V-JEPA HAND: SIMPLE, ACCURATE AND FAST 3D HAND AND CAMERA TRAJECTORY RECOVERY FROM EGOCENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

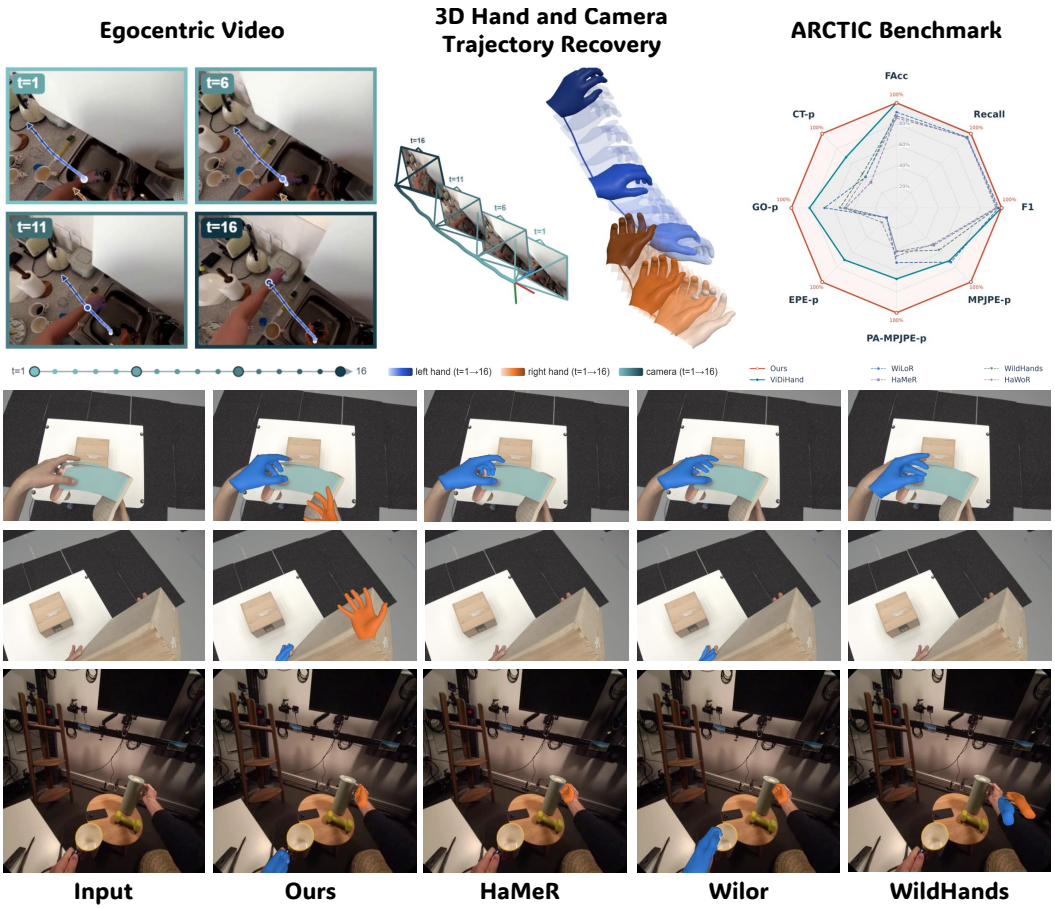


Figure 1: **V-JEPA Hand: joint hand and camera trajectory recovery from egocentric video. Top left and centre:** From a video clip, V-JEPA Hand jointly recovers bilateral 3D hand motion and camera motion in a single forward pass. Blue, orange and teal denote the left hand, right hand and camera, respectively; colour intensity indicates time. **Top right:** ARCTIC detection and coverage-aware reconstruction results across eight metrics, normalized to our result (100%); larger is better. **Bottom:** Qualitative comparisons with HaMeR, WiLoR and WildHands, illustrating differences in hand pose accuracy, missed hands and spurious predictions.

ABSTRACT

Egocentric video supports diverse applications, including robot learning, augmented reality, and human activity understanding. Recovering metric 3D hand trajectories from such video requires estimating both hand motion and camera egomotion to express the hands in a consistent reference frame. Many existing pipelines recover hand and camera motion in separate stages, combining hand detection, hand reconstruction, camera tracking, and trajectory refinement. We present V-JEPA Hand, a simple feed-forward video model that jointly reconstructs bilateral hand motion and camera motion from egocentric clips. It uses a self-supervised V-JEPA 2.1 encoder to extract spatio-temporal features from full-frame video and a query-based predictor to decode these features into per-frame

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

hand poses, hand presence, and camera motion. We use the predicted camera motion to express the hand trajectories in a common world frame. The model accepts known focal lengths or predicts them when unavailable. It reconstructs each clip in a single forward pass. We train both components end to end in a single stage on a mixture of diverse pseudo-labelled videos and high-precision 3D annotations. A single model achieves the lowest coverage-aware mean per-joint position error (MPJPE) among comparable baselines on ARCTIC, HOT3D, and EgoStandard. It also yields lower world-frame trajectory errors than the compared SLAM-based pipelines. Ablations support the benefits of video pretraining, temporal context, and mixed supervision. V-JEPA Hand achieves a forward-pass throughput of 91.3 FPS on a single RTX A6000. We will publicly release our code, model weights, and training scripts.

1 INTRODUCTION

Egocentric video provides a natural record of human manipulation and supports a wide range of applications. Recent VLA approaches use 3D hand motion from egocentric demonstrations for pretraining and robot-policy adaptation (Kareer et al., 2025a;b; Luo et al., 2025; Yang et al., 2025; Bi et al., 2026; Zheng et al., 2026; Li et al., 2025). Hand motion recovery also supports AR/VR interaction (Han et al., 2022), first-person activity understanding (Kwon et al., 2021), animation and digital humans (Romero et al., 2017), and wrist-kinematics assessment (Smeragliuolo et al., 2016). These applications benefit from accurate, temporally coherent 3D hand motion. Apparent hand motion in egocentric video can result from movement of the hand, the camera, or both. Recovering metric hand trajectories in a consistent reference frame therefore requires estimating both hands' positions, orientations, finger articulation, and presence, together with camera motion.

Single-image models such as HaMeR and WiLoR (Pavlakos et al., 2024; Potamias et al., 2025) reconstruct hands from detected crops. Applied frame by frame, they cannot use neighboring frames to improve reconstruction under occlusion, motion blur, or intermittent visibility. Temporal methods add cross-frame modeling (Zhang et al., 2024), while world-frame approaches combine hand reconstruction with camera tracking, motion infilling, or optimization (Zhang et al., 2025; Chen et al., 2025). More recently, ViDiHand (Wang et al., 2026) uses pretrained video diffusion features for detector-free, full-frame hand reconstruction. Video provides temporal cues for hand reconstruction and scene context for camera motion. We further investigate *how effectively a self-supervised video encoder can support joint hand–camera recovery through end-to-end, single-stage adaptation*.

Our central finding is that *a simple video encoder–predictor architecture can achieve accurate and fast joint hand–camera reconstruction*. We instantiate this design as V-JEPA Hand, pairing the self-supervised V-JEPA 2.1 encoder (Mur-Labadia et al., 2026) with a query-based predictor and fine-tuning both end to end. Cross-frame information is incorporated within the visual encoder, and hand and camera states are decoded from the same spatio-temporal features. The predictor jointly estimates per-frame bilateral hand poses, hand presence, and camera motion; it also accepts known clip-level focal lengths or predicts them when unavailable. The entire clip is reconstructed in one forward pass without a separate hand detector, hand crops, SLAM, or test time optimization. The model thus retains a simple feedforward structure rather than assembling estimates from independent hand and camera systems.

Training such a model requires accurate 3D annotations across diverse capture conditions. ARCTIC (Fan et al., 2023) and HOT3D (Banerjee et al., 2024) provide high-quality hand-motion annotations using dedicated motion-capture systems, while H2O (Kwon et al., 2021) uses calibrated multi-view RGB-D recordings to annotate bimanual interactions. These resources provide valuable geometric supervision, but their hardware and calibration requirements make collection costly to extend across everyday environments. In contrast, Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2018) capture diverse, unscripted activities without dense metric hand-motion labels. Reconstruction offers a complementary source of supervision: HaWoR (Zhang et al., 2025) recovers world-space hand trajectories, while the VITRA annotation pipeline (Li et al., 2025) combines hand reconstruction with metric camera tracking to generate frame-aligned hand-action labels at scale. Such automatically recovered trajectories are useful but imperfect pseudo-labels. We therefore train in a single stage on pseudo-labelled Ego4D and our in-house recordings, mixed

with a comparatively small pool of precisely annotated videos from ARCTIC and HOT3D. The pseudo-labelled videos broaden scene and action coverage, while the high-quality annotations provide reliable geometric supervision for transfer beyond the capture environments.

We evaluate the same trained model on ARCTIC (Fan et al., 2023), HOT3D (Banerjee et al., 2024), and EgoStandard (Lightwheel AI, 2025), whose commercial annotation pipeline combines hardware-assisted capture with multi-stage optimization. EgoStandard is excluded from our reconstruction training. Across all three benchmarks, V-JEPA Hand combines high hand-detection coverage with the lowest coverage-aware MPJPE among the previous baselines. It also yields lower world-frame trajectory errors than the compared SLAM-based pipelines. Data ablations show that mixed supervision improves reconstruction on unseen videos, while component ablations support the value of pretrained features and temporal context. V-JEPA Hand achieves forward-pass throughput of 91.3 FPS on a single RTX A6000.

We summarize our contributions as follows:

- A simple yet effective feed-forward framework that jointly recovers bilateral 3D hand motion, hand presence, camera egomotion, and clip-level focal lengths from a shared video representation, without separate hand detection, SLAM, or test-time optimization.
- A single-stage training recipe combining diverse pseudo-labelled egocentric video with high-precision annotations, with demonstrated benefits for reconstruction on unseen data.
- A systematic evaluation of detection coverage, hand pose, world-frame trajectories, cross-domain generalization, and inference efficiency, together with ablations of key training and modeling choices.

2 Related Work

Monocular 3D hand reconstruction. Estimating 3D hand pose and shape from monocular images has advanced rapidly with parametric models such as MANO (Romero et al., 2017). Most existing methods detect then regress, cropping each hand with a 2D detector before predicting mesh or kinematic parameters (Pavlakos et al., 2024; Potamias et al., 2025; Dong et al., 2024; Chen et al., 2022; Moon, 2023; Park et al., 2022). In egocentric videos, however, this pipeline struggles with frequent occlusions, motion blur, and boundary truncations, while discarding global context and inter-hand spatial relations. To bypass cropping, detector-free alternatives directly regress hand poses from full frames (Prakash et al., 2024) or leverage video diffusion features (Wang et al., 2026), though diffusion models incur substantial computational overhead. In contrast, V-JEPA Hand recovers both hands and their presence probabilities in a single forward pass via a lightweight query-based decoder, avoiding separate detectors, crops, or iterative diffusion.

Temporal and world-coordinate hand motion recovery. Applying single-frame estimators independently across video frames yields temporally jittery trajectories. While temporal networks and transformers help smooth camera-relative predictions (Yang et al., 2020; Zhang et al., 2024), they do not resolve camera egomotion. To express manipulation in a consistent world frame, current methods decouple hand estimation from camera tracking by combining per-frame regressors, external visual SLAM such as DROID-SLAM (Teed & Deng, 2021), and motion infilling or test-time trajectory optimization (Zhang et al., 2025; Chen et al., 2025; Fan et al., 2024), mirroring pipelines in world-grounded body capture (Ye et al., 2023; Shin et al., 2024). However, SLAM frequently drifts under aggressive head motion, and multi-stage optimization introduces latency and cascading errors. Instead, V-JEPA Hand unifies camera-space bilateral hand motion, hand presence, camera egomotion, and camera intrinsics into a single feedforward prediction, avoiding external SLAM or test-time optimization while achieving superior world-frame accuracy.

Self-supervised video representations and foundation models. Self-supervised video pretraining provides powerful motion and scene priors. While masked autoencoders such as MAE (He et al., 2022), VideoMAE (Tong et al., 2022), and VideoMAE V2 (Wang et al., 2023) reconstruct masked pixels, pixel-level objectives often overemphasize high-frequency texture details over geometric and physical dynamics. The Joint-Embedding Predictive Architecture (JEPA) family (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025; Mur-Labadia et al., 2026) addresses this by

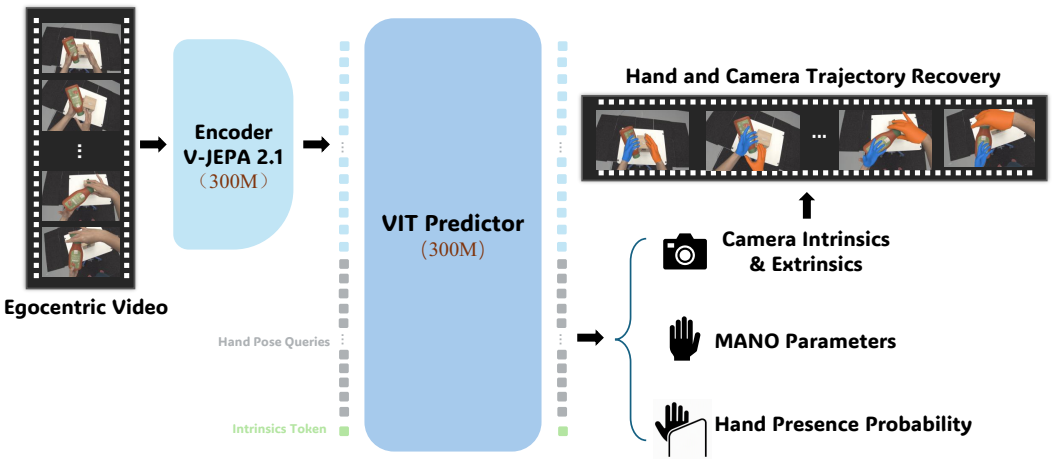


Figure 2: **Overview of V-JEPA Hand.** A V-JEPA 2.1 encoder extracts spatio-temporal features from an egocentric video clip. A ViT predictor jointly processes these features, one learnable hand-pose query per frame, and an intrinsics token. The output queries predict bilateral MANO pose, hand presence probability, and camera motion. The entire clip is reconstructed in a single forward pass; blue and orange meshes denote the reconstructed left and right hands, respectively.

predicting representations in latent space; notably, V-JEPA 2.1 (Mur-Labadia et al., 2026) learns rich spatio-temporal features and dense motion priors. Although generative video foundation models (Blattmann et al., 2023) also encode temporal priors (Wang et al., 2026), their generative nature requires heavy computation. In comparison, V-JEPA 2.1 offers a discriminative representation that supports high-throughput (91.3 FPS) inference and outperforms VideoMAE, DINOv3 (Siméoni et al., 2025), and Perception Encoder (Bolya et al., 2025) on both camera- and world-frame metrics.

3 Method

As illustrated in Figure 2, V-JEPA Hand reads a raw video clip and emits, in a single forward pass, the complete bilateral hand trajectory of that clip together with the camera motion and focal lengths. The model has two components: a self-supervised video encoder that turns the clip into a spatio-temporal token grid (Section 3.2), and a query-based trajectory predictor that decodes one output token per frame into the per-frame state defined below. Section 3.3 describes our single-stage training with multi-source supervision, combining large-scale pseudo-labelled video with high-quality data under a shared training objective.

3.1 Problem formulation

The input is a clip of T RGB frames $\mathbf{V} = (\mathbf{I}_1, \ldots, \mathbf{I}_T)$, $\mathbf{I}_t \in \mathbb{R}^{3\times H\times W}$, recorded by a freely moving egocentric camera. Here $t \in \{1, \ldots, T\}$ indexes frames, and H and W are the image height and width in pixels, respectively. We write $\mathcal{C}_t$ for the camera coordinate frame of frame t and index the two hands by $s \in \{\mathrm{L}, \mathrm{R}\}$, where L and R denote the camera wearer's left and right hand. Hats denote model predictions.

Model output. The model returns a per-frame hand–camera trajectory together with clip-level camera focal lengths,

$$\hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_T) \in \mathbb{R}^{T\times 119}, \qquad \hat{\mathbf{y}}_t = \big(\underbrace{\hat{\mathbf{x}}_t^{\mathrm{L}}, \hat{\mathbf{x}}_t^{\mathrm{R}}}_{110}, \underbrace{\hat{\mathbf{g}}_t}_{9}\big), \qquad \hat{\mathbf{k}} \in \mathbb{R}^2, \tag{1}$$

where $\hat{\mathbf{Y}}$ stacks the per-frame state vectors $\hat{\mathbf{y}}_t$. Each $\hat{\mathbf{x}}_t^s \in \mathbb{R}^{55}$ is the state of hand s, $\hat{\mathbf{g}}_t \in \mathbb{R}^9$ is the camera motion at frame t, and $\hat{\mathbf{k}}$ estimates the clip-level focal-length vector $\mathbf{k} = (f_x, f_y)$, where

f_x and f_y are the horizontal and vertical focal lengths in pixels at the input resolution. Our notation retains only these two varying intrinsic parameters. The per-frame states are decoded from the T query tokens, one row $\hat{\mathbf{y}}_t$ per token. The model accepts known focal lengths or predicts them from a dedicated token when unavailable (Section 3.2). The hand and camera motion states are defined next.

Hand state. For frame t and hand s the model predicts

$$\hat{\mathbf{x}}_t^s = \big(\, \hat{\mathbf{d}}_t^s,\ \hat{\mathbf{r}}_t^s,\ \hat{\boldsymbol{\theta}}_t^s,\ \hat{c}_t^s \,\big) \in \mathbb{R}^{55}, \tag{2}$$

whose four blocks are: the wrist translation $\hat{\mathbf{d}}_t^s \in \mathbb{R}^3$, i.e. the position of the hand root in meters expressed in $\mathcal{C}_t$; the wrist orientation $\hat{\mathbf{r}}_t^s \in \mathbb{R}^6$, a prediction of the first two rows of the rotation matrix $\mathbf{R}_t^s \in \mathbb{R}^{3\times3}$ that maps the canonical MANO hand frame to $\mathcal{C}_t$; the articulation $\hat{\boldsymbol{\theta}}_t^s \in \mathbb{R}^{45}$, the Euler angles of the 15 MANO finger joints, each relative to its parent joint; and the presence score $\hat{c}_t^s \in \mathbb{R}$, a logit that is high when hand s is actually in frame t. The two hands together therefore account for the 110 **hand dimensions** of $\hat{\mathbf{y}}_t$. We do not predict the MANO shape coefficient vector $\boldsymbol{\beta}$, and all hands are evaluated at the neutral shape $\boldsymbol{\beta} = \mathbf{0}$.

Camera state. The camera state encodes the rigid transform $\mathbf{T}_{t\to1} \in SE(3)$ from the current camera frame $\mathcal{C}_t$ to the first-frame anchor $\mathcal{C}_1$, where $SE(3)$ is the group of 3D rigid transformations:

$$\hat{\mathbf{g}}_t = \big(\, \hat{\mathbf{r}}_t^{\text{cam}},\ \hat{\mathbf{u}}_t \,\big) \in \mathbb{R}^9, \qquad \mathbf{u}_t = \operatorname{sgn}(\mathbf{a}_t) \odot \log\big(1 + |\mathbf{a}_t|\big), \tag{3}$$

where the superscript cam distinguishes camera rotation from hand rotation. The vector $\hat{\mathbf{r}}_t^{\text{cam}} \in \mathbb{R}^6$ predicts the first two rows of the camera rotation matrix $\mathbf{R}_t^{\text{cam}} \in \mathbb{R}^{3\times3}$, which maps vectors from $\mathcal{C}_t$ to $\mathcal{C}_1$. The vector $\mathbf{a}_t \in \mathbb{R}^3$ is the translation component of $\mathbf{T}_{t\to1}$, expressed in $\mathcal{C}_1$ in metres. We represent this translation by its signed-log encoding $\mathbf{u}_t \in \mathbb{R}^3$, and $\hat{\mathbf{u}}_t$ is the model's prediction of that encoding, rather than the translation in metres. Here sgn is the sign function, $\odot$ denotes element-wise multiplication, and the absolute value and natural logarithm are also applied element-wise.

3.2 Architecture

Video encoder. We initialize the video encoder with self-supervised pretrained V-JEPA 2.1 ViT-L weights (Mur-Labadia et al., 2026) and jointly optimize all encoder and predictor parameters on the mixed training data. The encoder cuts the video clip into non-overlapping space–time patches, each spanning 2 consecutive frames and a 16×16 pixel square, and embeds every patch into one token of dimension d, so the clip becomes a sequence

$$\mathbf{Z} \in \mathbb{R}^{N\times d}, \qquad N = \underbrace{\tfrac{T}{2}}_{\text{time}} \cdot \underbrace{\tfrac{H}{16} \cdot \tfrac{W}{16}}_{\text{space}}. \tag{4}$$

Here $\mathbf{Z}$ is the encoder token sequence, N is the number of space–time patches, and d is the token feature dimension. Each token is located by a rotary encoding of its (τ, h, w) patch coordinate, where τ, h, and w index the temporal, vertical, and horizontal positions in the patch grid, respectively.

Query-based trajectory predictor. The predictor is a ViT that operates on the concatenation of the encoder grid, T learnable *query tokens* $\mathbf{Q} \in \mathbb{R}^{T\times d}$, one per output frame, and a single intrinsics token $\mathbf{q}_k \in \mathbb{R}^d$, whose subscript k denotes its association with camera intrinsics, under joint self-attention:

$$\big[\tilde{\mathbf{q}}_k;\ \tilde{\mathbf{Z}};\ \tilde{\mathbf{Q}}\big] = \mathrm{ViT}\,\big([\mathbf{q}_k;\ \mathbf{Z};\ \mathbf{Q}]\big). \tag{5}$$

Here $[\cdot;\cdot]$ denotes concatenation along the token dimension, and ViT is the predictor's vision transformer. Tildes denote the corresponding output tokens: $\tilde{\mathbf{q}}_k \in \mathbb{R}^d$ for intrinsics, $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times d}$ for the video features, and $\tilde{\mathbf{Q}} \in \mathbb{R}^{T\times d}$ for the per-frame queries. Each query is temporally aligned with its corresponding video frame through rotary positional encoding. A shared MLP maps each output query to the 119-dimensional hand–camera state for that frame. During training, the intrinsics token is set to an embedding of the ground-truth intrinsics with probability 0.5 and a learned token otherwise. A linear head reads the output token $\tilde{\mathbf{q}}_k$ to predict $\hat{\mathbf{k}}$, the two focal-length components, with supervision applied only to clips whose intrinsics were withheld.

3.3 TRAINING

Single-stage multi-source training. We train V-JEPA Hand end to end in a single stage using four data sources: Ego4D (Grauman et al., 2022), our in-house egocentric recordings, ARCTIC (Fan et al., 2023), and HOT3D (Banerjee et al., 2024). Ego4D and our in-house recordings provide a large-scale pool of pseudo-labelled video spanning diverse scenes, activities, and camera wearers. We mix this noisy pseudo-labelled data with high-quality data from the training splits of ARCTIC and HOT3D. Throughout training, each clip is sampled from the pseudo-labelled pool, ARCTIC, or HOT3D with probabilities 0.35, 0.45, and 0.20, respectively.

Training objective. We supervise hand parameters, 3D joint positions, 2D projections, temporal smoothness, hand presence, and camera estimates with the total training loss

$$\mathcal{L} = \underbrace{\lambda_d \mathcal{L}_d + \lambda_r \mathcal{L}_r + \lambda_\theta \mathcal{L}_\theta}_{\text{parameters}} + \underbrace{\lambda_j \mathcal{L}_j}_{\text{3D}} + \underbrace{\lambda_\pi \mathcal{L}_\pi}_{\text{2D}} + \underbrace{\lambda_a \mathcal{L}_a}_{\text{jitter}} + \underbrace{\lambda_c \mathcal{L}_c + \lambda_g \mathcal{L}_g + \lambda_k \mathcal{L}_k}_{\text{presence and camera}}. \tag{6}$$

Each λ coefficient is a nonnegative scalar weight for the loss term with the same subscript. The terms $\mathcal{L}_d$, $\mathcal{L}_r$, and $\mathcal{L}_\theta$ supervise wrist position, wrist orientation, and finger articulation, respectively. $\mathcal{L}_j$ penalizes errors in the 21 camera-space 3D joints obtained through differentiable MANO forward kinematics, while $\mathcal{L}_\pi$ penalizes errors in their image-plane projections. $\mathcal{L}_a$ penalizes the acceleration of the predicted 3D joints across consecutive frames, which suppresses jitter without penalizing smooth motion. $\mathcal{L}_c$ supervises whether each hand is present in each frame. $\mathcal{L}_g$ supervises camera rotation and signed-log translation relative to the first frame, and $\mathcal{L}_k$ supervises clip-level focal lengths when they are withheld from the input; auxiliary implementation details are given in Appendix D.

4 EXPERIMENTS

We evaluate V-JEPA Hand on egocentric two-hand reconstruction. We first describe the datasets and evaluation protocol (Section 4.1), then compare with prior work (Section 4.2). Our ablations test the training data, the loss terms, encoder pretraining and temporal context, and encoder choice (Section 4.3). We also report model size and inference cost (Section 4.4), and show qualitative results (Section 4.5).

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on three egocentric hand benchmarks: **ARCTIC** tests reconstruction under severe occlusion during two-hand object manipulation, while **HOT3D** tests robustness to motion blur from rapid head and hand movements. We additionally evaluate on **EgoStandard**, a commercially annotated egocentric dataset spanning diverse scenes. Its 3D motion annotations are produced by a pipeline that combines hardware-assisted capture with multi-stage optimization. We exclude EgoStandard entirely from training and use it to assess generalization to unseen egocentric data. The EgoStandard hand comparison and training-data ablation use the expanded 60.05-minute evaluation. The HOT3D hand comparison uses the HaWoR-aligned split with $1{,}101$ segments and $88{,}432$ scored frames after annotation filtering. Table 2 uses the same evaluation splits and V-JEPA Hand checkpoint, with all world methods rescored on their common true-positive support.

Metrics. Metrics follow the coverage-aware evaluation implementation of Wang et al. (2026), which accounts for missed hands. Metric definitions, table abbreviations, and evaluation details are provided in Appendix F.

4.2 COMPARISON WITH PRIOR WORK

Table 1 compares V-JEPA Hand with WiLoR (Potamias et al., 2025), HaMeR (Pavlakos et al., 2024), WildHands (Prakash et al., 2024), and HaWoR (Zhang et al., 2025) on all three datasets. All five methods are evaluated on the same splits, and V-JEPA Hand uses the same checkpoint across all three datasets. We also include published ViDiHand (Wang et al., 2026) results: its ARCTIC scores are directly comparable, while its HOT3D scores use a different split and serve only as a reference.

Table 1: **Comparison with prior work.** WiLoR, HaMeR, and WildHands are single-frame methods; HaWoR, ViDiHand, and ours are multi-frame methods. We follow ViDiHand's evaluation protocol, except that pose metrics without the -p suffix are computed on each method's own intersection of predictions and ground truth (true positives). ViDiHand results are taken from its paper (Wang et al., 2026); its HOT3D row (gray) uses a different split and is shown for reference only, without direct comparison or ranking.

	Detection			3D Pose				Orient. & Position						Temporal
Method	Acc↑	Rec↑	F1↑	MP-p↓	MP↓	PA-p↓	PA↓	EP-p↓	EP↓	GO-p↓	GO↓	CT-p↓	CT↓	Jit↓
ARCTIC														
WiLoR	0.909	0.949	0.971	21.11	15.52	12.74	6.70	65.56	28.84	17.68	11.74	0.077	0.050	25.97
HaMeR	0.873	0.943	0.958	30.78	25.10	16.02	9.44	64.85	28.45	25.14	18.67	0.093	0.063	14.78
WildHands	0.874	0.944	0.958	27.07	21.29	15.99	9.55	65.44	29.58	22.54	16.05	0.070	0.039	12.28
HaWoR	0.853	0.953	0.945	31.52	26.94	14.48	9.07	44.32	16.24	24.38	18.82	0.076	0.050	19.37
ViDiHand	**0.997**	**0.999**	**0.999**	21.67	-	9.82	-	12.41	-	14.64	-	0.047	-	**3.18**
Ours	**0.997**	0.998	**0.999**	**14.92**	**14.72**	**6.80**	**6.59**	**9.62**	**9.51**	**11.97**	**11.72**	**0.030**	**0.029**	6.23
HOT3D														
WiLoR	0.810	0.879	0.922	39.38	26.09	23.10	7.57	79.90	22.11	32.51	16.49	0.103	0.048	17.19
HaMeR	0.719	0.880	0.886	67.09	57.96	25.09	10.26	70.66	15.40	50.01	36.35	0.106	0.055	21.09
WildHands	0.425	0.600	0.609	80.23	42.99	61.89	12.40	275.39	37.14	84.35	38.82	0.250	0.076	25.30
HaWoR	0.733	0.895	0.875	28.79	16.71	20.68	7.64	66.60	8.16	25.01	10.50	0.086	0.039	13.41
ViDiHand	0.948	0.974	0.983	21.51	-	11.38	-	14.95	-	15.83	-	0.040	-	3.74
Ours	**0.969**	**0.978**	**0.980**	**14.96**	**12.47**	**8.38**	**5.74**	**18.45**	**7.08**	**10.44**	**6.94**	**0.029**	**0.017**	**6.10**
EgoStandard														
WiLoR	0.949	0.971	0.982	35.19	32.16	11.11	7.35	37.09	12.74	25.88	22.13	0.089	0.080	15.53
HaMeR	0.986	0.992	0.994	37.37	36.60	**8.27**	**7.26**	18.06	11.45	26.37	25.37	0.093	0.090	18.35
WildHands	0.514	0.658	0.659	69.22	33.83	55.04	12.29	325.25	53.66	74.03	31.84	0.214	0.099	20.10
HaWoR	0.883	0.901	0.937	33.38	23.36	20.11	8.62	91.58	**8.00**	26.58	12.81	0.089	0.058	13.95
Ours	**0.988**	**0.993**	**0.996**	**17.90**	**17.08**	8.42	7.52	**13.67**	8.02	**11.89**	**10.96**	**0.039**	**0.037**	**4.97**

Table 2: **World-coordinate reconstruction.** All methods are evaluated on the same hand-frames, using the intersection of true positives across the four compared methods. W-MPJPE and WA-MPJPE are in mm; RTE is in %. Lower is better.

	ARCTIC			HOT3D			EgoStandard		
Method	W-MP↓	WA-MP↓	RTE↓	W-MP↓	WA-MP↓	RTE↓	W-MP↓	WA-MP↓	RTE↓
WiLoR+SLAM	63.36	34.13	29.78	74.79	28.42	40.21	90.61	41.53	26.26
HaMeR+SLAM	69.05	33.78	28.21	105.13	35.08	47.31	94.82	42.72	27.14
HaWoR	106.05	55.38	48.92	63.08	27.60	51.02	90.23	45.01	35.92
Ours	**35.41**	**19.58**	**17.68**	**38.67**	**18.47**	**31.07**	**67.05**	**33.48**	**22.48**

From occlusion and blur to unseen scenes. Table 1 shows that V-JEPA Hand outperforms prior methods on the vast majority of comparable metrics across all three benchmarks. The improvements span hand detection, 3D pose, and camera-space placement, rather than being confined to a single aspect of reconstruction. On ARCTIC, this translates into accurate hand recovery despite severe occlusion. Crucially, the advantage extends to EgoStandard's diverse scenes without any training on that dataset: V-JEPA Hand leads in most metrics on expanded EgoStandard, including both coverage-aware and TP-only MPJPE. Together, these results show that the model's accuracy is not restricted to the benchmarks represented in its training data. A single video-based model supports both precise reconstruction on ARCTIC and HOT3D and transfer to unseen egocentric scenes. On the HaWoR-aligned HOT3D split, V-JEPA Hand leads all fourteen metrics among the four local baseline reproductions.

World-frame reconstruction through unified prediction. Table 2 shows that the simpler architecture also delivers competitive world-frame trajectories, where hand reconstruction must be combined with camera motion. V-JEPA Hand leads all three world-frame metrics on all three datasets. Because this comparison uses shared true-positive hand-frames, the gains reflect trajectory accuracy rather than differences in detection coverage. Unlike the SLAM-based baselines, V-JEPA Hand predicts hand and camera motion together from the same video representation, eliminating the need to assemble their estimates from separate systems. The results therefore support unified prediction as an accurate alternative to a multi-stage SLAM-based reconstruction pipeline on these benchmarks: recovering motion in a consistent world frame does not require a separate camera-tracking stage or test-time optimization.

4.3 Ablation studies

Effect of pseudo-labelled training data. Table 3 compares training on all four data sources with training on an equal mixture of ARCTIC and HOT3D only, keeping the architecture, initialization, and training budget fixed. We evaluate both models on EgoStandard, which neither sees during training. Adding the pseudo-labelled Ego4D and in-house data helps the model place hands more accurately in these unseen videos: reprojection and translation errors fall, along with pose and world-trajectory errors. The benefit is not simply that the model detects more hands—detection accuracy and recall are slightly lower, while F1 is comparable. Instead, the broader training data improves the geometry of the reconstruction on a benchmark outside either model's training set.

Loss ablation. Table 4 removes either the jitter loss or the 2D reprojection loss while keeping the rest of training unchanged. Without the jitter loss, predictions vary more from frame to frame even though individual-frame pose accuracy changes little. This shows why accurate poses alone do not guarantee a smooth sequence. Without the reprojection loss, joint-position error improves slightly, but the hands line up less well with the image and their wrist positions become less accurate. The two terms therefore serve different purposes: one discourages temporal jumps, while the other keeps the reconstruction consistent with the image.

Pretraining and temporal context. Table 5 changes one choice at a time: starting the encoder from random weights, freezing its pretrained weights, or replacing the video input with a single frame. The remaining training setup is shared. Training from scratch gives much less accurate poses and world trajectories on both benchmarks. Even a frozen pretrained encoder performs substantially better, and fine-tuning it improves the results further. Thus, pretraining provides useful features for hand reconstruction, while task-specific training helps the model make better use of them. Single-frame predictions remain reasonably accurate in each image, but are much less smooth over time. Seeing neighboring frames therefore helps not only with pose accuracy, but also with keeping successive predictions consistent.

Encoder ablation. Table 6 compares four pretrained ViT-L encoders with the same predictor and fine-tuning recipe. V-JEPA 2.1 gives the most accurate hand poses and world trajectories, and the smoothest predictions, on both benchmarks. Video encoders outperform image encoders on world trajectories, even when the latter provide features from multiple frames. This suggests that video features help recover motion beyond simply combining per-frame image features. Differences in pretraining data and input settings are detailed in Appendix F.

4.4 Model size and speed

V-JEPA Hand has 607.1M parameters and achieves 91.3 forward FPS on a single RTX A6000 while jointly recovering hand and camera trajectories. WiLoR, HaMeR and WildHands achieve 16.2, 19.7 and 21.1 forward FPS, respectively, on the same GPU model.

4.5 Qualitative results

Figure 4 in Appendix A shows challenging cases beyond controlled capture settings, including very low light and outdoor activities. V-JEPA Hand recovers hands that the baselines miss in the darkest example and produces plausible hand reconstructions in outdoor and cluttered scenes. These examples provide qualitative evidence of generalization to diverse in-the-wild conditions. Additional ARCTIC and HOT3D comparisons are provided in Appendix A (Figure 3).

Table 3: **Training-data ablation.** To assess how pseudo-labelled training data affects generalization, we compare models trained with and without it on EgoStandard, a benchmark unseen by either model during training.

	Detection			3D Pose				Orient. & Position						Temporal	World Frame	
Method	Acc↑	Rec↑	F1↑	MP-p↓	MP↓	PA-p↓	PA↓	EP-p↓	EP↓	GO-p↓	GO↓	CT-p↓	CT↓	Jit↓	W-MP↓	WA-MP↓
ARCTIC+HOT3D only	**0.996**	**0.998**	**0.998**	18.68	18.41	8.69	8.39	16.66	15.01	12.05	11.75	0.088	0.087	4.99	70.95	34.30
Ours (4-way)	0.991	0.996	**0.998**	**17.44**	**16.92**	**8.03**	**7.46**	**10.28**	**6.68**	**11.10**	**10.50**	**0.046**	**0.044**	**4.94**	**64.12**	**31.75**

Table 4: **Loss ablation: paired metric views.** ARCTIC.

	Detection			3D Pose				Orient. & Position						Temporal	World Frame	
Method	Acc↑	Rec↑	F1↑	MP-p↓	MP↓	PA-p↓	PA↓	EP-p↓	EP↓	GO-p↓	GO↓	CT-p↓	CT↓	Jit↓	W-MP↓	WA-MP↓
Full	**0.997**	**0.999**	**0.999**	15.79	15.63	**7.27**	**7.10**	**9.91**	**9.81**	11.55	11.35	**0.019**	**0.018**	**6.46**	**17.98**	**11.98**
w/o jitter	**0.997**	0.998	**0.999**	15.81	15.59	7.38	7.15	10.16	10.04	11.69	11.42	**0.019**	**0.018**	8.29	18.33	12.24
w/o 2D reproj.	**0.997**	**0.999**	**0.999**	**15.55**	**15.38**	7.32	7.14	11.54	11.45	**11.41**	**11.20**	0.021	0.020	6.71	18.07	12.14

Table 5: **Pretraining and temporal context: paired metric views.** ARCTIC and HOT3D. †Single-frame world errors use ground-truth camera rotation and translation (oracle), rather than predicted cameras, and are excluded from world-column ranking.

	Detection			3D Pose				Orient. & Position						Temporal	World Frame	
Method	Acc↑	Rec↑	F1↑	MP-p↓	MP↓	PA-p↓	PA↓	EP-p↓	EP↓	GO-p↓	GO↓	CT-p↓	CT↓	Jit↓	W-MP↓	WA-MP↓
ARCTIC																
Ours (16 frames)	**0.997**	**0.999**	**0.999**	**15.79**	**15.63**	**7.27**	**7.10**	**9.91**	**9.81**	**11.55**	**11.35**	**0.019**	**0.018**	**6.46**	**17.98**	**11.98**
Random init.	0.993	0.996	0.998	24.74	24.32	11.00	10.52	17.07	16.34	16.89	16.38	0.032	0.030	8.74	37.82	23.20
Frozen encoder	**0.997**	0.998	**0.999**	17.28	17.06	7.99	7.75	10.60	10.33	12.46	12.19	0.027	0.025	7.18	21.14	13.72
Single frame	0.996	0.998	**0.999**	17.79	17.54	8.42	8.15	11.71	11.10	12.70	12.44	0.021	0.020	14.71	21.27†	14.06†
HOT3D																
Ours (16 frames)	**0.989**	0.993	**0.997**	**13.59**	**12.75**	**7.46**	**6.57**	8.09	**5.81**	**7.98**	**7.09**	**0.020**	**0.017**	**6.91**	**19.84**	**11.49**
Random init.	0.969	0.984	0.988	29.02	27.04	12.15	9.89	18.97	12.34	22.27	20.23	0.045	0.037	9.91	48.81	26.87
Frozen encoder	**0.989**	**0.994**	**0.997**	15.24	14.50	7.92	7.13	**7.79**	6.41	9.41	8.58	0.024	0.021	7.90	23.75	13.40
Single frame	0.985	0.992	0.995	16.86	15.83	8.76	7.66	9.05	7.00	10.62	9.48	0.027	0.024	18.31	25.37†	15.63†

Table 6: **Encoder ablation: paired metric views.** ARCTIC and HOT3D.

	Detection			3D Pose				Orient. & Position						Temporal	World Frame	
Method	Acc↑	Rec↑	F1↑	MP-p↓	MP↓	PA-p↓	PA↓	EP-p↓	EP↓	GO-p↓	GO↓	CT-p↓	CT↓	Jit↓	W-MP↓	WA-MP↓
ARCTIC																
V-JEPA 2.1 (ours)	**0.997**	**0.999**	**0.999**	**15.79**	**15.63**	**7.27**	**7.10**	**9.91**	**9.81**	**11.55**	**11.35**	**0.019**	**0.018**	**6.46**	**17.98**	**11.98**
VideoMAE	0.978	0.988	0.994	20.26	18.62	10.34	8.57	16.65	11.56	15.00	13.28	0.038	0.030	7.96	21.69	14.42
DINOv3	0.994	0.997	0.998	18.66	18.32	8.58	8.20	11.79	11.02	13.34	12.95	0.030	0.028	8.47	35.76	20.73
PE-Spatial	0.994	0.997	0.998	20.72	20.41	9.58	9.24	13.28	12.56	14.35	13.99	0.026	0.024	8.94	37.60	22.02
HOT3D																
V-JEPA 2.1 (ours)	**0.989**	**0.993**	**0.997**	**13.59**	**12.75**	**7.46**	**6.57**	**8.09**	**5.81**	**7.98**	**7.09**	**0.020**	**0.017**	**6.91**	**19.84**	**11.49**
VideoMAE	0.984	0.991	0.995	16.92	15.83	8.76	7.60	9.84	6.86	10.96	9.80	0.025	0.021	8.58	22.26	12.95
DINOv3	0.983	0.990	0.994	17.85	16.55	9.01	7.62	10.50	7.35	11.78	10.38	0.027	0.023	9.95	45.93	25.24
PE-Spatial	0.981	0.989	0.993	20.47	19.16	9.76	8.34	11.86	8.27	14.11	12.74	0.030	0.026	10.31	49.21	27.22

5 Conclusion

We presented V-JEPA Hand, a simple model that jointly reconstructs bilateral hand motion, camera motion and intrinsics from egocentric video in one forward pass, without hand detection, SLAM or test-time optimization. Built on V-JEPA 2.1 and trained in a single stage on mixed pseudo-labelled and high-quality data, it achieves the lowest coverage-aware MPJPE among reproduced baselines on ARCTIC, HOT3D and the unseen EgoStandard benchmark, and lower world-frame MPJPE than SLAM-based pipelines. Ablations show that mixing in pseudo-labelled video improves generalization, and validate video pretraining, temporal context and encoder choice. Running at over 90 frames per second, V-JEPA Hand enables large-scale bilateral hand-motion annotation of egocentric video to support robot imitation learning in embodied AI.

AI USE STATEMENT

We used large language models to assist with language editing, code development and debugging, and evaluation and result analysis. The authors take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

The architecture and training procedure are described in Section 3, with hyperparameters in Appendix C. Section 4.1 and Appendix E document the evaluation protocol, data splits, and baseline sources. We plan to release the training code, evaluation scripts, and configuration files.

ETHICS STATEMENT

This work uses existing egocentric datasets and in-house egocentric recordings, and does not perform person identification. Our goal is to support embodied-AI research by turning egocentric videos into hand-motion annotations for robot imitation learning, with potential applications in assistive robotics and human–robot collaboration. Any release will respect the source datasets' licenses and usage restrictions.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mahmoud Assran, Adrien Bardes, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *arXiv preprint arXiv:2406.09598*, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)*, 2024. V-JEPA.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18135–18143, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhengdi Chen et al. Dyn-HaMR: Recovering 4D interacting hand motion from a dynamic camera. *arXiv preprint arXiv:2412.12861*, 2025.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3D hand reconstruction with graph-guided bi-scanning mamba. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–9, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025a.

Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025b.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10118–10128. IEEE, 2021.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

Lightwheel AI. EgoStandard: Hardware-assisted egocentric multi-modal hand motion benchmark. `https://lightwheel.ai/egosuite`, 2025.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2409.12259.

Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3D hand pose estimation in everyday egocentric images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 36(6), 2017.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Anna H Smeragliuolo, N Jeremy Hill, Luis Disla, and David Putrino. Validation of the leap motion controller using markered motion capture technology. *Journal of biomechanics*, 49(9):1742–1750, 2016.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yuxi Wang, Chengkai Jin, Yufei Liu, Wenqi Ouyang, Tianyi Wei, Zhiwei Zeng, Siyuan Huang, Zhiqi Shen, and Xingang Pan. The surprising effectiveness of video diffusion models for hand motion reconstruction. *arXiv preprint arXiv:2606.30308*, 2026.

John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. SeqHAND: RGB-sequence-based 3D hand pose and shape estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. *arXiv preprint arXiv:2501.02973*, 2025.

Xin Zhang, Yuqi Ren, Di Kang, Junyu Xie, et al. OmniHands: Towards robust 4D hand mesh recovery via A versatile transformer. *arXiv preprint arXiv:2405.20330*, 2024.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

APPENDIX

CONTENTS

C Hyperparameters 17

D Training objective details 17

E Evaluation protocols and baseline setup 18

F The two metric views 19

B Encoder Feature Visualization 16

A ADDITIONAL QUALITATIVE COMPARISONS

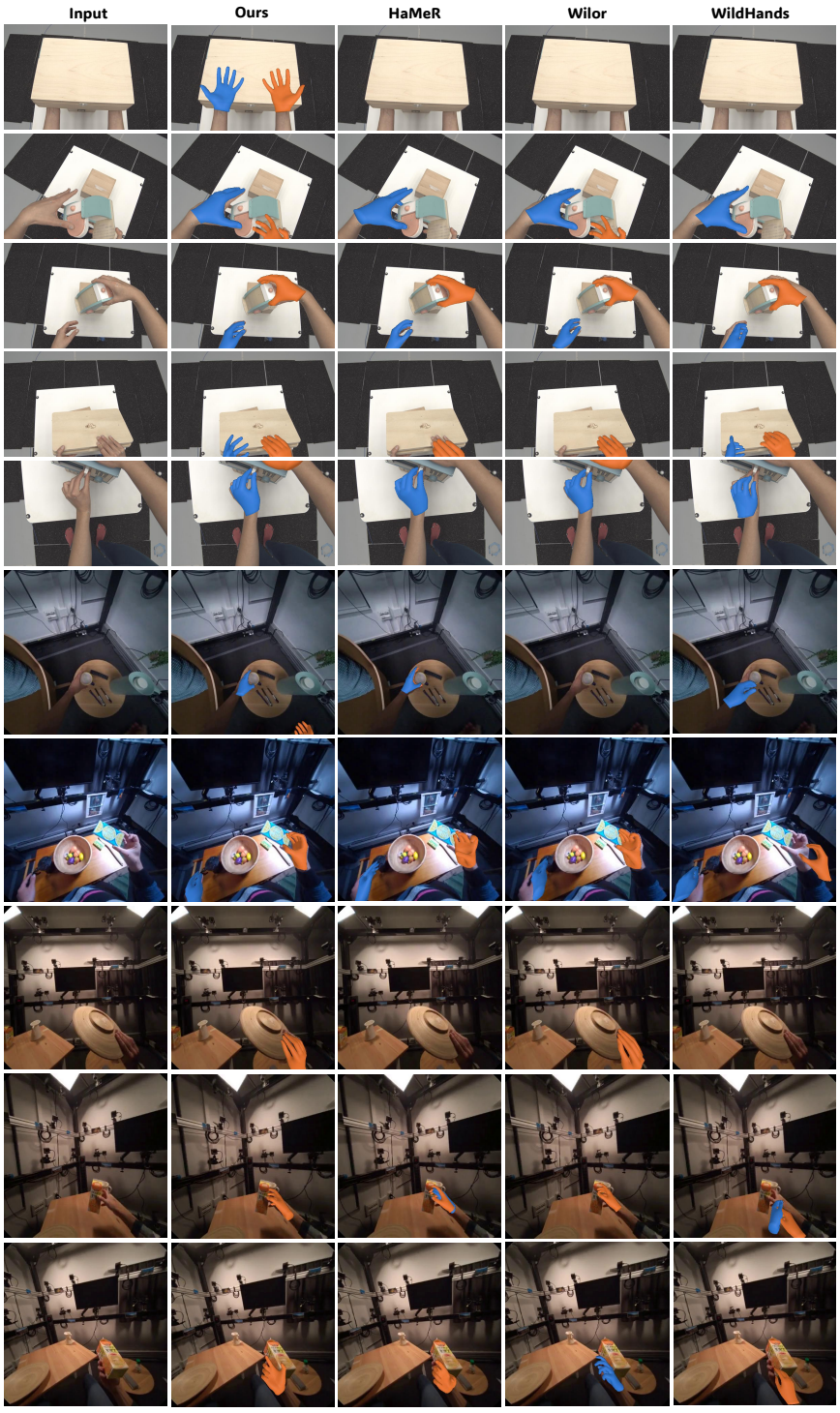


Figure 3: **Qualitative comparisons on ARCTIC and HOT3D.** Hand reconstructions from V-JEPA Hand, HaMeR, WiLoR, and WildHands.

Figure 4 shows challenging in-the-wild cases, including very low light, outdoor activities, and cluttered scenes.

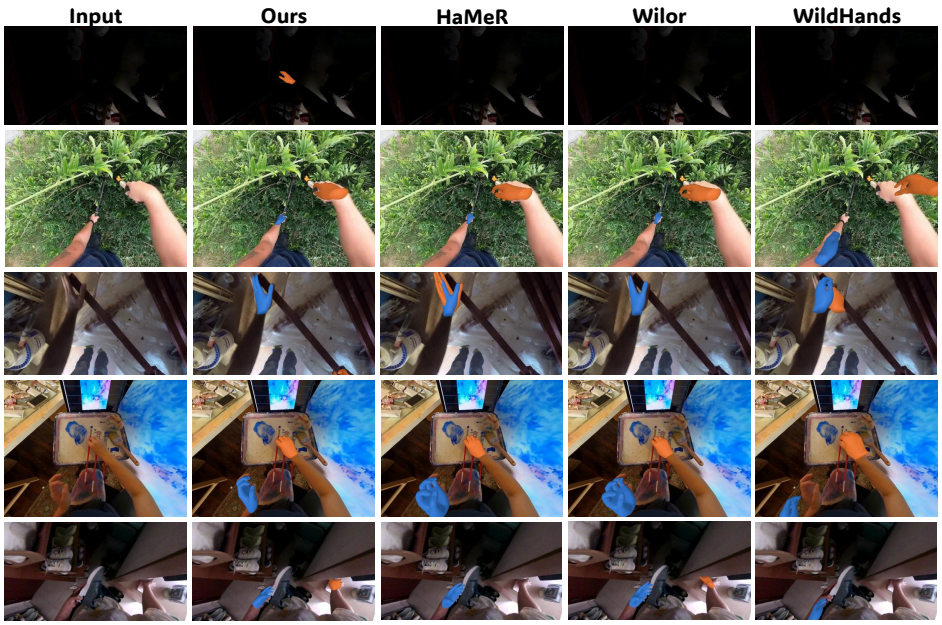


Figure 4: **Challenging in-the-wild cases.** Comparisons with HaMeR, WiLoR, and WildHands in low-light, outdoor, and cluttered scenes.

B ENCODER FEATURE VISUALIZATION

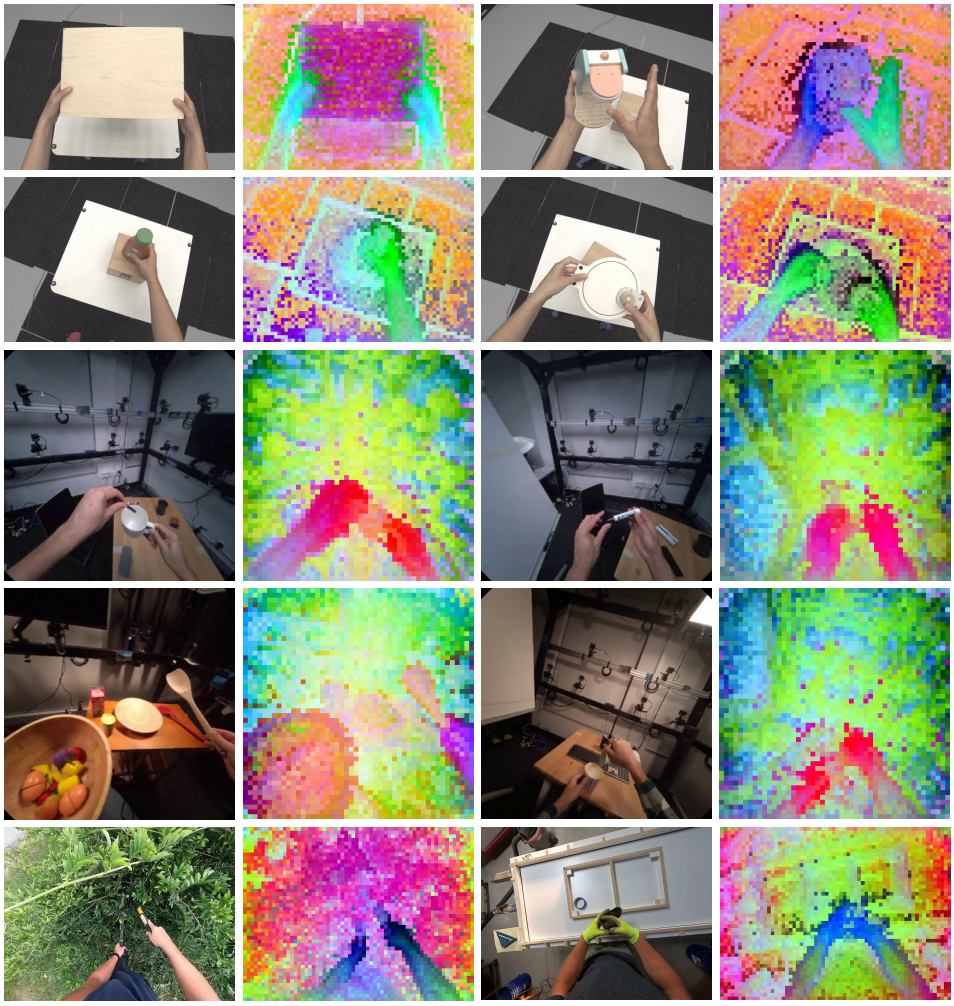


Figure 5: **Qualitative results.** Visualizations of encoder feature.

C HYPERPARAMETERS

Table 7 lists the architecture, loss, optimization and data settings of the reported V-JEPA Hand model, which is trained in a single stage on the mixture of pseudo-labelled and high-quality data.

Table 7: **Hyperparameters.** Model and training settings for V-JEPA Hand. For smooth-L_1 regression, δ is the absolute error at which the penalty changes from quadratic to linear.

Group	Hyperparameter	Value
Encoder	backbone	V-JEPA 2.1 ViT-L/384 (pretrained)
	tubelet / patch	2 (time) / 16×16 (space)
	tokens per clip	$8 \times 24 \times 24 = 4608$, dim 1024
	fine-tuning	full (unfrozen)
Predictor	type / depth	ViT-L / 24 layers
	width / heads	1024 / 16, MLP ratio 4
	query tokens	16 (one per frame)
	attention	joint self-attention
	stochastic depth	0.1
Targets	end-effector pose	9-dim (3 trans + 6D rot), per hand
	articulation	45-dim MANO Euler angles, per hand
	presence	1 logit per hand per frame
Loss	regression	smooth-L_1; weight 1.0 per term
	transition δ	0.05 (wrist and camera rotation) 0.03 (wrist / camera translation, 3D joints) 0.02 (articulation, 2D reproj., focal lengths)
	joint geometry	neutral-shape MANO FK
	jitter	L_1 joint acceleration, weight 0.2
	presence	BCE, weight 0.5
Optimization	optimizer	AdamW, β=(0.9, 0.95), wd 0.05
	lr (predictor)	6×10^{-4}
	lr (encoder)	2×10^{-5}
	schedule	1,800-step linear warmup + cosine decay, min ratio 5×10^{-4}
	training steps	20,000
	effective batch	64 clips (8 GPUs × grad-accum 8)
Data	sources (prob.)	pseudo-labelled 0.35, ARCTIC 0.45, HOT3D 0.20
	clip length / size	16 frames / 384×384
	augmentation	frame stride $\{1, 2, 4, 8\}$, horizontal flip, aspect ratio $[1, 2]$
	intrinsics input	withheld with probability 0.5

D TRAINING OBJECTIVE DETAILS

We detail the losses in Equation 6 using the notation of Section 3.1: t indexes frames and $s \in \{\mathrm{L}, \mathrm{R}\}$ indexes hands. Hats denote predictions and unhatted quantities denote targets. We write ρ_δ for the standard smooth-L_1 loss averaged over coordinates, where δ is its fixed quadratic-to-linear transition threshold. For hand geometry, $\langle\cdot\rangle_{v^s}$ averages over frames with a valid, visible target for hand s among the video clips processed together in one forward/backward pass. The two hands' losses are summed.

Hand parameters. We directly supervise wrist translation $\mathbf{d}_t^s \in \mathbb{R}^3$ in camera-space metres, wrist orientation $\mathbf{r}_t^s \in \mathbb{R}^6$ in the 6D rotation representation, and finger articulation $\boldsymbol{\theta}_t^s \in \mathbb{R}^{45}$ as Euler

angles in radians:

$$\mathcal{L}_d = \sum_s \langle \rho_{0.03}(\hat{\mathbf{d}}_t^s, \mathbf{d}_t^s) \rangle_{v^s},$$
$$\mathcal{L}_r = \sum_s \langle \rho_{0.05}(\hat{\mathbf{r}}_t^s, \mathbf{r}_t^s) \rangle_{v^s}, \tag{7}$$
$$\mathcal{L}_\theta = \sum_s \langle \rho_{0.02}(\hat{\boldsymbol{\theta}}_t^s, \boldsymbol{\theta}_t^s) \rangle_{v^s}.$$

The subscripts of ρ give the fixed threshold for each term.

Forward kinematics. The differentiable MANO forward-kinematics layer FK maps the predicted hand state $\hat{\mathbf{x}}_t^s$ to $\hat{\mathbf{J}}_t^s = \mathrm{FK}(\hat{\mathbf{x}}_t^s) \in \mathbb{R}^{21\times 3}$, the 21 camera-space joints in metres, using neutral hand shape. We supervise these positions with

$$\mathcal{L}_j = \sum_s \left\langle \rho_{0.03}\big(\hat{\mathbf{J}}_t^s, \mathbf{J}_t^s\big) \right\rangle_{v^s}. \tag{8}$$

The target $\mathbf{J}_t^s \in \mathbb{R}^{21\times 3}$ is the dataset's measured 3D joints when available and $\mathrm{FK}(\mathbf{x}_t^s)$ otherwise. This term couples wrist pose and finger articulation through their effect on joint positions.

Reprojection. Let $\pi_{\mathbf{k}}$ project 3D joints to image coordinates using the ground-truth calibration, with $\mathbf{k} = (f_x, f_y)$ denoting the horizontal and vertical focal lengths. The reprojection loss encourages image alignment:

$$\mathcal{L}_\pi = \sum_s \left\langle \rho_{0.02}\big(\pi_{\mathbf{k}}(\hat{\mathbf{J}}_t^s), \pi_{\mathbf{k}}(\mathbf{J}_t^s)\big) \right\rangle_{v^s}. \tag{9}$$

Ground-truth intrinsics remain available to this loss even when they are withheld from the predictor.

Temporal smoothness. The jitter loss $\mathcal{L}_a$ is the mean absolute second-order difference of predicted joints,

$$\hat{\mathbf{J}}_{t+1}^s - 2\hat{\mathbf{J}}_t^s + \hat{\mathbf{J}}_{t-1}^s, \qquad t = 2, \ldots, T-1, \tag{10}$$

where T is the clip length. We average over joint coordinates and triplets with valid targets in all three sampled frames, then sum over hands. Constant-velocity motion in camera coordinates incurs no penalty.

Hand presence. The presence loss $\mathcal{L}_c$ applies binary cross-entropy to the predicted logit $\hat{c}_t^s$ and the binary hand-presence target. It includes both present and absent hands, but excludes unknown targets and frames outside the anchor recording. We average over valid frames for each hand and sum the two losses.

Camera motion. The camera loss $\mathcal{L}_g$ supervises the transform from each camera frame to the first-frame anchor. It sums smooth-L_1 losses on the 6D rotation and signed-log translation defined in Equation 3, with thresholds 0.05 and 0.03, respectively. Each block is averaged over its coordinates and frames with valid camera targets in the anchor recording.

Intrinsics loss. The intrinsics loss $\mathcal{L}_k$ supervises the horizontal and vertical focal lengths (f_x, f_y) using smooth-L_1, averaged over the two components. It is applied only to clips whose intrinsics were withheld from the predictor.

Loss weights. All regression terms have unit weight, including both camera-motion blocks. The jitter and presence losses have weights 0.2 and 0.5, respectively.

E Evaluation protocols and baseline setup

Baseline implementations. We run WiLoR, HaMeR, WildHands and HaWoR for the comparisons in Table 1. ViDiHand's ARCTIC/HOT3D detection, coverage-aware pose and jitter scores are taken from Table 1 of Wang et al. (2026). Since its pretrained weights are not publicly available, we cannot reproduce its true-positive-only scores or evaluate it on EgoStandard; the unreported scores are marked with dashes.

Hand evaluation. We follow ViDiHand's evaluation protocol (Wang et al., 2026). For ARCTIC, we use the dataset's official validation split; for HOT3D, we follow the validation split used by HaWoR (Zhang et al., 2025). The evaluated clips in Table 1 cover approximately 13 minutes of ARCTIC, 49 minutes of HOT3D and 60 minutes of EgoStandard. We use the same trained V-JEPA Hand model for all three benchmarks.

World-coordinate comparison. Table 2 uses WiLoR+SLAM, HaMeR+SLAM, HaWoR, and V-JEPA Hand. It uses the same evaluation splits and trained V-JEPA Hand model as Table 1. The +SLAM baselines use the DROID-SLAM camera implementation distributed with HaWoR, with metric scale estimation; their camera inference is independent of the hand model. They are separate from the standalone WiLoR and HaMeR rows in Table 1. We first retain clips with completed predictions from all four displayed methods, then evaluate on their common true-positive hands in each frame. At least three selected frames are required per hand track. Gaps do not restart the alignment, and errors are weighted by the number of selected hand-frames, not averaged equally across tracks.

World-coordinate errors are therefore conditional on successful reconstruction by all compared methods. Missing predictions remain included in the coverage-aware hand evaluation rather than receiving zero error.

W-MPJPE fits an SE(3) transformation to the 21 joints in the first selected frame and evaluates the entire selected track. WA-MPJPE fits one SE(3) transformation to all selected joints across the track. Neither fits scale. RTE is the mean wrist error after the latter alignment, divided by the maximum ground-truth wrist displacement from the first selected frame, and multiplied by 100. Tracks with an extent below 1 mm are excluded from RTE only. This full-segment evaluation is distinct from the 16-frame world horizon of the historical ablations.

Training-data comparison. Table 3 compares four-source training with training on ARCTIC and HOT3D alone. Both models use the same V-JEPA 2.1 initialization, training duration and HOT3D training-window filter. The two-source model samples ARCTIC and HOT3D equally, without Ego4D or in-house data. This ablation uses a longer training schedule than the main comparison. We evaluate unsmoothed predictions on EgoStandard against MANO-fit targets, using projected MANO mesh bounding boxes for matching. World-coordinate errors use the two models' common true positives, with the same rigid alignment and full-track metric definitions as Table 2.

Comparability across methods. The following differences matter when interpreting Table 1:

- **HOT3D split.** The official test sequences have no public ground-truth MANO. Both evaluations therefore hold out annotated data, but use different subsets. Table 1 uses the HaWoR-aligned split; ViDiHand uses a random 5% of the validation sequences. Its HOT3D scores are therefore not directly comparable with ours; Table 1 shows them in gray as a reference only and excludes them from ranking.
- **HOT3D unlabelled hands.** HOT3D's 3D tracker can label hands even when they are outside the image. A missing target does not establish that a hand is absent. We therefore ignore unlabelled hands rather than treating them as negatives: predictions there receive neither credit nor penalty. We apply this rule separately to each hand, so the other hand in the frame is still scored. Labelled off-screen hands use the protocol's off-screen filter. A prediction is still penalized if its box does not match the corresponding off-screen ground-truth hand. The ignore rule leaves true positives and false negatives unchanged; only FAcc and F1 are affected.
- **Training exposure.** The baselines have different training exposure. WiLoR and HaWoR train on both ARCTIC and HOT3D. WildHands trains on ARCTIC. HaMeR sees neither benchmark. V-JEPA Hand includes ARCTIC and HOT3D in training.

F THE TWO METRIC VIEWS

We report two complementary views of pose accuracy, following the coverage-aware evaluation implementation of Wang et al. (2026) for metrics with the $-p$ suffix. The *coverage-aware* view replaces each unmatched ground-truth hand with a canonical rest-pose MANO placeholder before

averaging. These columns carry the -p suffix and reflect both detection coverage and reconstruction accuracy. The *true-positive-only* view averages pose errors over matched prediction–ground-truth hands. FAcc, recall and F1 report detection separately.

The main comparison and all main-text ablations pair coverage-aware and true-positive-only pose columns, separating pose quality from coverage. Comparisons must be made within the same column and metric view; a coverage-aware score must not be compared against a true-positive-only score: coverage-aware errors depend on both detection recall and pose accuracy.

Table 1 includes both views for V-JEPA Hand and the four local baseline reproductions. Only the quoted ViDiHand true-positive-only cells remain blank, because these scores cannot be recovered from published coverage-aware averages. Tables 4, 5 and 6 include both views for all ablation rows. The loss ablation is evaluated only on ARCTIC; the other two include both benchmarks.

Both views are computed from the same predictions; neither is inferred from the other.

Table notation and units. Table headers abbreviate frame accuracy (FAcc) as **Acc**, recall as **Rec**, MPJPE as **MP**, PA-MPJPE as **PA**, EPE as **EP**, and Jitter as **Jit**. F1, global-orientation error (**GO**), and camera-translation error (**CT**) retain their short names. The pairs MP-p/MP, PA-p/PA, EP-p/EP, GO-p/GO, and CT-p/CT report coverage-aware and true-positive-only errors, respectively. In the world comparison and ablations, **W-MP** and **WA-MP** denote W-MPJPE and WA-MPJPE (Zhang et al., 2025). The world comparison additionally reports relative trajectory error (**RTE**) as a percentage; alignment and scoring details are given in Appendix E. Across all tables, MP, PA, W-MP, and WA-MP are in millimetres; EP is in pixels, GO in degrees, CT in metres, and Jit in mm/frame^2. The $-p$ columns use the same units as their unsuffixed counterparts. Detection, Jit, and world-frame metrics are reported once rather than as paired columns. Detection scores lie in $[0, 1]$ and are higher-is-better; all other metrics are lower-is-better. Jitter is true-positive-only in both metric views. Bold marks column bests where applied, and shading runs from white (worst) to yellow (best) within each benchmark block. Colours are normalized within each column and do not indicate the absolute size of an improvement.

Shared ablation setup. The loss, pretraining and encoder ablations use a smaller two-phase recipe: training on half of the Ego4D pseudo-labels, followed by fine-tuning on ARCTIC and HOT3D. Each ablation changes one component of this shared setup and is evaluated after both training phases. This setup differs from the single-stage four-source model in Table 1 in its training data, schedule and HOT3D training-window filter. The main model requires both hands to have ground truth and project inside the image throughout each training window. The component ablations do not use this filter and are comparable to one another, but not directly to Table 1. The training-data ablation (Table 3) instead compares two models with matched training schedules on EgoStandard, using the hand protocol of Table 1 and a separate two-model common-support world evaluation. The final two columns of the loss, pretraining and encoder ablation tables report world-frame trajectory errors in millimetres over a fixed 16-frame evaluation horizon, using rigid alignment without scale fitting. W-MPJPE aligns predicted and ground-truth joints at the first selected frame and averages errors over all selected frames; WA-MPJPE aligns all frames before averaging joint errors. The single-frame variant cannot predict relative camera motion from one frame. Its world-frame errors therefore use ground-truth camera rotation and translation to lift independently predicted hands into world coordinates. The camera poses are used only for scoring, not as input to the hand model. These oracle-camera results are marked with a dagger and excluded from world-column ranking. This evaluation uses the same historical HOT3D split as the other component ablations, not the main-table split.

Encoder substitution details. All four backbones use ViT-L and are fine-tuned with the same self-attention predictor design. V-JEPA 2.1 and VideoMAE use 16 input frames and two-frame tubelets; DINOv3 and PE-Spatial use 8 input frames and one temporal token group per frame. All thus produce eight temporal groups, but their spatial grids and input durations differ. Output query counts match the input frame counts. The configured training image size is 384 for V-JEPA 2.1 and 224 for the alternatives; PE-Spatial's positional grid is interpolated from its pretrained 448-pixel configuration. For DINOv3 and PE-Spatial, frames are processed as a batch of images, prefix tokens are removed, and patch tokens are concatenated in temporal order for the predictor. Temporal attention is therefore learned in the predictor rather than applied within the image backbone. World-

frame scoring uses the same 16-frame evaluation horizon regardless of the model's input window length.

The encoders also differ in pretraining data and supervision. VideoMAE uses Kinetics-400 masked reconstruction. DINOv3 uses the LVD-1689M ViT-L/16 checkpoint distilled from its image self-supervised teacher. Gram anchoring is used in the teacher's training, not directly in the small-model distillation stage. PE-Spatial-L/14 is distilled from the spatially aligned PE-Spatial-G model, rather than used as an unmodified contrastive encoder. Our V-JEPA 2.1 ViT-L is distilled from a teacher trained on the VisionMix image–video corpus with dense latent prediction and deep self-supervision. These are comparisons of transferred pretrained representations, not equal-data pretraining trials.