WristCompass: Kinematic Coupling as a Learnable Visual Concept for Ego-Camera Orientation

Anonymous CVPR submission

Paper ID *****

Abstract

Recovering ego-camera orientation from manipulation video is a prerequisite for disentangling hand motion from camera motion, a key step in imitation learning from egocentric demonstrations. The obvious approach, inferring orientation from scene geometry, fails when hands occlude the frame: VGGT, a 1B-parameter scene reconstruction model, scores worse than a constant predictor on the TACO benchmark. We identify an alternative visual concept that is present precisely when scene geometry is absent: kinematic coupling dynamics, the structured physical relationship between wrist motion and camera orientation imposed by the arm-shoulder-head chain. We find that this concept is compact (4D inter-wrist features outperform 126D full hand keypoints), temporal (requiring a GRU over short windows rather than per-frame retrieval), and physically grounded (transferring zero-shot across datasets because it is rooted in anatomy rather than scene appearance). Trained only on tabletop manipulation, WristCompass transfers zero-shot to Epic Kitchens cooking video, achieving 14.3° median geodesic error and approaching the performance of a 1B-parameter scene model at 200K GRU parameters.

1. Introduction

Egocentric manipulation video is an observation from a joint space: ego motion (camera orientation in $SO(3)$) entangled with world state (hand and object configuration). Disentangling these components enables ego-motion-compensated hand trajectories, a prerequisite for learning manipulation policies from egocentric human demonstrations [5, 20, 21]. Identifying the right visual concept for each component is critical, but the camera orientation has proven surprisingly hard to recover. Scene-based approaches fail where hands occlude the frame: VGGT [17], a 1B-parameter scene reconstruction model, scores 23.98° median geodesic error on the bimanual manipulation benchmark TACO [10], worse than a constant predictor (21.22°). The scene-geometry concept simply does not exist when the scene is occluded.

We identify the concept that does exist and is sufficient for recovering the camera orientation: *kinematic coupling dynamics*. The arm-shoulder-head chain imposes a structured physical relationship between wrist motion and camera orientation. This suggests that egocentric video contains a compact, structured representation of ego motion encoded in body dynamics, a visual concept that is present precisely when scene geometry is absent.

This concept has three properties that characterize it as a structured visual concept. **Compact:** 4D inter-wrist features outperform 126D full hand keypoints (17.5° vs. 13.80°), consistent with the signal concentrating in the inter-wrist vector. **Temporal:** a nearest-neighbor lookup on per-frame features scores 26.69°, worse than constant, while a GRU over 12-frame windows achieves 13.80°; the concept is not accessible to static retrieval, only to temporal models. **Physically grounded:** trained only on TACO tabletop manipulation, WristCompass transfers zero-shot to Epic Kitchens [3]. It achieves 14.32°, approaching the performance of VGGT at 1000× lower parameter count, because the concept is grounded in anatomy rather than scene appearance.

Contributions: (1) identifying kinematic coupling dynamics as a compact, structured, temporally-grounded visual concept sufficient for camera orientation recovery in manipulation video; **(2)** WristCompass, realizing this concept from bare RGB; **(3)** a zero-shot transfer result demonstrating that the concept generalizes across datasets because it is grounded in anatomy rather than scene appearance.

2. Related Work

Ego-camera pose estimation. Recovering ego-camera orientation from video has been approached primarily through scene structure. Classical approaches such as SLAM systems [2, 15] track sparse keypoints across frames, while structure-from-motion methods [14] and learning-based reconstruction approaches [17, 18] reconstruct camera trajectories from scene geometry. These methods require sufficient scene texture and viewpoint diversity, assumptions that break

on close-up manipulation video where hands occlude most of the frame. As we show in Sec. 4, VGGT, a 1B-parameter scene reconstruction model, inherits the same failure mode despite state-of-the-art performance on standard benchmarks. IMU-based approaches can recover orientation but require dedicated hardware unavailable in existing RGB-only video archives. WristCompass targets precisely this regime, recovering orientation post-hoc from bare monocular RGB.

Ego-body, head, and hand pose estimation. Recovering body, head, or hand pose from egocentric video represents a complementary line of work, but each approach assumes information that is unavailable in close-up manipulation. EgoAllo [19] and EgoEgo [9] recover head or body pose from egocentric video but assume a wide-field view in which the body is partially visible, an assumption that fails in close-up manipulation. In the hand pose domain, WiLoR [13] recovers 3D hand meshes from monocular RGB (we use it as our keypoint extractor) and HaWoR [21] extends this to world-space trajectories via SLAM-based tracking. Both treat hand pose as output requiring camera pose as input. WristCompass reverses both directions: we recover camera orientation from wrist dynamics observable in the ego frame, rather than inferring body pose from camera motion or requiring camera pose to estimate hand trajectories.

3. Method

Input representation. Given an egocentric video, we extract 3D hand keypoints using WiLoR [13]. From each frame, we take the two wrist positions, right wrist and left wrist, and compute a 4D inter-wrist feature vector:

$$\mathbf{f}_t = \left[\|\mathbf{d}_t\|,\ \frac{\mathbf{d}_t}{\|\mathbf{d}_t\|} \right] \in \mathbb{R}^4 \tag{1}$$

where $\mathbf{d}_t = \mathbf{w}_t^L - \mathbf{w}_t^R$ is the left-minus-right wrist difference vector. The first component is inter-wrist distance; the remaining three form a unit direction vector on S^2. Features are z-score normalized using training set statistics and mean-centred per video at test time (using per-video statistics computed from the test video itself) to remove postural bias without leaking cross-video information. We use 4D inter-wrist features rather than full hand keypoints. The orientation signal concentrates in the inter-wrist vector while finger articulations encode subject-specific style that hurts cross-subject generalization, as we validate empirically in Sec. 4.

Temporal model. We process the feature sequence with a two-layer GRU:

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{f}_{t-W:t};\, \theta), \quad W = 12 \tag{2}$$

where W=12 frames (≈0.4s at 30fps), selected by sweeping over window sizes (Table 1). A linear head maps $\mathbf{h}_t \in \mathbb{R}^{128}$ to a 6D rotation representation [22], projected to $SO(3)$ via Gram-Schmidt orthogonalization. We train by minimizing geodesic loss against TACO ground-truth rotations from NOKOV optical motion capture, using Adam (lr=5×10^{-4}) with early stopping on a held-out validation split (frame-level 80/20, used solely for early stopping).

Inference. WristCompass has 200K parameters (excluding WiLoR as a shared frozen feature extractor). At inference, we run WiLoR on each frame and compute $\mathbf{f}_t$ when both wrists are detected; frames with only one detected wrist are dropped (not interpolated). On Epic Kitchens (≈50% bimanual detection), evaluation is computed over bimanual frames only; GRU windows span non-consecutive frames when detections are sparse. The full pipeline runs in real time on CPU from bare monocular RGB.

4. Experiments

4.1. Datasets and Evaluation

TACO [10] is our primary training and evaluation benchmark, providing 17 subjects performing 15 tool-action-object activities (5,210 frames) with a helmet-mounted RealSense L515 and NOKOV optical motion-capture ground truth for camera rotation. Evaluation uses Procrustes alignment, a single global rotation applied to all predictions within a video to minimize mean geodesic error, measuring relative orientation structure rather than absolute pose. We report median geodesic error, averaged over 5 random seeds on a fixed train/val split (80/20, stratified by subject), with early stopping on the held-out portion.

Epic Kitchens [3] is a large-scale egocentric cooking dataset with a chest-mounted GoPro that serves as our zero-shot transfer target. We use EPIC Fields [16] COLMAP poses as ground truth and evaluate on 36 participants, 62 videos, 16,609 frames (minimum thresholds: 70% COLMAP coverage, 10° constant baseline per video). COLMAP poses are a proxy for ground truth and may exhibit drift in textureless or heavily occluded regions.

4.2. TACO In-Distribution Results

Table 1 and Figure 2 compare WristCompass against all baselines and report architecture ablation results on TACO. WristCompass achieves **13.80°** ± **0.14°**, outperforming all baselines.

Three findings stand out. First, VGGT (23.98°) is worse than the constant predictor (21.22°), confirming that scene-based approaches fail on close-up manipulation video. Second, the 126D full-keypoint MLP (17.5°) is worse than 4D wrist geometry. A controlled comparison with an identical GRU architecture on 126D input confirms this: 4D wins on all 5 folds of subject-level cross-validation, ruling out architecture and capacity as confounds. Third, NN retrieval on the same 4D features scores 26.69°, worse than constant, show-

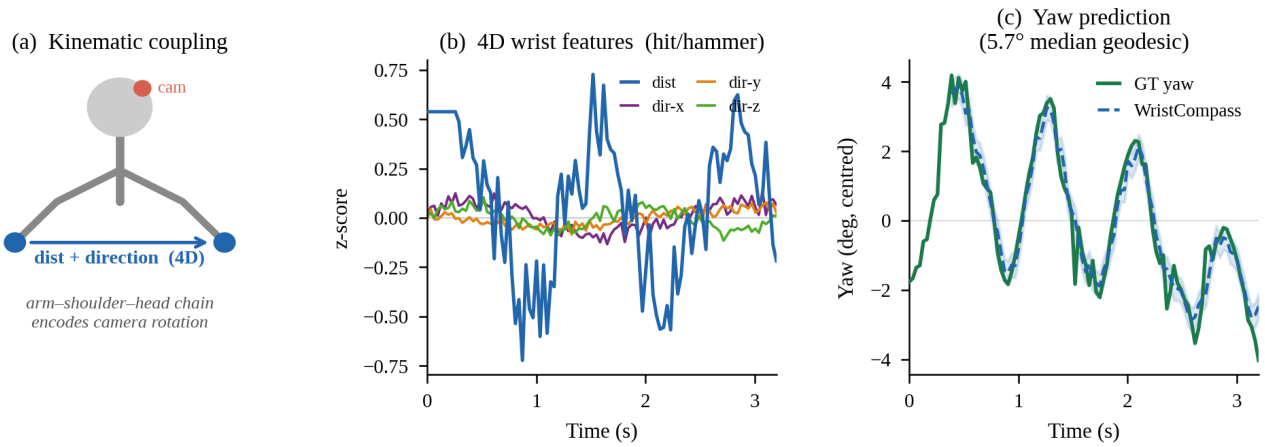

Figure 1. **WristCompass overview.** (a) Kinematic coupling: the arm-shoulder-head chain couples wrist motion dynamics to ego-camera rotation. (b) 4D inter-wrist features for a representative session: only the distance feature carries temporal variation at this timescale, direction components encode relative hand geometry. (c) WristCompass predicted yaw vs. ground-truth (GT) yaw (5.7° median geodesic). Shaded region indicates prediction uncertainty. The model tracks head rotation from wrist dynamics alone, with no scene information.

Table 1. **TACO results and ablation.** Top: baselines including scene-based (VGGT), static (NN retrieval), and full-keypoint (MLP 126D) approaches. A controlled comparison (identical GRU on 126D input) confirms 4D wins on all 5 folds of subject-level CV. Bottom: step-by-step architecture ablation, all GRU rows use 4D inter-wrist features; LR and val split were tuned jointly. All values are median geodesic error in degrees so lower is better.

Method	Geodesic (°) ↓
Constant-R baseline	21.2
VGGT 1B [17]	24.0
NN retrieval (4D, static)	26.7
MLP, 42 joints (126D)	17.5
GRU W=32 (init)	16.9
→ W=16	15.8
→ LR 5×10^{-4} + val split	13.8
→ W=12	13.7
WristCompass (5 seeds)	**13.8 ± 0.14**

ing the orientation signal is not accessible to static retrieval and requires temporal context.

Per-activity analysis. Supplementary Figure 4 shows per-activity results. WristCompass beats the constant baseline on 11/15 activities. It performs best on cyclic motions: stir/spoon (5.9°), brush/brush (6.5°). It does worst on activities where head orientation decouples from wrist dynamics: smear/glue-gun (37.3°), measure/ruler (24.3°). In these cases, the subject's wrists remain nearly stationary while the head rotates to inspect different workspace regions, breaking the kinematic coupling assumption.

4.3. Zero-Shot Transfer to Epic Kitchens

Trained exclusively on TACO, WristCompass achieves **14.32**° zero-shot on Epic Kitchens, with 33 of 36 participants beating the constant baseline. Despite different ground-truth sources (mocap vs. COLMAP), camera mountings (helmet vs. chest), and activity domains, performance is comparable to in-distribution TACO (13.80°). Figure 3 shows the per-participant breakdown. P14 (constant 66.1° → GRU 5.3°) demonstrates strong signal extraction when head movement is rich. The three failures (P10, P13, P31) are single-video participants with limited COLMAP pose quality.

Table 2. **Epic Kitchens comparison** (36 participants, zero-shot for WristCompass). Parameter count for WristCompass refers to the GRU only; WiLoR keypoint extractor is a shared prerequisite not counted in this comparison.

Method	Params	Geodesic (°) ↓
Constant-R	—	18.5
VGGT [17]	1B	12.8
WristCompass	**200K**	**14.3**

As shown in Table 2, compared to VGGT (12.83°), WristCompass (14.32°) trails by 1.5° at 1000× lower GRU parameter count. COLMAP ground truth may favor VGGT, which shares its scene-geometry assumptions. Both substantially beat the constant baseline (18.54°). The methods have complementary failure modes: VGGT wins on discrete manipulation tasks where scene structure changes predictably, while WristCompass wins on cyclic tasks where inter-wrist dynamics are more stable. On TACO, where scene features are largely occluded, WristCompass wins by over 10°.

5. Discussion and Limitations

Failure modes and scope. WristCompass works best when head movement is rich (constant baseline > 15°) and both hands are consistently visible. P14 (constant 66°, GRU 5°) shows the potential ceiling. Cyclic tasks (stir, brush) are particularly well-suited. The model cannot improve on the constant baseline when head orientation is near-static: on

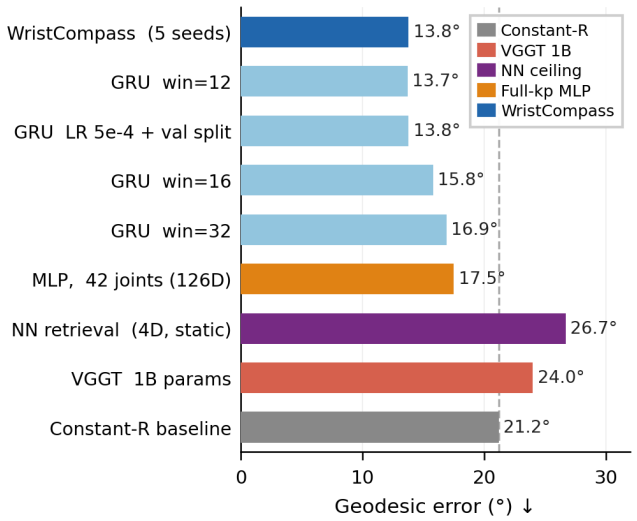

Figure 2. **TACO results and ablation.** NN retrieval (26.7°) is worse than the constant baseline (21.2°); the orientation signal is not accessible to static retrieval. WristCompass (13.8°) outperforms VGGT 1B (24.0°) at 200K GRU parameters.

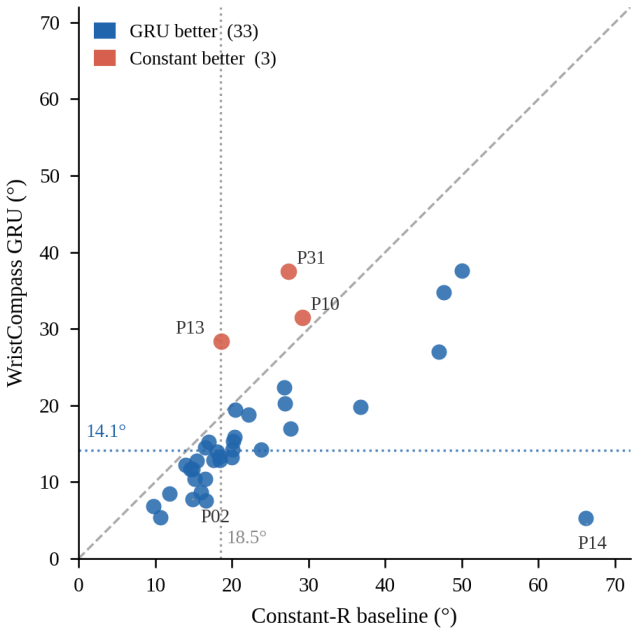

Figure 3. **Epic Kitchens zero-shot (36 participants).** Points below diagonal: WristCompass beats constant baseline (33/36, blue). Points above: failures (3/36, red). P14 (constant 66°, GRU 5°) is a striking outlier.

close-up manipulation datasets with limited head movement such as HOT3D (5.1°) and ARCTIC (5.5°), there is insufficient rotational signal to recover [1, 4]. Activity-level failures (smear 37.3°, measure 24.3°) occur when wrist position is constrained independently of head orientation, breaking kinematic coupling. Both of the datasets we evaluate, TACO and Epic Kitchens, involve standing manipulation, transfer to seated or full-body activities remains untested.

Kinematic coupling as a physical visual concept. Most concept learning methods, including β-VAE [7], Slot Attention [12], and Concept Bottleneck Models [8], discover structure from co-occurrence statistics in training data, so their generalization is bounded by training distribution diversity. Inductive-bias approaches impose architectural priors such as competition and capacity limits [11], but these are still learned constraints rather than physical ones. Kinematic coupling occupies a third position: its 4D inter-wrist representation is not learned from data but dictated by biomechanics—the physical structure of the human arm-shoulder-head chain. This makes it *physically grounded*: it generalizes across environments and activity domains because human anatomy is shared, not because the training distribution covers the test distribution. It is also *intrinsically temporal*—the NN retrieval ablation shows the concept is invisible in any single frame and only emerges over time. The zero-shot transfer from TACO to Epic Kitchens provides direct empirical evidence for both properties.

Limitations. Several limitations of the current approach point toward future directions. WristCompass recovers relative camera orientation, which is how the camera rotates over time, rather than absolute pose. This is sufficient for downstream tasks such as ego-motion-compensated hand trajectory extraction, where action representations depend on frame-to-frame rotation rather than global heading, but absolute calibration remains an open problem. Single-hand frames (≈48% of Epic Kitchens frames) currently fall back to a constant predictor; improving single-hand fallback is a clear direction for future work. Epic Kitchens evaluation uses raw WiLoR-mini keypoints without smoothing (≈50% bimanual detection); Kalman–RTS smoothing improves TACO (≈100% detection) but degrades Epic Kitchens by replacing temporal signal with near-constant interpolations, suggesting that detection-aware smoothing strategies warrant further investigation.

6. Conclusion

We identify kinematic coupling dynamics, the temporal relationship between bimanual wrist motion and ego-camera orientation imposed by the arm-shoulder-head chain, as a compact, physically-grounded visual concept for recovering the camera orientation of manipulation video. WristCompass realizes this concept with 200K GRU parameters from bare monocular RGB, outperforming a 1B-parameter scene model on close-up manipulation and generalizing zero-shot to kitchen video. Future directions include depth integration, explicit ego-world disentanglement [6], and downstream robot policy learning.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Ba-

sol, Richard A. Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *ArXiv*, abs/2406.09598, 2024. 4

[2] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 2021. 1

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *ECCV*, 2018. 1, 2

[4] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2022. 4

[5] Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos. *ArXiv*, abs/2601.01050, 2026. 1

[6] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. 4

[7] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 4

[8] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020. 4

[9] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023. 2

[10] Yun Liu, Haolin Yang, Xu Xu, Mingsheng Ding, Weicheng Li, Yuxin Li, Ziyu Liu, Jianyu Luo, Jing Cheng, and Li Cheng. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *CVPR*, 2024. 1, 2

[11] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019. 4

[12] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020. 4

[13] Rolandos Alexandros Potamias, Jinglei Shu, German Barquero, Cristina Palmero, Sergio Escalera, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *ECCV*, 2024. 2

[14] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[15] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Neural Information Processing Systems*, 2021. 1

[16] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Sheratt, Mike Brookes, Roberto Cipolla, Spyridon Leonardos, and Dima Damen. EPIC Fields: Marrying 3D geometry and video understanding. *NeurIPS*, 2023. 2

[17] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded deep structure from motion. In *CVPR*, 2025. 1, 3

[18] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2023. 1

[19] Jiaman Ye, Yiye Ye, Manolis Savva, Angel X. Chang, and Li Yi. EgoAllo: Egocentric human motion estimation via explicit alignment with a grounded world coordinate system. In *ECCV*, 2024. 2

[20] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dynhamr: Recovering 4d interacting hand motion from a dynamic camera. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27716–27726, 2024. 1

[21] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1815, 2025. 1, 2

[22] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 2

Supplementary Material

A. Controlled 4D vs. 126D comparison. Table 3 reports a head-to-head comparison using identical GRU architectures (W=12, hidden=128, 2 layers) on 4D inter-wrist vs. 126D full-keypoint input, evaluated via 5-fold subject-level cross-validation. 4D outperforms 126D on all 5 folds, ruling out architecture and capacity as confounds. The 126D model overfits to subject-specific finger articulations that do not transfer to held-out subjects.

Table 3. **4D vs. 126D GRU** (5-fold subject-level CV, 5 seeds per fold). 4D wins on every fold.

Fold	4D (°)	126D (°)	Δ
0	4.67	6.81	−2.14
1	6.44	7.98	−1.54
2	3.68	7.82	−4.14
3	8.56	12.14	−3.58
4	2.68	6.66	−3.98
Mean	**4.77 ± 0.24**	**7.84 ± 0.25**	**−3.07**

B. Per-axis error decomposition. On TACO, WristCompass achieves per-axis median errors of yaw 8.44°, pitch 5.28°, roll 3.87°. Yaw (left-right head rotation) carries the most error, consistent with the inter-wrist vector being most informative about lateral head movement. Pitch and roll are better constrained by the arm-shoulder-head kinematic chain.

C. All-frames Epic Kitchens evaluation. The main paper reports 14.32° on bimanual frames only (≈52% of frames). For a deployment-realistic estimate, we blend GRU predictions on bimanual frames with the constant-R baseline on single-hand/no-hand frames:

Table 4. **Epic Kitchens blended evaluation** (62 videos, 16,609 total frames).

Metric	Geodesic (°) ↓
Constant-R (all frames)	18.54
GRU-only (bimanual, 52%)	14.32
Blended median (per-video)	16.56

The 2.0° gap between GRU-only and blended reflects the 48% of frames where only one or no hands are detected. Improving single-hand fallback is a clear direction for future work.

D. Per-activity breakdown. Figure 4 shows per-activity TACO results.

E. Failure case: kinematic decoupling. Figure 5 shows a representative failure (empty/bowl/plate, session 8, 21.3° geodesic). Between t=4–6s, the subject's head rotates ≈14° in yaw while wrists remain stationary — the subject looks away from their hands to inspect the workspace. WristCompass cannot track this head rotation because the kinematic coupling between wrist motion and head orientation is broken. This failure mode is systematic: it accounts for the worst per-activity results (smear/glue-gun 37.3°, measure/ruler 24.3°) where wrist position is constrained independently of gaze direction.

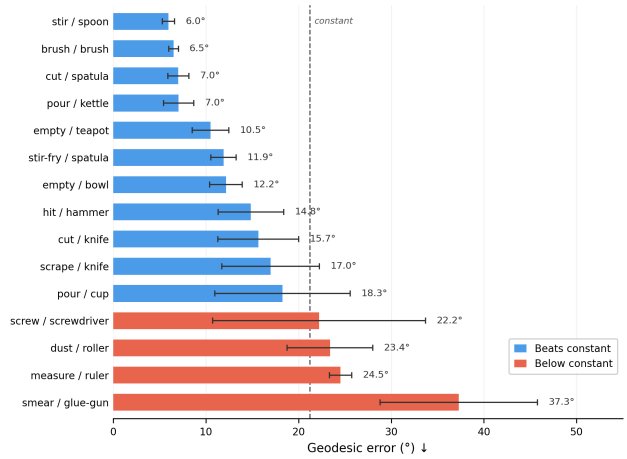

Figure 4. **Per-activity TACO results** (5 seeds, IQR error bars). Blue: beats constant (11/15). Red: below constant (4/15). Cyclic motions (stir, brush) are easiest; decoupled motions (smear, measure) are hardest.

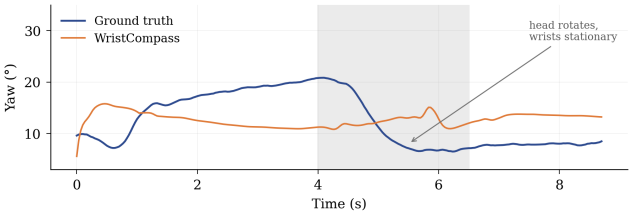

Figure 5. **Failure case: kinematic decoupling.** GT yaw (blue) drops 14° between t=4–6s while WristCompass prediction (orange) remains flat. The subject's head rotates independently of their stationary wrists.