TATE: Teleoperation-Anchored Transformation of Egocentric Data for Robot Learning

Le Xu[1,2], Xiaoquan Sun[1,2], Yijun Hong[2,3], Mingqi Yuan[1,2,†], Jiayu Chen[1,2,†]
[1]HKU [2]INFIFORCE [3]SUSTech

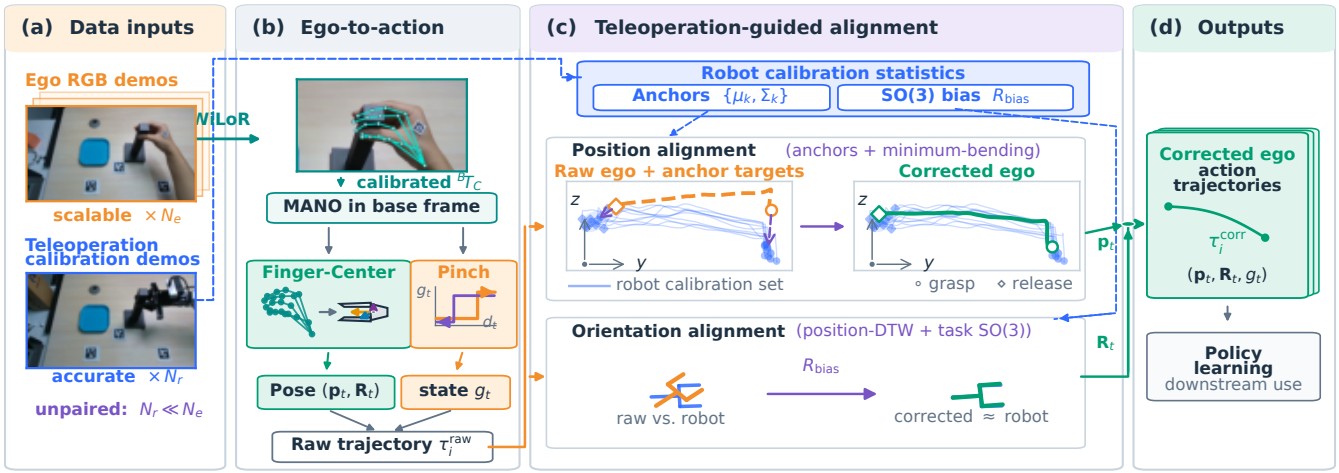

Fig. 1. Overview of TATE. Fixed ego-view RGB demonstrations are converted into raw virtual gripper trajectories with WiLoR and Finger-Center. A small teleoperation set provides grasp/release anchor statistics and a task-level orientation correction, producing calibrated ego-derived action labels.

***Abstract*—Egocentric human video is scalable but lacks reliable robot action labels. We introduce TATE (*Teleoperation-Anchored Transformation of Egocentric Data*), which uses a small teleoperation set to calibrate 6-DoF end-effector trajectories recovered from fixed ego-view RGB videos. TATE maps monocular hand reconstructions to virtual gripper poses with Finger-Center, aligns grasp/release anchors through minimum-bending position correction, and removes orientation bias with a task-level $\mathrm{SO}(3)$ rotation. On three stacking tasks, TATE reduces normalized pose discrepancy from 2.65, 4.23, and 4.73 to 1.64, 1.69, and 1.62, approaching robot-to-robot variation.**

***Index Terms*—Egocentric vision, robot learning, human-to-robot transfer, trajectory alignment.**

I. Introduction

Embodied intelligence is moving from task-specific policies trained on small, closed datasets toward foundation models built from open, heterogeneous, and continuously expanding data resources [1], [2]. This shift makes scalable action supervision a bottleneck: real-robot teleoperation gives accurate embodiment-matched states and actions but is costly, while egocentric human video is cheap and diverse but rarely contains robot-compatible 6-DoF labels. Existing ego pipelines either rely on expensive instrumented capture or plain RGB video whose actions are absent, weakly inferred, lack absolute scale, or ignore camera geometry [3]–[5]. These data help pretraining, but are weak as direct action supervision.

† Corresponding authors: Mingqi Yuan (my017@hku.hk) and Jiayu Chen (jiayuc@hk.hk).

We therefore ask whether a small teleoperation set can upgrade low-cost fixed ego-view video into scalable 6-DoF labels. We propose **TATE, Teleoperation-Anchored Transformation of Egocentric Data**, which reconstructs hands with WiLoR [6], maps hand geometry to a virtual gripper through Finger-Center, and uses task-matched teleoperation as a compact geometric prior. Grasp/release events anchor minimum-bending position calibration, while a task-level rotation in $\mathrm{SO}(3)$ removes systematic orientation bias. On three stacking tasks, TATE narrows the ego-to-robot gap toward held-out robot variation.

II. Method

TATE uses many unpaired fixed ego-view RGB demonstrations and a smaller task-matched teleoperation set. Ego videos provide visual coverage, while teleoperation supplies accurate end-effector poses in the robot base frame. As shown in Fig. 1, TATE recovers a raw virtual gripper trajectory from each human video, calibrates its position and orientation with robot statistics, and outputs RGB, pose, and gripper state for action supervision.

A. Ego-View to Virtual Gripper

For each ego-view frame, WiLoR [6] reconstructs MANO hand landmarks [7], and a calibrated camera-to-base transform expresses the virtual end-effector pose in robot coordinates. Valid frames contain position, orientation, and binary gripper state.

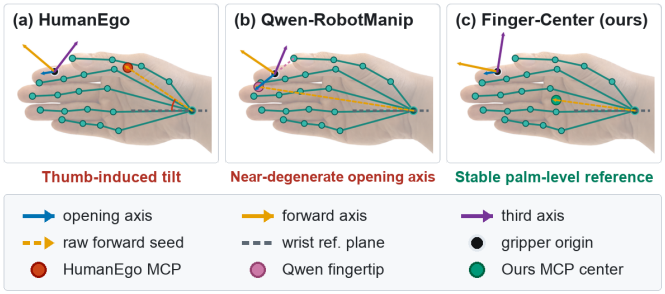

Fig. 2. Hand-to-gripper frame constructions. Finger-Center avoids the thumb-induced forward-axis bias of HumanEgo and the near-closure instability of Qwen-RobotManip's fingertip-derived axes.

Fig. 2 compares hand-to-gripper mappings. HumanEgo uses a forward seed from the wrist to the thumb–index MCP midpoint, which can tilt upward in tabletop pinches [4]. Qwen-RobotManip derives axes from fingertip geometry; near closure, its opening baseline becomes short and noise sensitive [5]. TATE instead uses *Finger-Center*: the origin and opening direction follow thumb–index pinch geometry, but the forward seed points from the wrist to the center of the four non-thumb MCP joints. Projecting this seed against the opening axis yields a stable palm-level frame for level, oblique, and vertical grasps.

We infer gripper state from the normalized thumb–index fingertip distance. Let $\mathbf{f}_T(t)$ and $\mathbf{f}_I(t)$ be the thumb and index fingertips, and let $\mathbf{w}(t)$ and $\mathbf{m}_M(t)$ be the wrist and middle-finger MCP joint. We compute $d_t = \|\mathbf{f}_T(t) - \mathbf{f}_I(t)\|_2 / \|\mathbf{w}(t) - \mathbf{m}_M(t)\|_2$ and apply hysteresis:

$$g_t = \begin{cases} 1, & d_t \le \theta_{\text{close}}, \\ 0, & d_t \ge \theta_{\text{open}}, \\ g_{t-1}, & \text{otherwise}, \end{cases} \qquad \theta_{\text{close}} < \theta_{\text{open}}. \tag{1}$$

A minimum-duration filter removes flicker, and stable transitions serve as grasp/release anchors.

B. Teleoperation-Anchored Calibration

For position, TATE fits a Gaussian location distribution to each robot anchor. Each ego trajectory samples target grasp/release anchors from these distributions, avoiding collapse to a single mean. Let $\boldsymbol{\delta}_{i,t}$ be the displacement applied to ego trajectory i at frame t. Subject to the sampled anchor constraints, we estimate the smoothest correction by minimizing discrete bending energy:

$$\boldsymbol{\delta}_i^\star = \arg\min_{\boldsymbol{\delta}_i} \sum_{t=1}^{T_i-2} \|\boldsymbol{\delta}_{i,t+1} - 2\boldsymbol{\delta}_{i,t} + \boldsymbol{\delta}_{i,t-1}\|_2^2 . \tag{2}$$

The corrected position is $\mathbf{p}_{i,t}^e + \boldsymbol{\delta}_{i,t}^\star$. This deforms the whole trajectory toward the sampled anchors while preserving the original motion trend.

For orientation, TATE estimates one task-level rotation in $\mathrm{SO}(3)$. Because human and robot demonstrations move at different speeds, task-critical segments are aligned with position-based dynamic time warping (DTW) [8]. Relative rotations along the DTW correspondences are averaged as the task bias, and its inverse is applied to all ego-derived gripper orientations. Combined with position calibration, this forms the final TATE trajectory.

TABLE I
HAND-TO-GRIPPER ORIENTATION ERROR IN DEGREES BEFORE AND AFTER TASK-LEVEL $\mathrm{SO}(3)$ CALIBRATION; LOWER IS BETTER.

Conversion	Level		Oblique		Vertical	
	Raw	SO(3)	Raw	SO(3)	Raw	SO(3)
HumanEgo	32.1	10.1	30.6	11.7	47.9	**11.6**
Qwen	28.2	11.9	41.1	12.7	**35.5**	21.5
Finger-Center	**14.6**	**10.0**	**16.5**	**11.5**	41.7	**11.6**

TABLE II
TATE CALIBRATION RESULTS WITH FINGER-CENTER CONVERSION. LOWER ρ_{pose} IS BETTER, AND η_{disp} SHOULD APPROACH ONE.

Task	ρ_{pose}	D_{pos}	D_{rot}	η_{disp}
Level	$2.65 \to \mathbf{1.64}$	$45.9 \to \mathbf{25.3}$	$14.6 \to \mathbf{10.1}$	$1.75 \to \mathbf{1.10}$
Oblique	$4.23 \to \mathbf{1.69}$	$92.1 \to \mathbf{21.0}$	$16.5 \to \mathbf{11.5}$	$2.73 \to \mathbf{1.52}$
Vertical	$4.73 \to \mathbf{1.62}$	$72.9 \to \mathbf{31.5}$	$41.7 \to \mathbf{11.4}$	$1.72 \to \mathbf{1.54}$

III. EVALUATION

We evaluate TATE on *Level*, *Oblique*, and *Vertical* stacking tasks, covering tabletop-parallel, inclined, and near-upright grasps under a common pick-and-place objective. Each task uses 30 fixed ego-view demonstrations and 20 teleoperation demonstrations, with 10 robot trajectories for calibration and 10 for held-out evaluation. Metrics use the grasp-to-release segment.

After resampling and position-DTW, we report position error D_{pos}, orientation error D_{rot}, normalized pose discrepancy $\rho_{\text{pose}} = \sqrt{\rho_{\text{pos}}\rho_{\text{rot}}}$, and dispersion ratio η_{disp}. Values near one mean ego-to-robot error and spread are close to robot-to-robot variation. Because evaluation is unpaired, DTW handles unmatched timing; evaluating orientation on the same path checks the stage-aligned gripper attitude.

Table I isolates hand-to-gripper conversion. Finger-Center gives the lowest raw orientation error on Level and Oblique, where thumb bias and near-closure fingertip instability are most visible. The Vertical task favors Qwen-RobotManip before calibration, but after the task-level $\mathrm{SO}(3)$ correction Finger-Center is best or tied on all tasks, while Qwen-RobotManip retains a $21.5°$ residual on Vertical.

Table II reports full calibration. TATE reduces ρ_{pose} on all tasks, showing complementary effects from position anchoring and task-level $\mathrm{SO}(3)$ correction. Oblique has the largest position gain (92.1 mm to 21.0 mm), suggesting that anchors compensate for camera and reconstruction offsets. Vertical gives the clearest rotation case ($41.7°$ to $11.4°$), so the remaining Finger-Center error is largely task-level bias. Finally, η_{disp} moves toward one without falling below it, indicating controlled spread rather than trajectory collapse.

REFERENCES

[1] Open X-Embodiment Collaboration, A. O'Neill *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2024.

[2] M. J. Kim, K. Pertsch, S. Karamcheti *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," *arXiv preprint arXiv:2410.24221*, 2024.

[4] Z. Wang, B. He, K. Yu, S. Lee, R. Gao, F. Huang, and Y. Aloimonos, "Humanego: Zero-shot robot learning from minutes of human egocentric videos," *arXiv preprint arXiv:2605.24934*, 2026.

[5] H. Yuan, Z. Liang, A. Chen *et al.*, "Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models," *arXiv preprint arXiv:2606.17846*, 2026.

[6] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 12 242–12 254.

[7] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 245:1–245:17, 2017. [Online]. Available: https://mano.is.tue.mpg.de/

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.