baseline N=16 trial matrix (two orientations for each quadrant, repeated twice), for the task of "Pick up the black marker and place it in the box" - the same prompt used during teleoperation recording. As shown in Table 4, the 100k checkpoint was empirically selected across all architectures, as it consistently yielded the highest physical success rates and lowest execution times (see Table 4).

Table 4
Checkpoint steps evaluation (*N* = 16 per policy and checkpoint)

Policy	Checkpoint steps	Task score mean	Task time mean
Control croup	100k	0.68	25.33
	80k	0.54	26.45
	60k	0.62	24.88
	40k	0.51	28.51
Co-trained	100k	0.81	21.28
	80k	0.68	26.38
	60k	0.62	25.42
	40k	0.64	26.03
Cross-embodiment (two-stage)	100k	0.86	21.70
	80k	0.76	25.15
	60k	0.78	23.08
	40k	0.68	26.62

To increase statistical confidence, each optimal 100k checkpoint was subjected to an expanded *N*=32 evaluation matrix encompassing all four spatial quadrants and four distinct target orientations (north, east, south, west). As synthesized in Table 5, empirical results demonstrate that explicit geometric retargeting not only preserves existing kinematic knowledge but yields statistically significant improvements in downstream task proficiency. The Control Group policy achieved an average task score of 0.69, a perfect run success rate of only 40.6%, with a mean execution time of 24.50 seconds.

Conversely, the Co-Trained policy, which integrates just 50 episodes of human-retargeted grasping data (grasping a similar black marker) demonstrated a significant performance leap. It achieved an average score of 0.84, translating to a 65.6% success rate of perfect execution. The Mann-Whitney U test confirms that this performance increase is statistically significant ($r = 0.2793$ and $p = 0.036$ against the Control Group). Furthermore, the Co-Trained model reduced execution time to 21.31 seconds ($r = 0.2910$ and $p = 0.035$).

The Cross-Embodiment (Two-Stage) policy exhibited the highest physical execution metrics across the entire experiment. Pre-trained on 1000 diverse DexYCB episodes before fine-tuning, this model achieved a peak average score of 0.88 ($r = 0.3691$ and $p = 0.004$) and a remarkable 78.1% absolute success rate. Moreover, the execution time was significantly reduced to 20.25 ($r = 0.3213$ and $p = 0.023$).

This validates the core hypothesis: broad geometric priors extracted from human manipulation across diverse, OOD objects fundamentally improve the underlying dexterity and confidence of the policy, even if those geometric priors do not exactly match the downstream task.

Table 5
Aggregate experiment 1 performance (*N* = 32 per policy)

Policy	Metric evaluation	Performance (Mean ±Var [95% CI])	Statistical significance against Control Group (Rank-Biserial r, and Mann-Whitney U test p)
Control group	Task score	0.69 ± 0.08 [0.58, 0.79]	-
	Execution time (seconds)	24.50 ± 55.04 [21.82, 27.18]	-
	Success rate	40.6% ± 0.25 [23.7, 59.4]	-
Co-Trained	Task score	0.84 ± 0.07 [0.74, 0.93]	r = 0.2793, p = 0.036
	Execution time (seconds)	21.31 ± 61.50 [18.48, 24.14]	r = 0.2910, p = 0.035
	Success rate	65.6% ± 0.24 [46.8, 81.4]	r = 0.2500, p = 0.048
Cross-embodiment (two-stage)	Task score	0.88 ± 0.06 [0.79, 0.97]	r = 0.3691, p = 0.004
	Execution time (seconds)	20.25 ± 46.09 [17.80, 22.70]	r = 0.3213, p = 0.023
	Success rate	78.1% ± 0.18 [60.0, 90.7]	r = 0.3750, p = 0.003

Figure 6 provides a granular breakdown of the specific failure modes encountered during physical execution. The Control Group frequently exhibited early grasping failures, with approximately half of all its attempted runs terminating prematurely at either the “Approach” (0.25) or “Contact” (0.50) milestones. This visualizes a limitation in fine motor control and geometric alignment when the policy relies solely on a small teleoperation dataset. In contrast, the injection of explicitly retargeted human data systematically compressed this failure distribution. Most notably, the Co-Trained and Cross-Embodiment policies drastically reduced the incidence of “Contact” (0.50) failures—the primary failure mode of the baseline. This demonstrates that the geometric approach priors extracted from the DexYCB dataset successfully stabilized the gripper's spatial alignment before closure. Ultimately, the two-stage Cross-Embodiment policy not only nearly doubled the perfect execution rate (1.0), but uniformly suppressed errors across all intermediate milestones, showcasing a robust, end-to-end alignment of the Vision-Language-Action pipeline.

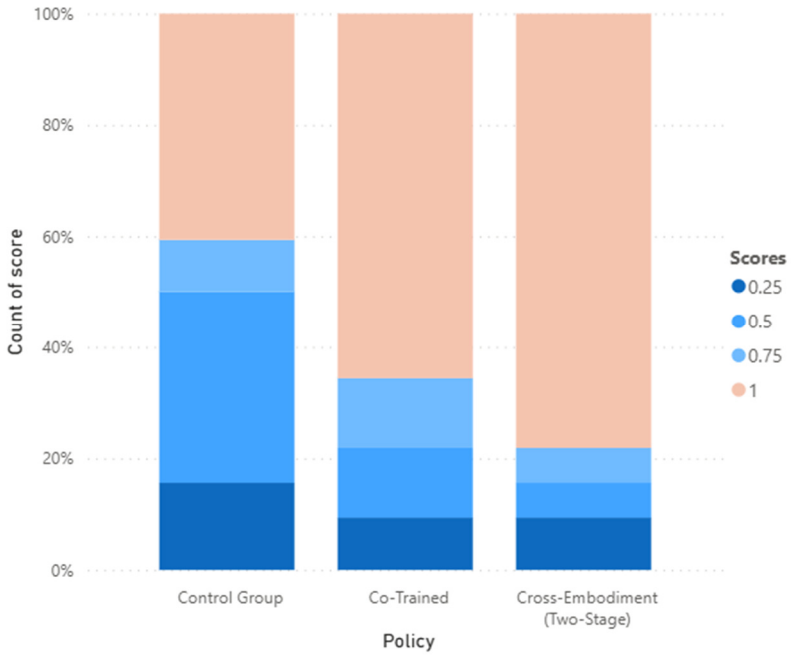

Fig. 6. Experiment 1 score breakdown by policy

4.2 *Single Distractor (Experiment 2)*

To isolate failure modes, two distinct distractor profiles were tested, namely an ID (a banana, featured in the DexYCB dataset) and an OOD object (a blue duct tape roll, absent from the retargeting data). As shown in Table 6, the presence of the ID distractor exposed a severe vulnerability in the Control Group model. When the banana was introduced, the model suffered a catastrophic collapse, yielding a 0% absolute success rate and a mean task score of just 0.40. Qualitative observations (e.g., "Hallucinated mid-air", "Tried to grasp next to the object") indicate severe attention bleed from the target object. One explanation is that, although the VLA recognized the banana's semantic embedding, it failed to isolate the target marker, resulting in kinematic failures from the beginning.

Table 6
Aggregate experiment 2 performances (*N*=16 for each policy and distractor type)

Policy	Evaluation Metric	ID Distractor		OOD Distractor	
		Performance (Mean ± Var [95% CI])	**Statistical significance against Control Group**	**Performance (Mean ± Var [95% CI])**	**Statistical significance against Control Group**
Control group	Task score	0.40 ± 0.04 [0.30, 0.51]	-	0.60 ± 0.12 [0.41, 0.79]	-
	Success rate	0% ± 0.0 [0.0, 0.0]	-	18.7% ± 0.16 [4.0, 45.6]	-
Co-trained	Task score	0.61 ± 0.09 [0.45, 0.77]	r = 0.3867, p = 0.051	0.81 ± 0.09 [0.65, 0.96]	r = 0.3867, p = 0.056
	Success rate	25% ± 0.20 [7.3, 52.4]	r = 0.2500, p = 0.038	56.2% ± 0.26 [29.9, 80.2]	r = 0.3750, p = 0.033
Cross-embodiment (two-stage)	Task score	0.80 ± 0.07 [0.67, 0.94]	r = 0.7539, p = 0.0002	0.74 ± 0.09 [0.57, 0.90]	r = 0.1875, p = 0.362
	Success rate	43.7% ± 0.26 [19.8, 70.1]	r = 0.4375, p = 0.004	18.7% ± 0.16 [4.0, 45.6]	r = 0.0, p = 1

Conversely, the policies augmented with explicit human retargeting demonstrated profound semantic anchoring. The Cross-Embodiment model successfully ignored the banana, elevating the mean score to 0.80 and achieving a 43.75% success rate. A Mann-Whitney U test confirms this recovery is highly statistically significant (r = 0.7539, p = 0.0002 against the Control Group for mean score and r = 0.4375, p = 0.004). By pre-training on DexYCB retargeted episodes with a frozen visual backbone, the model learned to strictly associate the action sequence with the specific language instruction, rather than firing grasping primitives at the nearest recognized object.

When exposed to the novel object (blue duct tape), the Control Group model performed marginally better than it did against the banana, achieving a mean score of 0.60 and an 18.75% success rate. This confirms that the banana was acting as a semantic trap rather than just a physical obstacle.

However, the retargeted policies maintained superior robustness even against this unlearned visual noise. The Co-Trained model exhibited the highest OOD immunity, achieving a mean score of 0.81 and an impressive 56.25% success rate. This suggests that while the larger two-stage policy (Cross-Embodiment) is optimal for navigating known semantic clutter, the single-stage Co-Trained injection of human spatial priors is highly efficient at filtering out pure background noise, maintaining the policy's kinematic focus on the target object.

A relevant observation is that the Cross-Embodiment model successfully grasped the marker near the duct tape most of the time, attaining a score of 0.74. However, it struggled to return to base after dropping the marker in the box (receiving a score of 0.9), which explains the success rate of 18.7%. The Co-Trained model was therefore the most successful against the duct tape.

4.3 *Dense Clutter (Experiment 3)*

Table 7 presents the aggregate performance results from experiment 3. Under extreme visual clutter, the Control Group suffered severe performance degradation. In the ID configuration, it achieved a mean score of only 0.58, with a perfect success rate of just 18.7%. In the OOD configuration, it failed completely to achieve a single perfect run, therefore achieving 0% success and a mean score of 0.50. The dense visual noise overwhelmed the model's spatial reasoning, leading to frequent early termination at the “Approach” or “Contact” milestones.

Table 7
Aggregate experiment 3 performances. N=16 for each policy and distractor type

Policy	Evaluation metric	ID Distractor		OOD Distractor	
		Performance (Mean ± Var [95% CI])	**Statistical significance against Control Group**	**Performance (Mean ± Var [95% CI])**	**Success Rate Statistical significance against Control Group**
Control group	Task score	0.58 ± 0.11 [0.40, 0.94]	-	0.50 ± 0.17 [0.28, 0.72]	-
	Success rate	18.7% ± 0.16 [4.0, 45.6]	-	0% ± 0.0 [0.0, 0.0]	-
Co-trained	Task score	0.80 ± 0.04 [0.69, 0.90]	r = 0.3594, p = 0.079	0.65 ± 0.16 [0.44, 0.87]	r = 0.2344, p = 0.235
	Success rate	25% ± 0.20 [7.3, 52.4]	r = 0.0625, p = 0.693	12.5% ± 0.12 [1.6, 38.3]	r = 0.1250, p = 0.164
Cross-Embodiment (two-stage)	Task score	0.66 ± 0.07 [0.52, 0.80]	r = 0.0859, p = 0.685	0.65 ± 0.12 [0.47, 0.84]	r = 0.1562, p = 0.433
	Success rate	0% ± 0.0 [0, 0.0]	r = 0.1875, p = 0.080	0% ± 0.0 [0.0,0.0]	r = 0, p = 1

The Co-Trained policy demonstrated the highest overall robustness to extreme noise. In the dense ID configuration, it achieved 0.80, the highest mean score of the experiment. Even under OOD clutter, it maintained the highest task score of 0.65, demonstrating that a targeted dataset of just 50 human approach vectors provides a highly stabilized geometric anchor for the VLA, preventing early kinematic collapse.

The most critical architectural insight in this experiment emerged from the behaviour of the two-stage Cross-Embodiment model. As identified in Table 7 and Figure 7, this model yielded a 0% absolute success rate under both ID and OOD dense clutter. Its overall mean task scores, 0.66 for ID clutter and 0.65 for OOD clutter, failed to achieve statistical significance over the Control Group. However, a granular analysis of the failure milestones reveals that the model successfully completed the physical objective (grasping and dropping the marker into the box) in nearly 50% of its runs. Instead of achieving a perfect score of 1.0, these runs uniformly terminated with a score of 0.90. The

robot flawlessly executed the manipulation but failed to return to its neutral base position, instead lingering over the box or executing erratic secondary movements.

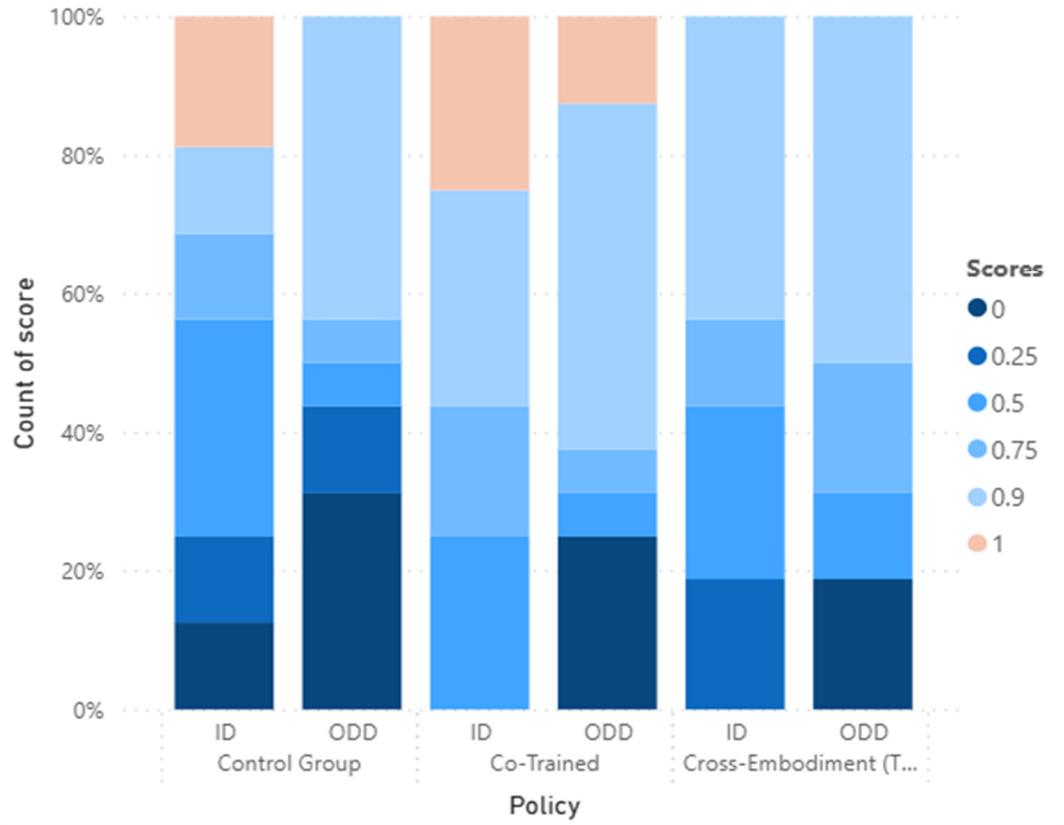

Fig. 7. Experiment 3 score breakdown by policy

Therefore, this severe temporal failure mode is formally characterized as TSA. It is defined as a condition where a generative continuous-control policy successfully achieves the physical manipulation objective but fails to output a definitive zero-velocity termination sequence, instead entering an unbounded kinematic loop. To quantify this vulnerability, we propose the TSA rate (TSA_R), calculated as the ratio of successful physical manipulations that fail to terminate safely within a standard episode horizon (T_{max}):

$$TSA_R = \frac{N_{no\_term}}{N_{success}} * 100 \quad (5)$$

where $N_{success}$ is the total number of episodes where the target objective was successfully manipulated (reaching the 0.90 score threshold), and N_{no_term} is the subset of those episodes where the policy failed to yield the end-of-episode state.

Architecturally, it is hypothesized that this ambiguity may be driven by visual distribution shifts acting as a latent routing mechanism within the fine-tuned Action Head. Analogous to how Mixture-of-Experts (MoE) architectures route tokens to specific sub-networks based on input context, the Action Head learns a conditional kinematic prior. When the frozen Vision Encoder processes extreme clutter, it outputs embeddings heavily correlated with the visually dense human retargeting dataset (DexYCB), which contains thousands of open-ended episodes that routinely cut off mid-air. Consequently, the Action Head routes the action generation toward this open-ended human kinematic prior, overshadowing the explicit semantic prompt condition ("Pick up the black marker

and place it in the box"). Because the generative VLA is forced to rely on these foundational human priors under severe visual noise, it bypasses the rigid "return to initial position" trajectories found strictly in the visually sparse teleoperation dataset, directly resulting in the observed action loops.

5. Discussion

5.1 Study Implications

Theoretically, the empirical outcomes of this study support the view that human manipulation demonstrations contain transferable geometric priors that can improve robot policy learning, even when visual embodiments differ substantially. This contributes to the understanding of cross-embodiment learning, demonstrating that explicit kinematic alignment can effectively complement large-scale emergent VLA transfer.

In practical terms, the proposed pipeline offers a pathway for scaling VLA fine-tuning in data-scarce regimes. By augmenting sparse robotic datasets with kinematically retargeted human demonstrations, this framework drastically reduces the field's systemic dependence on high-cost, logistically bottlenecked teleoperation data. Instead of relying on a linear 1-to-1 accumulation of expensive robot-specific trials, this paradigm introduces an alternative data-scaling trajectory. It unlocks the ability to leverage abundant, diverse human video repositories as direct action supervision.

5.2 Research Limitations

This study presents robust evidence for explicit kinematic retargeting in low-data regimes; however, several primary limitations contextualize the empirical findings. First, all physical experiments were restricted to a single 5-DoF manipulator (the SO-101). Consequently, these results serve as a proof-of-concept for accessible, low-DoF hardware rather than absolute validation for highly redundant robotic systems.

Second, the downstream task evaluation was limited strictly to pick-and-place operations.

Third, the real teleoperation dataset contained only 100 episodes. While this reflects the targeted low-data regime, it bounds the statistical power of the conclusions and may increase the model's sensitivity to the specific distribution of the collected demonstrations.

Fourth, the human demonstration source was restricted to the DexYCB dataset, which inherently does not capture the full diversity of manipulation behaviors required for broader, unconstrained robotic deployment. Finally, as formalized in Section 4.3, explicitly retargeted continuous-control policies remain susceptible to TSA due to the severe covariate shift between visually sparse and dense datasets.

5.3 Guidelines for Future Research

Future work should focus on resolving these morphological and data bottlenecks to further scale cross-embodiment transfer. First, to verify that the observed performance gains persist at scale, researchers must validate the pipeline on additional robotic embodiments, more diverse manipulation tasks, and significantly larger teleoperation datasets encompassing different camera viewpoints and object categories. Extending this methodology to higher-DoF manipulators should include exploring the relaxation of the strict positional-only IK constraint ($\lambda = 0$) to reintroduce rotational penalty terms.

Second, to explicitly address the temporal vulnerability of TSA, future policy training frameworks should test whether automated synthetic return-to-base segments, explicit discrete stop actions, or terminal-state classifiers can successfully enforce the termination conditions.

Finally, to rigorously quantify the contribution of individual pipeline components, future studies should conduct the systematic ablation plan outlined in Table 8. By isolating the impacts of the IK rotation weight (λ), the EMA smoothing filter (α), sequential seeding, and the grasp noise threshold (r), researchers can definitively map the mathematical trade-offs between kinematic responsiveness, temporal continuity, and overall downstream policy robustness.

Table 8
Suggested ablation plan

Component	Baseline value	Ablation values	Expected effect
IK rotation weight (λ)	0	0.1, 0.5, 1.0	Tests orientation/position trade-off
EMA Smoothing filter (α)	0.35	0.1, 0.5, 1.0	Tests jitter versus responsiveness
Sequential seeding	Enabled	Disabled	Tests temporal continuity
Noise threshold (r)	0.045 m	0.025, 0.065 m	Tests false versus delayed grasp trigger

6. Conclusions

This research aimed to overcome the embodiment gap in VLA policy training by proposing a lightweight, explicit kinematic retargeting pipeline. By translating human demonstrations into robot joint trajectories via IK, the proposed method successfully generates smooth, mathematically grounded spatial priors without relying on computationally expensive visual synthesis.

Physical evaluations on the pi0.5 architecture demonstrate that explicit retargeting significantly enhances baseline teleoperation. The two-stage Cross-Embodiment pretraining successfully resolved early-stage grasping bottlenecks, increasing the absolute task success rate from 40.6% to 78.1%. Furthermore, the explicitly retargeted policies exhibited visual robustness, successfully maintaining kinematic focus and filtering out both In-Distribution semantic traps and Out-of-Distribution visual noise.

Additionally, the evaluation under extreme visual clutter highlighted a severe temporal vulnerability in generative architectures, formally characterized in this study as TSA. As quantified in the dense clutter experiments, this limitation demonstrates that while the VLA can perfectly execute physical manipulation using human geometric priors, it fails to learn a definitive termination condition. This failure is driven by a fundamental task distribution mismatch: extreme visual clutter acts as a latent trigger, biasing the fine-tuned action head to soft-route action generation toward open-ended human trajectories rather than the controlled, zero-velocity sequences of teleoperation data. Therefore, future research exploring similar frameworks should address this ambiguity, potentially by integrating explicit boundary constraints or synthetic trajectory reversals into the retargeted action sequences. Replicating the termination distribution will be a critical step toward maximizing the scalable, safe transfer of human demonstrations to continuous robotic control.

Conflict of Interest
The author declares no conflict of interest.

Acknowledgment
This work was supported by national funds through FCT, I.P. (Fundação para a Ciência e a Tecnologia,I.P.), under the project - UID/04152/2025 - Centro de Investigacão em Gestão de

Informacão (MagIC)/NOVA IMS (DOI: 10.54499/UID/04152/2025), and by the European Union – NextGenerationEU under the project UID/PRR/04152/2025 (DOI: 10.54499/UID/PRR/04152/2025). This work was funded by the European Union through the project 101084013 - DIGITAL4Business. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. This research was not funded by any grant

References

[1] Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., ... & Zhu, Y. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734. (arXiv:2503.14734). arXiv. https://doi.org/10.48550/arXiv.2503.14734.

[2] Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., ... & Zhilinsky, U. (2025). π0.5: A Vision-Language-Action Model with Open-World Generalization (arXiv:2504.16054). arXiv. https://doi.org/10.15607/RSS.2025.XXI.010.

[3] Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications. *IEEE Access, 13*, 162467-162504. https://doi.org/10.1109/ACCESS.2025.3609980.

[4] Sapkota, R., Cao, Y., Roumeliotis, K. I., & Karkee, M. (2025). Vision-Language-Action Models: Concepts, Progress, Applications and Challenges (arXiv:2505.04769). arXiv. https://doi.org/10.48550/arXiv.2505.04769.

[5] Xiao, X., Liu, J., Wang, Z., Zhou, Y., Qi, Y., Jiang, S., He, B., & Cheng, Q. (2025). Robot learning in the era of foundation models: A survey. *Neurocomputing, 638*, 129963. https://doi.org/10.1016/j.neucom.2025.129963.

[6] Lepert, M., Fang, J., & Bohg, J. (2025). Masquerade: Learning from In-the-wild Human Videos using Data-Editing (arXiv:2508.09976). arXiv. https://doi.org/10.48550/arXiv.2508.09976.

[7] Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A. C., ... & Wang, X. (2025). EgoVLA: Learning Vision-Language-Action Models from Egocentric Human Videos (arXiv:2507.12440). arXiv. https://doi.org/10.48550/arXiv.2507.12440.

[8] Feng, Z., Li, Q., Liang, H., Yang, R., Shen, Y., Du, Z., Zhang, Z., Deng, Y., Zhao, L., Zhao, H., Lu, Z., Mees, O., Pollefeys, M., Yang, J., & Guo, B. (2026). From Human Videos to Robot Manipulation: A Survey on Scalable Vision-Language-Action Learning with Human-Centric Data. https://doi.org/10.36227/techrxiv.177126525.54038135/v1.

[9] Chao, Y. W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., ... & Fox, D. (2021). Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 9040-9049). IEEE. https://doi.org/10.1109/CVPR46437.2021.00893.

[10] Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., ... & Levine, S. (2023). Bridgedata v2: A dataset for robot learning at scale. In Conference on robot learning (pp. 1723-1736). PMLR.

[11] O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., ... & Chen, M. (2024, May). Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation* (ICRA) (pp. 6892-6903). IEEE. https://doi.org/10.1109/ICRA57147.2024.10611477.

[12] Qiu, R.-Z., Yang, S., Cheng, X., Chawla, C., Li, J., He, T., ... & Wang, X. (2025). Humanoid Policy ~ Human Policy (arXiv:2503.13441). arXiv. https://doi.org/10.48550/arXiv.2503.13441.

[13] Darvish, K., Tirupachuri, Y., Romualdi, G., Rapetti, L., Ferigo, D., Chavez, F. J. A., & Pucci, D. (2019). Whole-Body Geometric Retargeting for Humanoid Robots. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots* (Humanoids), 679-686. https://doi.org/10.1109/Humanoids43949.2019.9035059.

[14] Yin, Z.-H., Wang, C., Pineda, L., Bodduluri, K., Wu, T., Abbeel, P., & Mukadam, M. (2025). Geometric Retargeting: A Principled, Ultrafast Neural Hand Retargeting Algorithm (arXiv:2503.07541). arXiv. https://doi.org/10.1109/IROS60139.2025.11247700.

[15] Fu, Z., Zhao, Q., Wu, Q., Wetzstein, G., & Finn, C. (2024). HumanPlus: Humanoid Shadowing and Imitation from Humans. (arXiv:2406.10454). arXiv. https://doi.org/10.48550/arXiv.2406.10454.

[16] Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., ... & Xu, D. (2025). Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation* (ICRA) (pp. 13226-13233). IEEE. https://doi.org/10.1109/ICRA55743.2025.11127989.

[17] Jeong, T., Byun, T., Kim, J., Choi, K., Oh, J., Lee, S., Darwish, O., Kim, J., & Choi, S. (2024). Robust Robot Motion Retargeting: Rig Unification and Application to Diverse Robots. *Research Square*. https://doi.org/10.21203/rs.3.rs-4544618/v1.

[18] Huang, H., Bethala, G. C. R., Yuan, S., Wen, C., Tzes, A., & Fang, Y. (2025). One-shot Humanoid Whole-body Motion Learning (arXiv:2510.25241). arXiv. https://doi.org/10.48550/arXiv.2510.25241.

[19] Yan, Y., Mascaro, E. V., & Lee, D. (2024). ImitationNet: Unsupervised Human-to-Robot Motion Retargeting via Shared Latent Space (arXiv:2309.05310). arXiv. https://doi.org/10.1109/Humanoids57100.2023.10375150.

[20] Park, S., Lee, S., Choi, M., Lee, J., Kim, J., Kim, J., & Joo, H. (2025). Learning to Transfer Human Hand Skills for Robot Manipulations (arXiv:2501.04169). arXiv. https://doi.org/10.48550/arXiv.2501.04169.

[21] Canh, T. N., Tran, T.-T., Zhang, H., Gao, Z., Chong, N. Y., & HoangVan, X. (2026). Human-to-Robot Interaction: Learning from Video Demonstration for Robot Imitation (arXiv:2602.19184). arXiv. https://doi.org/10.48550/arXiv.2602.19184.

[22] Lang, X., Wang, Y., Zhou, Y., Ni, C., Li, K., Zhu, J., … & Zhu, Z. (2026). VAG: Dual-Stream Video-Action Generation for Embodied Data Synthesis (arXiv:2604.09330). arXiv. https://doi.org/10.48550/arXiv.2604.09330.

[23] Li, H., Zhang, I., Ouyang, R., Wang, X., Zhu, Z., Yang, Z., …. & Wang, X. (2025). MimicDreamer: Aligning Human and Robot Demonstrations for Scalable VLA Training (arXiv:2509.22199). arXiv. https://doi.org/10.48550/arXiv.2509.22199.

[24] Lepert, M., Fang, J., & Bohg, J. (2025). Phantom: Training Robots Without Robots Using Only Human Videos (arXiv:2503.00779). arXiv. https://doi.org/10.48550/arXiv.2503.00779.

[25] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., … & Zhilinsky, U. (2024). π0: A Vision-Language-Action Flow Model for General Robot Control (arXiv:2410.24164; Version 1). arXiv. https://doi.org/10.15607/RSS.2025.XXI.010.

[26] Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., & Lu, Z. (2025). Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos (arXiv:2507.15597). arXiv. https://doi.org/10.48550/arXiv.2507.15597.