StableHand: Quality-Aware Flow Matching for World-Space Dual-Hand Motion Estimation from Egocentric Video

Huajian Zeng[1] **Chaohua Yao**[2] **Yuantai Zhang**[1] **Jiaqi Yang**[1]
Rolandos Alexandros Potamias[3] **Xingxing Zuo**[1,*]

[1] Mohamed bin Zayed University of Artificial Intelligence
[2] University of Illinois at Urbana-Champaign
[3] Imperial College London

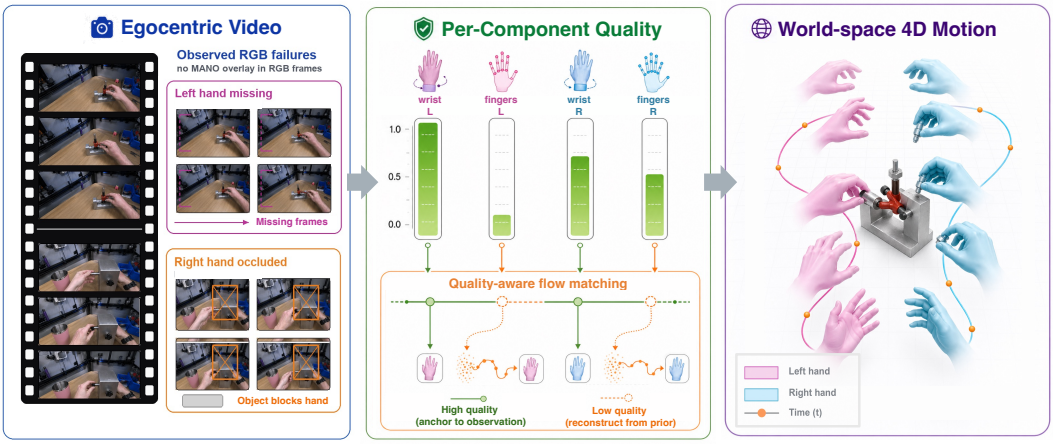


Figure 1: **StableHand recovers world space dual-hand motion from egocentric video.** We introduce StableHand, a quality-aware flow-matching framework driven by a per-component (wrist and fingers), per-hand quality signal $\mathbf{q} \in [0,1]^4$ (middle). Given egocentric inputs with missing or occluded hands (left), StableHand anchors reliable hand observations from a hand pose estimator and regenerates unreliable ones from a learned bimanual motion prior, yielding consistent world space dual-hand trajectories (right).

Abstract

Recovering world-space 4D motion of two interacting hands from egocentric video is a fundamental capability for supervising robot policy learning, where wrist trajectories track the end-effector and finger articulations specify the grasp pose. Two major challenges arise in this setting: hands frequently leave the camera view for extended periods due to head motion, and persistent hand–object interactions cause severe occlusions of one or both hands. Existing methods uniformly condition on noisy hand motion observations without accounting for their per-frame reliability, leading to substantial performance degradation. Our key insight is that accurate world-space hand motion esimation is tightly coupled with the quality of per-frame hand observations. To this end, we decompose the quality of hand motion observations extracted from an off-the-shelf hand pose estimator into four channels: wrist global translation and finger articulations for both hand. We propose StableHand, a

Project page: `https://huajian-zeng.github.io/projects/stablehand`
*Corresponding author.

Preprint.

quality-aware flow-matching framework conditioned on these four-channel quality signals, which are predicted by a learned quality network. We naturally incorporate the quality signals into the flow-matching process through a per-channel forward schedule, a quality-adjusted velocity target, AdaLN modulation of the DiT denoiser, and a quality-aware ODE initialization. This unified generative process preserves high-quality observations while reconstructing unreliable ones using a learned bimanual motion prior. Experiments on HOT3D and ARCTIC, two egocentric benchmarks featuring long missing-hand spans and persistent hand–object occlusions, show that StableHand achieves state-of-the-art performance across all reported metrics, reducing W-MPJPE by 20–25% compared to the strongest baseline, with the largest gains on heavily occluded ARCTIC sequences.

1 Introduction

Robot policy learning from large-scale egocentric video has scaled to dexterous and bimanual manipulation [16, 52], mobile manipulation [54], and vision-language-action pretraining [46, 12], with accurate world space 4D motion of two interacting hands as a key supervision channel [17, 14, 10]. On the robot end, the wrist trajectory drives the end-effector while the finger articulation specifies the contact and grip pose required for dexterous interaction. The egocentric capture setting poses significant challenges for recovering this signal: coupled head and hand motion often causes one or both hands to exit the camera frustum for prolonged intervals, while sustained hand–object interactions during dexterous bimanual manipulation introduce asymmetric occlusions that degrade hand observations.

For recovering accurate dual hand motions in world space, SLAM-based pipelines [31, 34, 44, 35] directly transform per-frame camera-frame hand poses into the world frame via monocular SLAM [43] without further refinement, drifting catastrophically across long missing-hand spans for lack of any temporal prior. Existing native world space methods [51, 49, 42] learn a sequence-level motion prior that jointly reasons about hand and camera, but condition uniformly on visual features and ignore the fact that wrist global drift is several times larger than finger articulation error, and that bimanual occlusion degrades the two hands asymmetrically.

In this work, we propose **StableHand** (Fig. 1), a quality-aware flow-matching framework that synthesizes the world space 4D motion of two interacting hands from egocentric video. The framework is conditioned on a four-channel quality signal, predicted at inference by a learned quality network from hand observations extracted from an off-the-shelf hand-pose estimator. Our key insight is that observation quality heterogeneity should serve as an explicit conditioning signal of the generative process: a single generative process can then anchor high-quality channels on the observation while regenerating low-quality channels from the prior.

Realizing this insight raises three central challenges: **(i) Heterogeneous quality decomposition.** A single scalar quality per hand would average wrist global drift with finger articulation error and ignores asymmetric bimanual occlusion, informing our four-channel signal indexed by hand and component. **(ii) Per-channel flow-matching pathway.** Truncated reverse processes [26, 25] and learned multivariate noise schedules [38] fail to anchor high-quality channels while regenerating low-quality ones, motivating our per-channel forward schedule whose induced velocity target vanishes on frozen channels and a quality-aware initialization that mirrors the training-time forward. **(iii) Quality signal prediction at inference.** Since the quality signal is derived from ground-truth joints at training time and unavailable at deployment, we pretrain a quality network on egocentric bimanual hand-pose corpora to predict it from observed hand parameters, hand-confidence flags, and camera pose.

In summary, our contributions are:

- We frame egocentric world space dual-hand recovery as per-component, quality-aware generation, identifying the two-axis (component, hand) heterogeneity of observation quality as the structure prior pipelines leave implicit.
- We design a quality-aware flow-matching pathway with four coupled mechanisms driven by the same four-value quality signal: a per-channel forward schedule, a quality-adjusted velocity target, AdaLN modulation of a Diffusion Transformer (DiT), and a quality-aware ODE initialization.

- We introduce a calibrated four-channel quality signal with a corpus-adaptive radial-basis bandwidth, predicted at inference from the egocentric visual stream by a learned quality network.
- On HOT3D and ARCTIC, two egocentric benchmarks targeting long missing-hand spans and persistent hand-object occlusion, StableHand achieves state-of-the-art world space dual-hand motion estimation across every reported metric, with 20–25% W-MPJPE reductions over the strongest baselines and the largest margins on the most occlusion-affected clips.

2 Related Work

Egocentric Hand Recovery. MANO [37] introduced a low-dimensional parametric hand model that enabled the single-image hand regressor [3] and inspired direct vertex prediction [19, 20]. Recent transformer-based pipelines such as HaMeR [31] and WiLoR [34] scale this paradigm with ViT backbones on internet-scale data, with specialized variants targeting egocentric viewpoints [35], differentiable global positioning [44], occlusion robustness [30, 8], or hand-object interaction [22, 4]. World-space hand recovery extends this line by coupling SLAM [43] or multi-stage optimization with a learned hand motion estimator [51, 49, 42, 5, 7, 48], but none exposes per-component observation quality as an explicit conditioning signal of the prediction process. Our work fills this gap with a calibrated per-component quality signal that decouples wrist global drift from finger articulation error and the two hands under asymmetric occlusion.

Quality-Conditioned Diffusion. Diffusion models [11] and conditional flow matching [21], often instantiated as DiTs [32] and applied across image and motion domains [29, 50], support diverse auxiliary conditioning signals. For visibility-aware generation, SDEdit [26] and RePaint [25] truncate or interleave the reverse process under a binary visibility distinction, while MuLAN [38] learns a multivariate input-conditional schedule end-to-end without any per-channel quality signal. None of these works factorizes the conditioning signal across anatomical components or hands. Our framework instead naturally embeds a per-component quality signal into the flow-matching generative process.

3 Methodology

Given an egocentric video $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^{T}$ of T frames containing two possibly interacting hands under a moving first-person camera, we aim to recover the world space 4D motion of both hands. As illustrated in Fig. 2, our framework consists of two learned modules. A quality network (Sec. 3.2) estimates a per-component (wrist and fingers of both hands) quality signal from the egocentric visual stream, and a quality-aware generative model (Sec. 3.3) synthesizes the world space dual-hand trajectory conditioned on this signal. In contrast to prior egocentric hand recovery that either trusts per-frame pose estimators [51] or treats every channel of the hand pose observation identically [5], this design conditions the generative process on a per-component quality signal, allowing high-quality components of the hand pose observation to anchor the trajectory while low-quality components are regenerated from the learned motion prior. The two modules are trained in sequence: the quality network is pretrained on multi-dataset hand-pose corpora and frozen, after which the generative model is trained with the predicted quality signal $\hat{\mathbf{q}}$ from the frozen quality network as conditioning.

Setup and notation. We represent the per-frame state of hand $h \in \{L, R\}$ at time t as $\mathbf{x}_t^h = (\mathbf{p}_t^h, \mathbf{r}_t^h, \boldsymbol{\theta}_t^h) \in \mathbb{R}^{54}$, with $\mathbf{p}_t^h \in \mathbb{R}^3$ the root translation in a world frame, $\mathbf{r}_t^h \in \mathbb{R}^6$ a continuous 6D wrist rotation [53], and $\boldsymbol{\theta}_t^h \in \mathbb{R}^{45}$ the MANO [37] finger axis-angle pose, yielding the dual-hand motion $\mathbf{x} \in \mathbb{R}^{T\times 108}$. A frozen hand pose estimator $\mathcal{M}$ [34] extracts a per-hand camera frame MANO observation $\mathbf{y}_t^h \in \mathbb{R}^{54}$ and a visual feature $\mathbf{f}_t^h \in \mathbb{R}^{1280}$ from each frame, while a frozen geometry model $\mathcal{G}$ [18] produces a per-frame camera pose $\mathbf{g}_t \in \mathbb{R}^9$ (6D rotation and 3D translation) and a scene-geometry token $\mathbf{s}_t \in \mathbb{R}^{3072}$ from the full video. The camera pose rigidly transforms each observation into the world frame to produce $\bar{\mathbf{y}} \in \mathbb{R}^{T\times 108}$, and is additionally consumed as cross-attention conditioning by the generative model. Throughout, $t \in \{1, \ldots, T\}$ indexes a frame in the input video, while $\tau \in [0, 1]$ indexes the flow-matching ODE time, with $\tau{=}1$ the clean target hand motion and $\tau{=}0$ the noise prior.

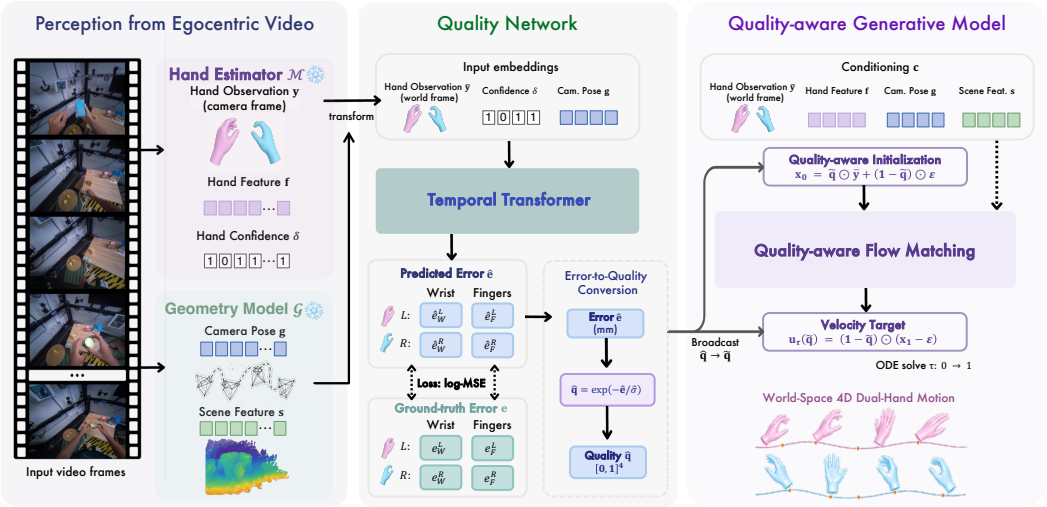


Figure 2: **StableHand pipeline.** From an egocentric video, frozen off-the-shelf modules [34, 18] produce per-hand MANO observations together with camera and scene context. Our two learned modules are a **quality network** that predicts a per-component error $\hat{\mathbf{e}}$ (wrist and fingers of each hand) converted to a quality signal $\hat{\mathbf{q}} \in [0,1]^{T\times 4}$ via the Error-to-Quality Conversion block (Eq. 3), and a **quality-aware flow-matching model** that synthesizes the world space dual-hand motion $\mathbf{x} \in \mathbb{R}^{T\times 108}$. The broadcast signal $\tilde{\mathbf{q}} \in [0,1]^{T\times 108}$ drives four coupled mechanisms (bottom-right): a per-channel noise schedule, a quality-adjusted velocity target that freezes high-quality channels while regenerating low-quality ones, AdaLN modulation inside the DiT, and a quality-aware initialization at $\tau{=}0$.

3.1 Adaptive Quality Signal

We associate every frame with a per-component quality signal $\mathbf{q}_t = [q_W^L, q_F^L, q_W^R, q_F^R]^\top \in [0,1]^4$ rating the quality of the wrist and fingers of each hand, kept as separate channels because their error scales decouple: wrist error is dominated by global translation drift on the tens-of-millimetre scale, whereas finger error lives at the millimetre scale of MANO articulation. Let $\bar{\mathbf{j}}_k^h, \mathbf{j}_k^h \in \mathbb{R}^3$ denote the observed and ground-truth 3D joint positions of hand h ($k=0$ wrist, $k=1,\dots,15$ fingers), and $\bar{\mathbf{R}}^h, \mathbf{R}^h \in SO(3)$ the corresponding wrist rotation matrices recovered from the 6D representation $\mathbf{r}_t^h$. The wrist channel sums the wrist translation error and the wrist rotation error scaled by the canonical MANO palm radius $r_0 = 85$ mm. The finger channel uses MPJPE [13] in the ground-truth wrist frame, which isolates finger articulation from wrist-orientation error via the alignment rotation $\mathbf{R}^h(\bar{\mathbf{R}}^h)^\top$:

$$e_W^h = \|\bar{\mathbf{j}}_0^h - \mathbf{j}_0^h\|_2 + r_0 \arccos\!\left[\tfrac{1}{2}\left(\mathrm{tr}(\bar{\mathbf{R}}^{h\top}\mathbf{R}^h) - 1\right)\right], \quad q_W^h = \exp(-e_W^h/\sigma_W), \tag{1}$$

$$e_F^h = \tfrac{1}{15}\sum_{k=1}^{15} \|\mathbf{R}^h(\bar{\mathbf{R}}^h)^\top(\bar{\mathbf{j}}_k^h - \bar{\mathbf{j}}_0^h) - (\mathbf{j}_k^h - \mathbf{j}_0^h)\|_2, \quad q_F^h = \exp(-e_F^h/\sigma_F). \tag{2}$$

Each entry follows an RBF kernel [39] $q = \exp(-e/\sigma)$ over a component-specific joint error. We calibrate the RBF bandwidth per corpus so that the worst 20% of frames (those with the largest errors) fall into the low-quality region $q < 0.1$. Equivalently, σ is chosen to satisfy $\exp(-e/\sigma) = 0.1$ at the 80-th-percentile error,

$$\sigma_\star = p_{80}(e_\star)/\ln 10, \quad \hat{\sigma}_\star = p_{80}(\hat{e}_\star)/\ln 10, \qquad \star \in \{W, F\}, \tag{3}$$

where $\hat{\sigma}$ is recovered at deployment from the quality network's predictions $\hat{\mathbf{e}}$ on a local-data calibration subset. Frames without a valid detection or annotation receive $q = 0$. The perturbation pool and the 20%-target sensitivity are in the supplemental material.

3.2 Quality Network

The ground-truth $\mathbf{q}$ defined by Eqs. 1–2 requires joint annotations unavailable at inference, motivating a learned predictor. In the quality network, we predict the per-component error $\hat{\mathbf{e}}$ (mm) directly rather

than the quality $\hat{\mathbf{q}}$, since errors live on a physical scale common to all corpora while $\hat{\mathbf{q}}$ depends on the corpus-specific bandwidth σ. A single $(\bar{\mathbf{y}}, \delta, \mathbf{g}) \to \hat{\mathbf{e}}$ mapping can therefore be learned across corpora whose σ values span more than an order of magnitude. At deployment, the quality signal is recovered by the Error-to-Quality Conversion (Fig. 2),

$$\hat{\mathbf{q}} = \exp(-\hat{\mathbf{e}}/\hat{\sigma}), \tag{4}$$

with the bandwidth $\hat{\sigma}$ adaptively calibrated from QN predictions on a local data subset as in Eq. 3. Hand observation quality depends on both spatial cues within a single frame (e.g., occlusion or motion blur) and temporal cues across consecutive frames (e.g., sustained detection gaps or rapid camera rotation). To capture both, our quality network jointly processes a T-frame window of world space MANO observations and hand-confidence flags (Fig. 2, middle panel).

Architecture. At each frame t, the per-hand input concatenates the world space MANO observation $\bar{\mathbf{y}}_t^h$, a binary hand-confidence flag δ_t^h, and the per-frame camera pose $\mathbf{g}_t$ that anchors wrist translation to head ego-motion. A learnable [GEO] token prepended to the temporal sequence aggregates trajectory-level context through self-attention and broadcasts it back to every per-frame token, so that each per-frame quality prediction is conditioned on the clip-level quality pattern rather than on local frames alone. A temporal self-attention encoder with RoPE [45, 41] processes the augmented sequence, after which a shared per-frame two-layer MLP outputs the log-space prediction $\log(1+\hat{\mathbf{e}}_t^h)$ that exponentiates to the per-hand error $\hat{\mathbf{e}}_t^h = (\hat{e}_W^h, \hat{e}_F^h) \in \mathbb{R}^2_{\geq 0}$ (mm), stacked across both hands as $\hat{\mathbf{e}}_t \in \mathbb{R}^4_{\geq 0}$. The frozen hand estimator's feature $\mathbf{f}_t^h$ is deliberately excluded from the input to suppress the dataset-style shortcut it injects under multi-corpus pretraining. Undetected frames ($\delta_t^h = 0$) hard-override $\hat{\mathbf{q}}_t^h = \mathbf{0}$ at the Error-to-Quality Conversion step (Fig. 2). The ablation of the visual-feature exclusion is provided in the supplemental material.

Training. The quality network is pretrained on eight egocentric bimanual hand-pose corpora ($\sim$10M frames, Sec. 4.1). A log-space mean-squared-error term supervises per-sample magnitude over frame-hand pairs $\mathcal{V}$ with valid annotations,

$$\mathcal{L}_{\text{MSE}} \;=\; \tfrac{1}{|\mathcal{V}|} \sum_{(t,h)\in\mathcal{V}} \left\| \log(1+\hat{\mathbf{e}}_t^h) - \log(1+\mathbf{e}_t^h) \right\|_2^2, \tag{5}$$

where $\mathbf{e}_t^h$ is the ground-truth per-component error of Eqs. 1–2 and the $\log(1+\cdot)$ transform balances per-corpus contributions whose error scales span over an order of magnitude. The MSE objective biases the QN toward the conditional mean and narrows the spread of $\hat{\mathbf{e}}$ relative to the ground-truth error. As a consequence, the deployment bandwidth $\hat{\sigma} = p_{80}(\hat{\mathbf{e}})/\ln 10$ falls below the training-time σ, and the deployment-time $\hat{\mathbf{q}}$ distribution is more polarized (concentrated near 0 or 1) than the one the DiT was trained on. To restore the distribution shape that the per-sample MSE compresses, we add an auxiliary 1D Wasserstein-1 distance [1] between the empirical distributions of $\log(1+\hat{\mathbf{e}})$ and $\log(1+\mathbf{e})$ within each batch. For scalar variables this distance admits the closed-form L1 difference between sorted empirical samples, providing a differentiable, distribution-level signal that per-sample MSE cannot supply:

$$\mathcal{L}_{\text{W}} \;=\; \tfrac{1}{N} \sum_{i=1}^{N} \Big| \text{sort}_i\big(\log(1+\hat{\mathbf{e}})\big) - \text{sort}_i\big(\log(1+\mathbf{e})\big) \Big|, \tag{6}$$

where $\text{sort}_i(\cdot)$ extracts the i-th smallest entry over the N valid frame-hand-component triples in the batch. This analytical 1D Wasserstein-1 distance aligns the full empirical CDF and matches $p_{80}(\hat{\mathbf{e}})$ to $p_{80}(\mathbf{e})$ by construction, recovering the training-time σ at deployment. The total quality-network loss combines the two terms,

$$\mathcal{L}_{\text{QN}} \;=\; \mathcal{L}_{\text{MSE}} + \lambda_{\text{W}}\, \mathcal{L}_{\text{W}}, \tag{7}$$

with $\lambda_{\text{W}} = 1.0$. After pretraining, the quality network is frozen and its predictions $\hat{\mathbf{e}}$ are converted to $\hat{\mathbf{q}}$ via Eq. 4 at both the generative model's training and inference. The regression target is a per-component MPJPE error in physical units (Eqs. 1–2), defined by dual-hand geometry and independent of the upstream estimator, so the same $(\bar{\mathbf{y}}, \delta, \mathbf{g}) \to \hat{\mathbf{e}}$ mapping transfers to a different regressor by retraining on its outputs and recalibrating $\hat{\sigma}$ via Eq. 3, which we validate by swapping WiLoR [34] for HaMeR [31] in the supplemental material.

3.3 Quality-Aware Hand Motion Generation

We instantiate the generative model as a quality-aware conditional flow-matching DiT (Fig. 2, right panel) and detail its four parts below: a per-channel forward process driven by $\tilde{\mathbf{q}}_t$, the resulting

training objective, a denoiser architecture with AdaLN modulation by $\tilde{\mathbf{q}}_t$, and a quality-aware inference initialization at τ=0.

Per-channel forward process. Standard conditional flow matching (CFM) [21] adopts a linear path $\mathbf{x}_\tau = \tau\,\mathbf{x}_1 + (1-\tau)\,\boldsymbol{\varepsilon}$ between the clean dual-hand motion $\mathbf{x}_1 \in \mathbb{R}^{T\times 108}$ and noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$ with velocity target $\mathbf{u}_\tau = \mathbf{x}_1 - \boldsymbol{\varepsilon}$, whose scalar schedule treats every channel identically.

In our work, we replace this scalar schedule with a per-channel one driven by the quality signal. We broadcast $\mathbf{q}$ onto $\tilde{\mathbf{q}} \in [0,1]^{T\times 108}$ by replicating q_W^h at time t onto the wrist slots $(\mathbf{p}_t^h, \mathbf{r}_t^h)$ and q_F^h onto the finger slots $\boldsymbol{\theta}_t^h$ of each hand, yielding the per-channel schedule and its associated velocity target,

$$\alpha(\tau, \tilde{\mathbf{q}}) = 1 - (1-\tau)(1-\tilde{\mathbf{q}}), \quad \mathbf{x}_\tau = \alpha \odot \mathbf{x}_1 + (1-\alpha)\odot\boldsymbol{\varepsilon}, \quad \mathbf{u}_\tau(\tilde{\mathbf{q}}) = (1-\tilde{\mathbf{q}})\odot(\mathbf{x}_1 - \boldsymbol{\varepsilon}). \tag{8}$$

where $\odot$ denotes the element-wise (Hadamard) product.

This per-channel schedule has two informative extremes that drive the design. For any motion channel c, $\tilde{q}_{t,c} = 0$ recovers standard flow matching, while $\tilde{q}_{t,c} = 1$ freezes channel c at $\mathbf{x}_1$ via $\alpha \equiv 1$ and $\mathbf{u} \equiv 0$, supervising the denoiser to leave high-quality channels untouched while regenerating low-quality channels from noise.

Training objective. A learned vector field $\mathbf{v}_\theta(\mathbf{x}_\tau, \tau; \mathbf{c})$ is trained to regress the quality-adjusted velocity target of Eq. 8 via

$$\mathcal{L}_{\mathrm{FM}} \;=\; \mathbb{E}_{\tau,\mathbf{x}_1,\boldsymbol{\varepsilon}}\big[\,\|\mathbf{v}_\theta(\mathbf{x}_\tau,\tau;\mathbf{c}) - \mathbf{u}_\tau(\tilde{\mathbf{q}})\|_2^2\,\big], \tag{9}$$

where $\mathbf{c} = \{\mathbf{c}_t\}_{t=1}^T$ collects per-frame cross-attention conditioning bundles, each $\mathbf{c}_t$ comprising the scene-geometry token $\mathbf{s}_t$ [18], the camera pose $\mathbf{g}_t$, and the per-hand observation $\mathbf{z}_t^h = [\bar{\mathbf{y}}_t^h;\, \mathbf{f}_t^h]$ for $h \in \{L, R\}$. On top of $\mathcal{L}_{\mathrm{FM}}$ we add a second-order temporal smoothness regularizer on the implied clean trajectory $\widetilde{\mathbf{x}}_1 = \mathbf{x}_\tau + (1-\tau)\,\mathbf{v}_\theta$ (which recovers $\mathbf{x}_1$ exactly via $1-\alpha(\tau,\tilde{\mathbf{q}}) = (1-\tau)(1-\tilde{\mathbf{q}})$) to suppress frame-to-frame jitter on channels generated from noise, and an auxiliary wrist-translation term $\mathcal{L}_{\mathrm{wrist}}$ to rebalance the 3-dimensional wrist slot otherwise dominated by the 45-dimensional finger component:

$$\mathcal{L}_{\mathrm{smooth}} = \tfrac{1}{T-2}\sum_{t=2}^{T-1}\|\widetilde{\mathbf{x}}_{1,t-1} - 2\widetilde{\mathbf{x}}_{1,t} + \widetilde{\mathbf{x}}_{1,t+1}\|_2^2, \quad \mathcal{L}_{\mathrm{wrist}} = \tfrac{1}{2T}\sum_{t=1}^{T}\sum_{h\in\{L,R\}}\|\widetilde{\mathbf{p}}_{1,t}^h - \mathbf{p}_t^h\|_2^2, \tag{10}$$

where $\widetilde{\mathbf{p}}_{1,t}^h$ extracts the wrist translation of $\widetilde{\mathbf{x}}_{1,t}$. The total training objective is $\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{smooth}}\mathcal{L}_{\mathrm{smooth}} + \lambda_{\mathrm{wrist}}\mathcal{L}_{\mathrm{wrist}}$ with $\lambda_{\mathrm{smooth}} = 0.1$ and $\lambda_{\mathrm{wrist}} = 5.0$.

Denoiser architecture. We instantiate $\mathbf{v}_\theta$ as a DiT [32] that processes each frame as four tokens, one per quality channel of $\tilde{\mathbf{q}}_t$ (left/right wrist and left/right fingers), letting AdaLN driven by $\mathbf{m}_t = \mathrm{MLP}_\tau(\tau) + \mathrm{MLP}_q(\tilde{\mathbf{q}}_t)$ scale wrist and finger updates to match the per-channel quality structure of the forward process. Temporal self-attention with RoPE [41, 45] spans all $4T$ tokens of the T-frame clip, while cross-attention injects the conditioning bundle $\mathbf{c}_t$ as complementary context.

Quality-aware inference initialization. At inference we replace the stochastic forward process with a *deterministic mapping* from the quality-aware initialization at τ=0 to the recovered trajectory at τ=1, realized by integrating the deterministic reverse ODE $\mathrm{d}\mathbf{x}/\mathrm{d}\tau = \mathbf{v}_\theta(\mathbf{x}, \tau; \mathbf{c})$ over $\tau \in [0,1]$ from

$$\mathbf{x}_0 \;=\; \tilde{\mathbf{q}} \odot \bar{\mathbf{y}} + (1-\tilde{\mathbf{q}})\odot\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I), \tag{11}$$

which matches the training-time $\mathbf{x}_\tau$ of Eq. 8 at τ=0 since $\alpha(0, \tilde{\mathbf{q}}) = \tilde{\mathbf{q}}$. We initialize each channel of the reverse ODE at the noise level matching its own quality, exactly as the training-time forward of Eq. 8 places it at τ=0: high-quality channels start near the observation and low-quality channels near pure noise. SDEdit-style [26] truncated reverse processes instead start every channel from a single intermediate noise level, which cannot reproduce the per-channel structure of the training-time forward (each channel sits at noise level $1 - \tilde{q}_c$ at τ=0). Our initialization eliminates this per-channel train-test mismatch by construction.

4 Experiments

We evaluate world space 4D dual-hand motion estimation against two complementary failure modes of egocentric hand observation: extended hand out-of-view spans under dynamic head-mounted camera

Table 1: **World-space hand motion estimation on HOT3D [2].** The upper block reports SLAM-based methods. The lower block reports native world space methods. ‡ numbers taken from the original paper (code not publicly available). Per-column **best** and second best are color-coded.

Method	PA-MPJPE [mm] (↓)	W-MPJPE [mm] (↓)	WA-MPJPE [mm] (↓)	AccEr [m/s^2] (↓)
HaMeR-SLAM [31]	9.07	295.32	73.90	13.81
WiLoR-SLAM [34]	8.60	171.24	54.34	12.30
HMP-SLAM [5]	11.72	131.19	42.77	5.83
Dyn-HaMR [49]	8.64	86.78	40.47	5.46
HaWoR [51]	8.65	71.89	30.20	7.80
UniHand‡ [42]	4.76	63.97	25.24	4.93
Ours	**4.02**	**57.83**	**21.02**	**3.83**

motion (HOT3D [2]), and dexterous bimanual manipulation with persistent hand-object occlusion and asymmetric per-component quality (ARCTIC [6]). Sec. 4.2 reports quantitative comparisons, and Sec. 4.3 ablates the core design choices on HOT3D.

4.1 Experimental Setup

Datasets. We evaluate on HOT3D [2] and ARCTIC [6]. The generative model is trained per benchmark while the quality network (Sec. 3.2) is shared across both. Splits and pretraining details are in the supplemental material.

Evaluation metrics. Following [47, 51, 49], we report PA-MPJPE, W-MPJPE, WA-MPJPE, and Acceleration Error (AccEr). On ARCTIC, we additionally report MRRPE [6, 27] for the two-hand relative spatial consistency. All metrics are computed on every GT-valid frame with no detection-failure filtering. Precise definitions are in the supplemental material.

Baselines. We compare against two families of prior work, distinguished by how camera information enters the world space estimation. *SLAM-based* methods directly transform a per-frame camera-frame hand pose into the world frame via a SLAM-estimated camera pose [43], with no further refinement: HaMeR [31], WiLoR [34], and HMP [5]. *Native world space* methods learn a world-frame hand motion model that jointly reasons about hand and camera, integrating SLAM (when used) as one component of a learned sequence-level pipeline rather than as a stand-alone projection: HaWoR [51], UniHand [42], and Dyn-HaMR [49]. Native world space baselines are retrained on the corresponding training split using their official codebases, while SLAM-based baselines use public checkpoints. UniHand numbers (‡ in Tab. 1) are quoted from the original paper since its code is unreleased.

Implementation details. Both networks are trained on a single H100 with AdamW [24] optimizer. Inference runs a 20-step Euler ODE solve over each $T{=}150$ clip on per-frame dual-hand observations from a frozen WiLoR estimator [34]. Full hyperparameters and architectural details are in the supplemental material.

4.2 Comparison on HOT3D and ARCTIC

HOT3D [2]: long missing-hand spans. Our method achieves the lowest error on every metric of Tab. 1, reducing W-MPJPE by $\sim$20% over HaWoR [51], the strongest baseline with public code, while SLAM-based baselines drift catastrophically for lack of any temporal prior. Two mechanisms drive this gap, visualized in Fig. 3 (top two rows): the quality network identifies missing-hand frames and the per-channel forward process regenerates them from the prior without disturbing the visible hand's observations. Stratifying test clips by missing-hand fraction (Fig. 4a) shows our W-MPJPE advantage concentrates in the high-missing regime that the quality-aware schedule targets.

ARCTIC [6]: persistent hand-object occlusion. Our method achieves the lowest error on every metric of Tab. 2, reducing W-MPJPE by $\sim$25% over the strongest baseline, with comparable margins on WA-MPJPE and MRRPE. The MRRPE gain reflects our dual-hand trajectory representation that lets the generative model reason jointly about the two hands' relative position. ARCTIC reverses the HOT3D family ranking, where native world space methods suffer disproportionately because their temporal coupling propagates one hand's noisy observation into the other's trajectory (Dyn-HaMR's W-MPJPE alone increases $8.5\times$ from HOT3D to ARCTIC). The aggregate metrics also mask a per-hand asymmetry (Fig. 4b): single-hand baselines collapse the low-quality hand to a generic prior,

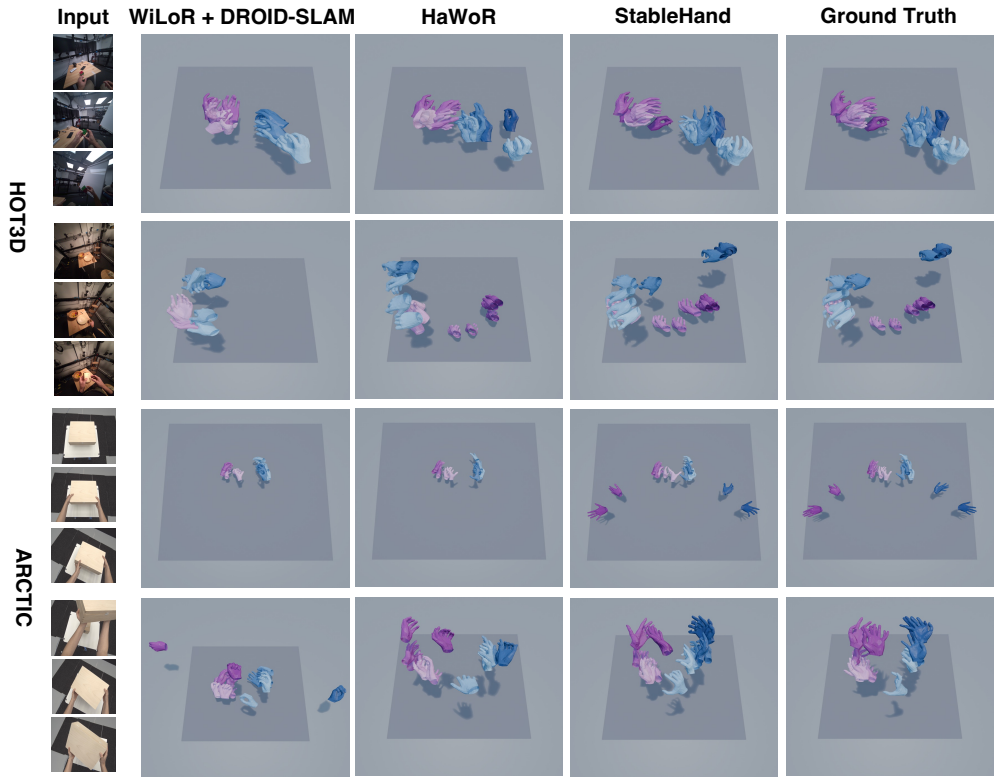
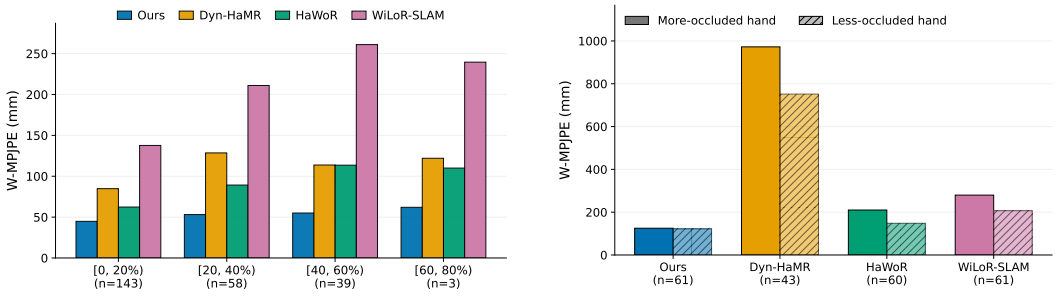


Figure 3: **Qualitative comparison on HOT3D (top two rows, long missing-hand spans) and ARCTIC (bottom two rows, persistent hand-object occlusion).** Each row shows three input frames and the world space dual-hand mesh trajectory of WiLoR [34]+DROID-SLAM [43], HaWoR [51], our StableHand, and Ground Truth (left hand magenta, right hand blue, with mesh shading dark→light encoding temporal order). StableHand preserves coherent trajectories and re-anchors when observations resume, while WiLoR+DROID-SLAM drifts and HaWoR over-smooths or collapses the occluded hand to a generic prior. Best viewed in the supplementary video.



(a) HOT3D, by per-clip missing-hand fraction

(b) ARCTIC, per-hand under detector failure

Figure 4: **Stratified evaluations.** (a) On HOT3D our W-MPJPE is lowest in every bin, and the gap to WiLoR-SLAM [34], Dyn-HaMR [49], and HaWoR [51] is largest in the high-missing regime. The lower baseline values in the 60–80% bin reflect small-sample statistics (per-bin clip counts annotated above each bar). (b) On ARCTIC, single-hand baselines exhibit a substantial W-MPJPE gap between the more- and less-occluded hand. Our model closes most of this gap, with a small residual difference reflecting that the less-occluded hand still carries cleaner observations.

while our model closes the per-hand gap within measurement noise via shared self-attention over all $4T$ tokens. Fig. 3 (bottom two rows) visualizes this across long contact phases. Additional qualitative comparisons and failure cases for both benchmarks are provided in the supplemental material.

Table 2: **World-space hand motion estimation on ARCTIC [6].** The upper block reports SLAM-based methods. The lower block reports native world space methods, retrained on the ARCTIC training split. Per-column **best** and second best are color-coded.

Method	PA-MPJPE [mm] (↓)	W-MPJPE [mm] (↓)	WA-MPJPE [mm] (↓)	AccEr [m/s^2] (↓)	MRRPE [mm] (↓)
HaMeR-SLAM [31]	10.99	232.48	59.35	10.82	163.46
WiLoR-SLAM [34]	8.92	183.72	50.03	8.80	122.42
HMP-SLAM [5]	13.73	166.21	41.62	7.13	131.33
Dyn-HaMR [49]	13.36	734.31	83.63	6.05	189.09
HaWoR [51]	10.30	171.61	51.98	11.35	112.36
Ours	**8.07**	**124.59**	**32.66**	**4.67**	**93.71**

4.3 Ablation Study

We ablate StableHand on HOT3D (Tab. 3) along (a) the quality network, (b) design choices of the quality-aware flow-matching pathway, and (c) the quality signal consumed at inference.

Table 3: **Ablation studies on HOT3D.** Three axes: (a) quality network, (b) generative model design, and (c) quality source at inference. Per-column **best** and second best are color-coded within each block. Per-corpus QN breakdown is in the supplemental material.

Setup	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	AccEr ↓
(a) Quality network				
Ours (full)	**4.02**	**57.83**	**21.02**	3.83
QN pretrained on HOT3D only	4.85	61.54	21.12	4.06
QN w/o temporal self-attention (per-frame MLP)	4.33	64.98	25.71	**3.82**
(b) Generative model design				
Ours (full)	4.02	**57.83**	**21.02**	3.83
w/o quality schedule ($\tilde{\mathbf{q}}\equiv\mathbf{0}$, standard FM)	4.55	92.32	30.22	4.87
w/o per-component $\mathbf{q}$ (single scalar q^h per hand)	5.06	61.23	22.37	4.89
w/o per-token AdaLN (single global $\mathbf{m}_t$)	4.44	64.66	23.28	**3.23**
w/o smoothness regularizer ($\lambda_{\text{smooth}}=0$)	4.62	63.57	23.26	3.60
w/o wrist-translation term ($\lambda_{\text{wrist}}=0$)	**3.95**	70.52	22.45	4.31
w/o four-token split (single per-frame token)	5.23	118.61	40.03	6.60
(c) Quality source at inference				
$\mathbf{q}$ = oracle per-component (upper bound)	**3.22**	**48.16**	**17.58**	**3.12**
$\mathbf{q} = \hat{\mathbf{q}}$ from QN (**Ours (full)**)	4.02	57.83	21.02	3.83
$\mathbf{q}$ = per-corpus median (constant baseline)	5.43	104.62	31.37	4.25
$\mathbf{q}$ = binary detection mask	6.19	135.35	43.82	9.39

Quality network (Tab. 3(a)). Restricting QN pretraining to HOT3D alone lifts W-MPJPE by 3.7 mm to 61.54, confirming that cross-corpus pretraining broadens failure-mode coverage. Replacing the temporal self-attention encoder with a per-frame MLP further lifts it to 64.98 (+7.2 mm), confirming that temporal context distinguishes sustained detection gaps from isolated low-confidence frames.

Generative model design (Tab. 3(b)). Setting $\tilde{\mathbf{q}}\equiv\mathbf{0}$ collapses Eq. 8 to a scalar schedule and lifts W-MPJPE to 92.32 (1.6×): without per-channel quality, the process regenerates every channel from noise instead of anchoring high-quality observations. Replacing the four tokens per frame with a single per-frame token, which strips AdaLN of any per-channel modulation, is the most damaging entry at 118.61 (2.1×), confirming that the four-token split is necessary for AdaLN to scale wrist and finger updates separately.

Quality source at inference (Tab. 3(c)). Replacing $\hat{\mathbf{q}}$ with a per-corpus median or a binary detection mask lifts W-MPJPE to 104.62 and 135.35, confirming that graded per-component quality cannot be approximated by a constant or detection threshold. The oracle $\mathbf{q}$ at 48.16 (−9.7 mm) sits only modestly below our predicted $\hat{\mathbf{q}}$, indicating that the quality network captures most of the available signal. The Wasserstein-loss ablation, RBF bandwidth sensitivity, QN calibration and input-perturbation analyses, and inference efficiency are in the supplemental material.

5 Conclusion

We presented StableHand, a quality-aware flow-matching framework for world space 4D dual-hand motion estimation, in which a four-value quality signal drives the forward schedule, velocity target, ODE initialization, and AdaLN modulation of a DiT. StableHand achieves state-of-the-art across every reported metric on HOT3D and ARCTIC, with the largest margins on the most occlusion-affected

clips. The quality signal is calibrated to a specific upstream estimator, and over long low-quality horizons the model may drift toward plausible but not faithfully grounded configurations. Joint object–hand modeling and end-to-end training are natural next steps.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024.

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019.

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12417–12426, 2021.

[5] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6353–6363, 2024.

[6] 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*, pages 12943–12954, 2023.

[7] Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Rolandos Alexandros Potamias, Taku Komura, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos. *arXiv preprint arXiv:2601.01050*, 2026.

[8] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23600–23611, 2023.

[9] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[10] Irmak Guzey, Haozhi Qi, Julen Urain, Changhao Wang, Jessica Yin, Krishna Bodduluri, Mike Lambeta, Lerrel Pinto, Akshara Rai, Jitendra Malik, et al. Dexterity from smart lenses: Multi-fingered robot manipulation with in-the-wild human demonstrations. *arXiv preprint arXiv:2511.16661*, 2025.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[14] 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)*, pages 13226–13233. IEEE, 2025.

[15] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[16] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6991–7003, 2025.

[17] 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.

[18] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021.

[20] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021.

[21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[22] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14687–14697, 2021.

[23] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[26] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[27] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020.

[28] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *Advances in Neural Information Processing Systems*, 36:17689–17701, 2023.

[29] Yuxuan Mu, Hung Yu Ling, Yi Shi, Ismael Baira Ojeda, Pengcheng Xi, Chang Shu, Fabio Zinno, and Xue Bin Peng. Stablemotion: Training motion cleanup models with unpaired corrupted data. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

[30] 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*, pages 1496–1505, 2022.

[31] 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*, pages 9826–9836, 2024.

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[33] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913, 2025.

[34] 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 Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025.

[35] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *European Conference on Computer Vision*, pages 183–202. Springer, 2024.

[36] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

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

[38] Subham S Sahoo, Aaron Gokaslan, Chris De, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. *Advances in Neural Information Processing Systems*, 37:105730–105779, 2024.

[39] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[40] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[41] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[42] Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. Unihand: A unified model for diverse controlled 4d hand motion modeling. *arXiv preprint arXiv:2602.21631*, 2026.

[43] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[44] Eugene Valassakis and Guillermo Garcia-Hernando. Handdgp: Camera-space hand mesh prediction with differentiable global positioning. In *European Conference on Computer Vision*, pages 479–496. Springer, 2024.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] 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.

[47] 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*, pages 21222–21232, 2023.

[48] Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *arXiv preprint arXiv:2602.22209*, 2026.

[49] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27716–27726, 2025.

[50] Huajian Zeng, Lingyun Chen, Jiaqi Yang, Yuantai Zhang, Fan Shi, Peidong Liu, and Xingxing Zuo. Flowhoi: Flow-based semantics-grounded generation of hand-object interactions for dexterous robot manipulation. *arXiv preprint arXiv:2602.13444*, 2026.

[51] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1805–1815, 2025.

[52] 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.

[53] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

[54] Lawrence Y Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. Emma: Scaling mobile manipulation via egocentric human data. *IEEE Robotics and Automation Letters*, 2026.

Supplementary Material

This supplementary document provides additional details and results that complement the main paper. Sec. A summarizes the full architecture and training hyperparameters. Sec. B reports additional robustness and calibration analyses spanning predicted-quality calibration against ground truth, quality-network input perturbations, and estimator generality across upstream pose estimators. Sec. C reports supplemental ablation studies covering the radial-basis bandwidth (Sec. C.1) and the quality network architecture (Sec. C.2). Sec. D reports inference efficiency including a per-stage runtime breakdown and an ODE-step sweep. Sec. E specifies the synthetic perturbation taxonomy used to augment the quality-network training distribution. Sec. F details the eight egocentric corpora used to pretrain the quality network. Sec. G provides formal definitions of the evaluation metrics. Sec. H concludes with additional qualitative results and failure-case analyses.

A Implementation Details

The DiT denoiser stacks 8 blocks of hidden dimension 512 with 8 attention heads and a $4\times$ GeGLU [40] feed-forward expansion. Each frame is tokenized into four tokens (left/right wrist, left/right fingers) and temporal RoPE [41] self-attention spans all $4T$ tokens. The quality network is a 4-layer Transformer encoder of hidden dimension 256 and 4 attention heads with camera-state cross-attention, and shares weights between the two hands. The total parameter count is 70.7M (DiT 66.8M + QN 3.6M + auxiliary heads 0.3M).

Both networks are trained on a single NVIDIA H100 GPU under bfloat16 autocast with AdamW [24] (learning rate 1×10^{-4}, weight decay 0.01, batch size 16, 1000-step linear warmup followed by cosine decay, gradient clipping $\|\nabla\|\leq1.0$, seed 42). A single benchmark reaches early-stop on *eval/W-MPJPE* in 6–8 H100-hours. The 20-step Euler ODE solve through the DiT denoiser consumes $\sim$230 ms per clip. The full pipeline including upstream estimators is detailed in Sec. D. The pose estimator and geometry model are the released WiLoR [34] and Depth-Anything-V3 [18] checkpoints, with no fine-tuning.

B Additional Quality-Network and Estimator Analyses

This section reports three robustness and calibration analyses that complement the ablation study of the main paper. Sec. B.1 reports a per-channel calibration analysis of the predicted quality signal against ground truth. Sec. B.2 probes the quality network's response to controlled noise on each of its input streams and the corresponding downstream impact. Sec. B.3 examines whether the framework transfers to a different upstream pose estimator (HaMeR). Sec. B.2 and Sec. B.3 report the 16-joint MANO MPJPE for direct comparison with Tab. 1 of the main paper, while Sec. B.1 color-codes its scatter by per-frame wrist-joint W-MPJPE for visual correspondence with the wrist quality channel.

B.1 Predicted-Quality Calibration

Fig. 5 reports a per-channel calibration analysis of the predicted quality signal against ground truth on the HOT3D test split, since the QN is the central conditioning signal of the generative model and its calibration directly bounds downstream behavior. For each of the four quality channels $(q_W^L, q_F^L, q_W^R, q_F^R)$, we plot a scatter of (predicted, ground-truth) pairs across all valid frames, color-coded by the per-frame downstream wrist-joint W-MPJPE, with an inset 4×4 confusion matrix that quantizes both axes into four bins. The Spearman rank correlation $\rho(\hat{\mathbf{q}}, \mathbf{q})$ is 0.734 for the left finger channel, 0.735 for the right finger channel, 0.570 for the left wrist channel, and 0.502 for the right wrist channel. Wrist channels are systematically harder to calibrate because wrist global trajectory error involves both translation and rotation drift in the world frame, while finger error is confined to the local articulation manifold.

Off-diagonal frames in the four-bin grid (where the predicted-q bin differs from the ground-truth bin) account for 22% to 44% of frames depending on the channel. The mean downstream wrist-joint W-MPJPE on these off-diagonal frames is within 2 to 3 mm of the on-diagonal mean, indicating that QN miscalibration is bounded in its downstream impact: even when the QN occasionally over- or

under-confidently classifies a frame, the generative process degrades gracefully rather than collapsing to the prior or anchoring to a noisy observation.

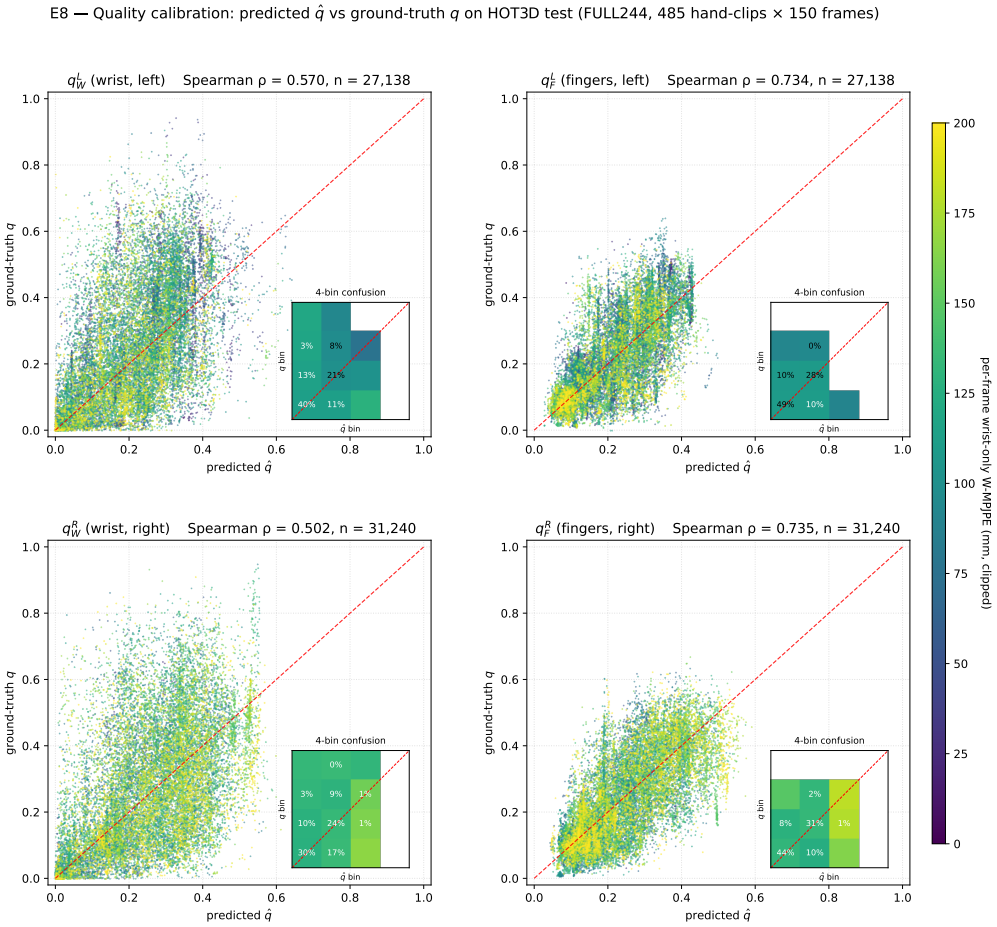


Figure 5: **Predicted-quality calibration on HOT3D test.** Each panel corresponds to one of the four quality channels. Each scatter point is one (frame, hand) pair, colored by the per-frame wrist-joint W-MPJPE (viridis colormap). The diagonal line marks perfect calibration, and the inset 4×4 matrix quantizes both axes into four bins (V-Low, Low, Mid, High) with cell percentages. Spearman rank correlations are reported per channel.

B.2 Quality-Network Input Perturbation

We probe the quality network's downstream value by injecting controlled noise into each of its input streams and measuring both the QN-side error degradation and the wrist trajectory response. Each of the three QN inputs (the per-hand observation $\bar{\mathbf{y}}$, the binary detection flag δ, and the per-frame camera pose $\mathbf{g}_t$) is perturbed at four severity levels covering deployment-realistic to extreme noise, with all other streams held clean. Inference uses the frozen quality network and frozen flow-matching denoiser without retraining, recording the change in predicted error $\Delta\hat{\mathbf{e}}$ and the change in downstream W-MPJPE relative to the unperturbed baseline, which matches the corresponding row of Tab. 1 to within rounding. For context, we further reference the oracle-q upper bound and constant-q lower bound from Tab. 3(c), obtained from separately trained DiTs using ground-truth and per-corpus-median quality at both training and inference.

Tab. 4 reports the full sweep, and Fig. 6 summarizes the same data graphically. At zero perturbation the predicted-q baseline of 57.83 mm sits between the oracle-q upper bound of 48.16 mm and the constant-q lower bound of 104.62 mm from Tab. 3(c). This recovers 83% of the gap an explicit per-frame quality signal can close relative to a constant-q ablation, leaving a 9.7 mm residual to the

Table 4: **Quality-network input perturbation sweep on HOT3D test.** $\bar{\hat{e}}$ denotes the mean predicted error magnitude across all components, and W-MPJPE is the downstream world space MPJPE under the 16-joint MANO MPJPE convention of Tab. 1. Δ columns report the change relative to the predicted-q baseline at zero perturbation.

Stream	**Level**	$\bar{\hat{e}}$ (mm)	$\Delta\hat{e}$ (mm)	W-MPJPE (mm)	ΔW (mm)
Reference anchors from Tab. 3(c) (separately trained DiTs)					
oracle-q (upper bound)	—	—	—	48.16	—
constant-q (lower bound)	—	—	—	104.62	—
Predicted-q (Ours) baseline + perturbation sweep					
baseline (no perturbation)	0	27.36	+0.00	**57.83**	+0.00
$\bar{\mathbf{y}}$ trans (mm)	1	27.36	−0.00	57.06	−0.77
	5	27.36	+0.01	57.17	−0.66
	10	27.39	+0.03	57.58	−0.25
	30	27.66	+0.30	59.62	+1.79
$\bar{\mathbf{y}}$ rot (deg)	0.5	27.35	−0.00	58.56	+0.73
	1	27.36	−0.00	57.20	−0.63
	2	27.37	+0.01	57.45	−0.38
	5	27.46	+0.10	57.57	−0.26
$\bar{\mathbf{y}}$ AA (rad)	0.01	27.39	+0.03	57.64	−0.19
	0.05	28.14	+0.79	58.57	+0.74
	0.1	30.08	+2.72	65.55	+7.72
	0.2	35.35	+8.00	**99.15**	+41.32
δ flip rate	0.05	27.79	+0.43	57.67	−0.16
	0.1	28.27	+0.92	57.98	+0.15
	0.2	29.09	+1.73	58.22	+0.39
	0.5	31.08	+3.72	60.90	+3.07
$\mathbf{g}_t$ trans (mm)	1	27.36	−0.00	57.77	−0.06
	5	27.36	+0.00	57.67	−0.16
	10	27.36	+0.00	57.02	−0.81
	30	27.40	+0.04	59.64	+1.81
$\mathbf{g}_t$ rot (deg)	0.5	27.36	+0.00	56.88	−0.95
	1	27.36	+0.01	56.89	−0.94
	2	27.39	+0.04	57.56	−0.27
	5	27.59	+0.24	**63.34**	+5.51
combined moderate	all level-2	29.19	+1.84	59.38	+1.55

oracle that reflects QN prediction error rather than an architectural limitation of the quality-aware pathway.

Across all but the most extreme perturbation level of each input stream in Tab. 4, downstream W-MPJPE stays within 1 mm of the unperturbed baseline.

At extreme noise the largest sensitivity is finger axis-angle perturbation at 0.2 rad ($+41.3$ mm), which scrambles articulation that the DiT cannot recover. Camera rotation at 5° ($+5.5$ mm) feeds the DiT cross-attention directly and similarly degrades the inference-swap oracle-q reference line in Fig. 6, indicating a DiT-side bottleneck rather than a QN failure. This experiment subsumes the controllably-corrupted-$\hat{\mathbf{q}}$ probe, since every realistic $\hat{\mathbf{q}}$ corruption arises from upstream input noise and the input-side perturbation curves bound it from above.

B.3 Estimator Generality on HaMeR Observations

The framework predicts the per-component error in physical units (mm) from a generic MANO observation, a binary detection flag, and a head-camera pose, none of which are estimator-specific, so the quality-aware pipeline transfers in principle to any pose estimator that produces MANO predictions. We empirically probe this generality by comparing the per-component error distributions of WiLoR and HaMeR on the same HOT3D test data, by running drop-in inference with HaMeR

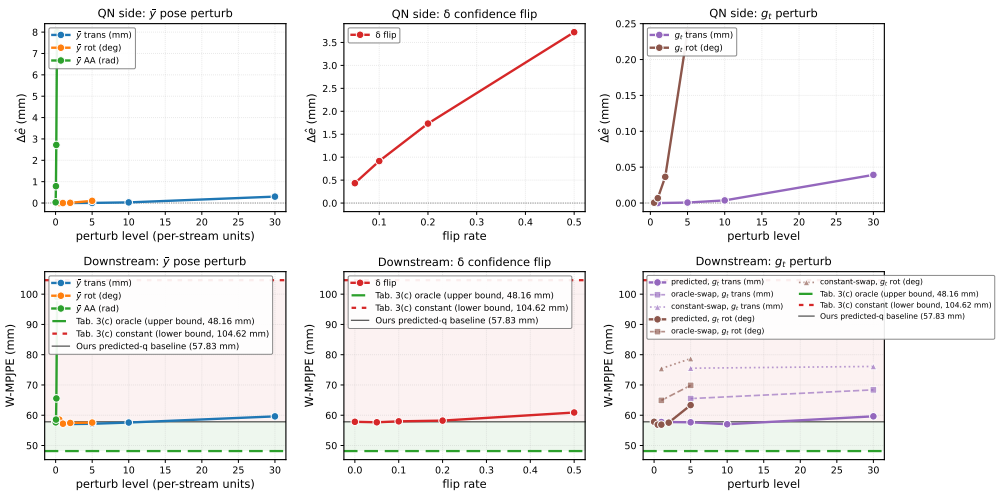


Figure 6: **Quality-network input perturbation sweep.** Top row: change in predicted error $\Delta\hat{\mathrm{e}}$ (mm) under controlled perturbation of each QN input stream; bottom row: downstream W-MPJPE (mm) under the same perturbations, with horizontal references for the predicted-q baseline (57.83 mm), the oracle-q upper bound (48.16 mm), and the constant-q lower bound (104.62 mm) from Tab. 3(c). Columns correspond to the per-hand observation $\bar{\mathbf{y}}$, the detection flag δ, and the camera pose $\mathbf{g}_t$, with sub-streams plotted within each column. The $\mathbf{g}_t$ panel additionally overlays oracle-q and constant-q reference curves to disentangle DiT-side from QN-side bottlenecks under camera-pose noise. See Tab. 4 for the full numerical sweep.

observations under two calibration settings (zero-shot and bandwidth-recalibrated), and by retraining the entire pipeline end-to-end on HaMeR observations.

Fig. 7 overlays the per-component error histograms produced by the two estimators on 15 HOT3D test clips with both estimator caches available, totaling roughly 3,125 wrist samples and 3,125 finger samples. Both estimators produce qualitatively similar long-tailed distributions on both components. The corpus-adaptive radial-basis bandwidth resolves to $\hat{\sigma}_W = 24.95$ mm and $\hat{\sigma}_F = 4.32$ mm for WiLoR, and to $\hat{\sigma}_W = 28.68$ mm and $\hat{\sigma}_F = 4.33$ mm for HaMeR. The slightly larger $\hat{\sigma}_W$ on HaMeR tracks its slightly higher 80th-percentile wrist error.

We evaluate four configurations of upstream-estimator handling on the full HOT3D test split (Tab. 5), reporting the standard 16-joint MANO MPJPE for direct comparison with Tab. 1. Row 1 anchors the WiLoR-trained baseline, rows 2 and 3 isolate zero-shot drop-in and bandwidth-recalibrated drop-in with HaMeR observations at inference, and row 4 retrains the entire pipeline (DiT + QN + adaptive $\hat{\sigma}$) end-to-end on HaMeR observations.

Table 5: **HaMeR drop-in and full-retrain evaluation on HOT3D.** Same Ours pipeline architecture, four configurations of the upstream observation source and training scheme: WiLoR-trained baseline, HaMeR zero-shot at inference, HaMeR with bandwidth recalibration, and HaMeR end-to-end retraining. Numbers are reported on the full HOT3D test set (244 clips, n=485 hand-clips) under the 16-joint MANO MPJPE convention of Tab. 1.

Method	W-MPJPE (mm)	WA-MPJPE (mm)	PA-MPJPE (mm)	AccEr (m/s^2)
Ours (WiLoR)	57.83	19.90	3.65	3.53
Ours (HaMeR, zero-shot)	104.81 (+46.98)	38.50	5.83	4.10
Ours (HaMeR, $\hat{\sigma}$-recalib.)	103.84 (+46.01)	38.49	5.81	4.09
Ours (HaMeR, full retrain)	62.16 (+4.33)	21.15	3.71	**3.49**

The HaMeR zero-shot drop-in degrades W-MPJPE by approximately 47 mm relative to the WiLoR baseline, an expected gap because the deployable pipeline was trained against WiLoR-distributed observations. Recomputing $\hat{\sigma}$ on HaMeR data changes the downstream W-MPJPE by only -0.97 mm (well within ± 1 mm), indicating that the bandwidth-adaptation mechanism operates in a flat regime in

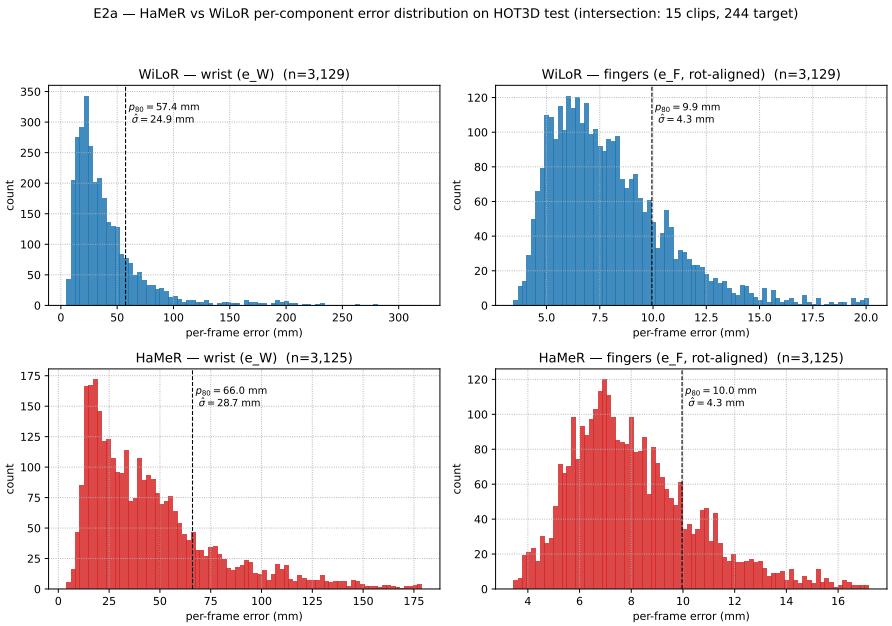


Figure 7: **Per-component error distributions on HOT3D under two upstream pose estimators.** Histograms of e_W (wrist error) and e_F (finger MPJPE in the wrist frame) for WiLoR and HaMeR, computed on the same 15 HOT3D test clips with both estimators cached. Vertical lines mark the 80th-percentile error per estimator.

this range and is robust to estimator-specific bandwidth shifts but does not by itself close the train-test gap.

Retraining the entire pipeline end-to-end on HaMeR observations closes 91% of the zero-shot gap, bringing W-MPJPE to 62.16 mm (only +4.33 mm above the WiLoR baseline). The remaining 4.33 mm gap reflects a regressor-quality difference rather than a framework limitation: HaMeR's training corpus does not include Aria-style egocentric imagery, while WiLoR's training corpus includes H2O [15] which provides comparable egocentric supervision. PA-MPJPE is essentially unchanged under HaMeR retraining (+0.06 mm, within noise), confirming that the gap is concentrated in wrist absolute positioning rather than finger articulation. AccEr is slightly better under HaMeR retraining (-0.04 m/s^2), suggesting comparable temporal smoothness.

C Supplemental Ablation Studies

This section reports a complementary ablation analysis to the main-paper studies of Sec. 4.3. The RBF bandwidth sensitivity in Sec. C.1 sweeps fixed σ around our adaptive calibration of Eq. 3, and the quality-network ablation in Sec. C.2 reports per-corpus Spearman correlation and aggregate metrics across architectural and corpus variants.

C.1 RBF Bandwidth Sensitivity

Tab. 6 sweeps the RBF bandwidth σ around the adaptive value $\sigma_\star = p_{80}(e_\star)/\ln 10$ used in the main paper. Each row retrains the generative model with the labeled σ applied to Eqs. 1–2, and uses the oracle per-component $\mathbf{q}$ at inference under the same σ to isolate the generative-model response to bandwidth choice from quality-network prediction.

W-MPJPE degrades monotonically as σ grows. The adaptive $\sigma_\star$ on HOT3D resolves to $\sigma_W \approx 26.2$ mm and $\sigma_F \approx 6.4$ mm, close to the fixed $\sigma = 30$ mm row but with a per-component split that the fixed sweep cannot reproduce. Larger σ flattens $\exp(-e/\sigma)$ toward 1 and erodes the contrast between high- and low-quality frames, removing the gradient that the per-channel forward schedule relies on. AccEr behaves non-monotonically: a moderately wider $\sigma = 30$ mm attains the lowest AccEr (3.31 vs. 3.83 m/s^2 for the adaptive $\sigma_\star$) at the cost of a 7.7 mm rise in W-MPJPE. This mirrors the multi-metric trade-off in the ODE-step sweep of Sec. D, where joint-position accuracy and trajectory smoothness

Table 6: **Sensitivity to the RBF bandwidth σ on HOT3D.** Each row retrains the generative model with the labeled σ applied to Eqs. 1–2. Inference uses the oracle per-component $\mathbf{q}$ computed under the same σ, isolating the generative-model response to bandwidth choice from quality-network prediction. Per-column **best** and second best are color-coded.

Setup	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	AccEr ↓
σ = adaptive $p_{80}/\ln 10$ (Ours)	**4.02**	**57.83**	**21.02**	3.83
σ = fixed 30 mm	4.54	65.52	23.36	**3.31**
σ = fixed 50 mm	5.79	87.51	32.02	3.44
σ = fixed 80 mm	6.67	118.70	41.56	3.43
σ = fixed 150 mm	6.86	128.04	45.48	4.40

Table 7: **Quality-network ablation evaluated by per-corpus Spearman correlation $\rho(\hat{\mathbf{q}}, \mathbf{q})$ and aggregate operational metrics.** Per-corpus ρ measures rank agreement between predicted and ground-truth per-component quality on each named held-out test split. ρ=1 matches the ground-truth ranking and ρ=0 is random. Aggregate metrics report q-MAE (mean absolute error of $\hat{\mathbf{q}}$ against the ground-truth $\mathbf{q}$ on the 8-corpus mixed validation), mm-MAE (best $\hat{\mathbf{e}}$ MAE on each ckpt's own training-time validation set, in mm), and gap-closed (the QN-vs-random fraction of the random-to-oracle quality-rejection AUC gap on the mixed validation). (a) Ablates the input cues, temporal architecture, and training loss, holding the pretraining corpus fixed. The Ours and *w/o Wasserstein loss* rows are evaluated under the v26 protocol while the architectural variants retain their original v22 measurements (cross-protocol comparison on a few cells is therefore loose). (b) Ablates the pretraining corpus, holding the architecture and training loss fixed. The mm-MAE column uses different validation sets for the two rows (HOT3D-only val for the single-corpus row, mixed val for the multi-corpus row), so cross-row comparison on this column is loose. Per-column **best** and second best within each block are color-coded.

	Per-corpus Spearman ρ ↑						Aggregate		
Setup	HOT3D	H2O	HOI4D	Re:InterHand	ARCTIC	Avg	q-MAE ↓	mm-MAE (mm) ↓	gap-closed ↑
(a) Quality-network architecture and training loss									
Full QN (Ours)	0.846	**0.906**	**0.859**	0.912	**0.775**	**0.862**	0.096	12.55	**+98.5%**
+ visual feature $\mathbf{f}_t^h$ (no exclusion)	0.775	0.855	0.814	0.883	0.649	0.825	0.120	24.07	+86.9%
w/o camera-pose input $\mathbf{g}_t$	0.797	0.857	0.801	0.880	0.667	0.836	0.116	27.90	+88.5%
w/o detection flag δ_t^h	**0.847**	0.887	0.834	**0.917**	0.613	0.859	0.104	18.64	+89.6%
per-frame MLP (no temporal self-attention)	0.819	0.857	0.830	0.892	0.697	0.838	**0.078**	22.10	+89.6%
w/o Wasserstein loss (MSE-only training)	0.831	0.880	0.840	0.907	0.754	0.844	0.096	**11.04**	+91.7%
(b) Pretraining corpus									
QN pretrained on HOT3D only	**0.850**	0.707	0.766	0.704	0.623	0.757	0.162	**9.48**	+83.9%
QN pretrained on full multi-corpus mixture (Ours)	0.846	**0.906**	**0.859**	**0.912**	**0.775**	**0.862**	**0.096**	12.55	**+98.5%**

pull against each other and motivate selecting $\sigma_\star$ on the joint-position metric while reporting both. Each variant trains and evaluates under the same σ, matching the deployment scenario but conflating the model's representational adaptation to the bandwidth with the change in $\mathbf{q}$ distribution it induces.

C.2 Quality Network Ablation

Tab. 3(a) of the main paper isolates the impact of two quality-network design choices on downstream world space hand recovery. This subsection drills further into the quality network's own predictive fidelity, evaluated by the Spearman rank correlation $\rho(\hat{\mathbf{q}}, \mathbf{q})$ between predicted and ground-truth per-component quality on held-out test splits of multiple egocentric corpora (Tab. 7).

Architecture and training loss (Block (a)). Each row of Tab. 7(a) ablates one design choice of the quality network. Adding the frozen WiLoR visual feature $\mathbf{f}_t^h$ back to the per-hand input token degrades cross-corpus average ρ by 0.037 and lifts the validation $\hat{\mathbf{e}}$ MAE from 12.55 mm to 24.07 mm. Even under the dataset-invariant log-error target, the 1280-dimensional ViT feature inflates input redundancy and dilutes the optimization signal rather than supplying a complementary cue. Removing the camera-pose input $\mathbf{g}_t$ costs 0.026 average ρ, with consistent drops of 0.05 to 0.10 on every in-train corpus, confirming that head-motion context contributes broadly to per-component error prediction across egocentric streams. Removing the binary hand-confidence flag δ_t^h costs only 0.003 average ρ, indicating that the flag is a near-redundant input recoverable from the proposal vector $\bar{\mathbf{y}}_t^h$ and its temporal context. Replacing the temporal self-attention encoder with a per-frame MLP raises the validation $\hat{\mathbf{e}}$ MAE from 12.55 mm to 22.10 mm and trims average ρ by 0.024, leaving rank correlation almost unchanged but losing the magnitude-refinement role of cross-frame attention. The per-frame MLP variant retains a competitive q-MAE (0.078 vs. 0.096), suggesting that the leaner predictor

Table 8: **Inference efficiency on HOT3D** (150-frame clips, single H100). We report wall-clock time per clip and peak GPU memory. Optimization-based baselines are an order of magnitude slower than feed-forward generative models. StableHand stays in the feed-forward envelope despite solving a 20-step ODE. Per-column **best** and second best are color-coded.

Method	Time / clip (s) ↓	Peak GPU mem (GB) ↓
HaMeR-SLAM [31]	125.91	12.50
WiLoR-SLAM [34]	30.12	12.50
Dyn-HaMR [49]	347.84	**9.82**
HaWoR [51]	28.45	14.56
StableHand (Ours)	**27.29**	10.32

overfits less to in-train corpora at the cost of magnitude refinement. Removing the Wasserstein loss of Eq. 7 (MSE-only training) lowers the validation $\hat{\mathbf{e}}$ MAE marginally to 11.04 mm but collapses the operational gap-closed metric from $+98.5\%$ to $+91.7\%$ and the cross-corpus average ρ from 0.862 to 0.844. This isolates the Wasserstein term's role: it trades a small per-sample magnitude regression for a large gain in distribution alignment, directly driving the deployment-time match between predicted and oracle σ. Across the architectural ablations, every input or encoder choice contributes between 0.003 and 0.037 average ρ, all dominated by the 0.105 swing between the multi-corpus and single-corpus pretraining of Block (b).

Pretraining corpus (Block (b)). A single-corpus quality network pretrained on HOT3D alone marginally beats the multi-corpus model on its own training distribution (ρ=0.850 vs. 0.846 on HOT3D). On every other corpus the single-corpus model degrades, with $\rho \in [0.704, 0.766]$ on H2O, HOI4D, and Re:InterHand and ρ=0.623 on the zero-shot ARCTIC test split. The multi-corpus model lifts these out-of-distribution cells to $\rho \geq 0.859$ on H2O, HOI4D, and Re:InterHand and to 0.775 on ARCTIC, raising the cross-corpus average from 0.757 to 0.862. Calibration degrades more sharply than rank: the single-corpus q-MAE averages 0.162 against the multi-corpus model's 0.096, and the operational gap-closed metric drops from $+98.5\%$ to $+83.9\%$. Broad pretraining is essential for the quality network to produce $\hat{\mathbf{q}}$ that is both well-ranked and well-calibrated on out-of-distribution egocentric corpora at deployment time.

D Inference Efficiency

Tab. 8 reports wall-clock time per 150-frame clip on a single H100 GPU. Optimization-based baselines (HaWoR, Dyn-HaMR) require per-clip SLAM and iterative refinement, while StableHand runs a single quality-network forward pass followed by a 20-step Euler ODE solve with no per-clip optimization loop. The runtime gap between StableHand and SLAM-based baselines is driven primarily by the choice of geometry foundation model (DA3 forward pass vs Droid-SLAM optimization) rather than by the quality-aware flow-matching pathway itself, and we report it here for completeness rather than as a contribution of this work.

Per-stage breakdown. Tab. 9 dissects the StableHand wall-clock cost into its four stages. The pipeline cost is dominated by the geometry foundation model (Depth-Anything-V3, 17.09 s per clip) and the per-frame WiLoR estimator (9.96 s per clip), while the quality network (<0.01 s) and the 20-step Euler ODE solve through the DiT denoiser (0.23 s) together account for less than 1% of the total wall-clock time. The quality-aware flow-matching pathway is essentially free relative to the upstream visual stack.

Table 9: **Per-stage inference-time breakdown on a single H100 GPU.** Average wall-clock time per T=150 clip, averaged over 10 HOT3D clips with batch size 1. Per-stage values are proportionally aligned to the wall-clock total of Tab. 8.

Stage	Time per clip (s)	Time per frame (ms)
Depth-Anything-V3 [18] (geometry, 1 call)	17.09	113.93
WiLoR [34] estimator ($\times$150 frames)	9.96	66.40
DiT $\times$ 20 ODE steps	0.23	1.51
Quality network (1 forward over T)	<0.01	<0.10
Total (full pipeline)	**27.29**	**181.93**

ODE-step sweep. Fig. 8 sweeps the Euler ODE step count $n_{\text{steps}} \in \{2, 5, 10, 20, 30, 50\}$ on HOT3D test, evaluated jointly on trajectory accuracy (W-MPJPE) and trajectory smoothness (AccEr). The two metrics behave oppositely with n_{steps}: W-MPJPE varies by under 9% across the entire sweep (54.04 mm at $n{=}2$ to 58.81 mm at $n{=}50$), while AccEr drops nearly $4\times$ from 13.97 at $n{=}2$ to 3.53 m/s^2 at $n{=}20$ as the dense Euler integration replaces stair-step trajectory reconstruction with a smooth one. Most of the AccEr improvement occurs by $n{=}20$, with diminishing returns thereafter (3.34 and 2.91 m/s^2 at $n{=}30$ and $n{=}50$ at $1.5\times$ and $2.5\times$ the DiT compute respectively). We adopt $n{=}20$ as the deployable default since temporal smoothness is largely converged at this setting and the marginal gain from higher step counts does not justify the extra cost, while the remaining real-time bottleneck lies in the upstream pose estimator and geometry foundation model rather than in the quality-aware flow-matching pathway proposed in this work.

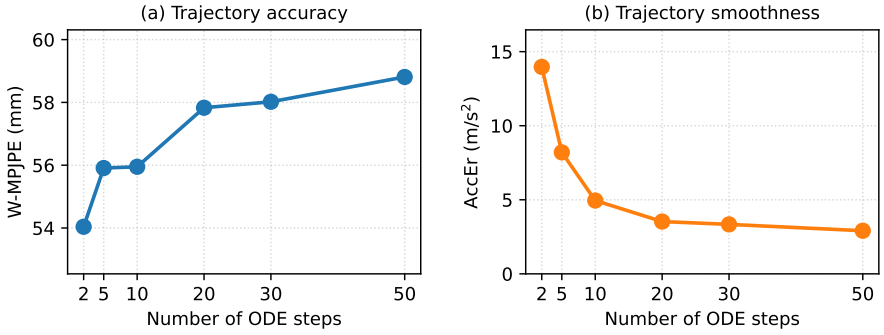


Figure 8: **ODE-step sweep on HOT3D test, multi-metric view.** $n_{\text{steps}} \in \{2, 5, 10, 20, 30, 50\}$ evaluated on (a) trajectory accuracy (W-MPJPE) and (b) trajectory smoothness (AccEr), with the $n{=}20$ row anchored to the main paper Tab. 1. Lower step counts attain marginally lower W-MPJPE (under 9% range) but degrade AccEr by nearly $4\times$ at $n{=}2$, motivating $n{=}20$ as the deployable default where AccEr is largely converged.

E Synthetic Perturbation Taxonomy

This section formalizes the quality-consistent observation augmentation introduced in Sec. 3.1. We define the augmentation as a stochastic operator on $(\bar{\mathbf{y}}_t^h, \mathbf{q}_t^h)$ pairs, derive the joint update rule that preserves the calibration of Eq. 3, and describe each of the five perturbation types together with the upstream-stream failure mode it imitates.

Augmentation operator. Let $\mathcal{A}$ denote the augmentation operator, applied independently per hand $h \in \{L, R\}$ at every training step. With probability $\frac{1}{2}$, $\mathcal{A}$ leaves the input unchanged. Otherwise, $\mathcal{A}$ samples a perturbation type uniformly from a five-element pool and applies it to a designated channel subset of the per-hand observation $\bar{\mathbf{y}}_t^h \in \mathbb{R}^{54}$, partitioned as $\bar{\mathbf{y}}_t^h = [\bar{\boldsymbol{\theta}}_{\text{wrist}}^h;\ \bar{\boldsymbol{\theta}}_{\text{fingers}}^h;\ \bar{\mathbf{p}}^h]$ with $\bar{\boldsymbol{\theta}}_{\text{wrist}}^h \in \mathbb{R}^6$ (rot6d), $\bar{\boldsymbol{\theta}}_{\text{fingers}}^h \in \mathbb{R}^{45}$ (finger axis-angles), and $\bar{\mathbf{p}}^h \in \mathbb{R}^3$ (wrist translation). The visual feature $\mathbf{f}_t^h$, camera pose $\mathbf{g}_t$, and scene-geometry token $\mathbf{s}_t$ are never perturbed.

Joint quality update. Continuous perturbations attenuate the affected quality channel multiplicatively,

$$q_\star^{h\prime} \;=\; q_\star^h \cdot \exp(-\Delta e_\star / \sigma_\star)\,, \qquad \star \in \{W, F\}, \tag{12}$$

where σ_W, σ_F are the per-corpus bandwidths of Eq. 3 and $\Delta e_\star \geq 0$ is the perturbation-induced increase in the corresponding component error of Eqs. 1–2. Eq. 12 follows from the RBF form $q{=}\exp(-e/\sigma)$: substituting the post-perturbation error $e_\star{+}\Delta e_\star$ factors $q_\star^{h\prime}$ into $q_\star^h{\cdot}\exp(-\Delta e_\star/\sigma_\star)$. Discrete perturbations instead clamp the affected quality channels to zero,

$$q_\star^{h\prime} \;=\; 0, \tag{13}$$

mirroring the inference-time treatment of frames whose hand-confidence flag $\delta_t^h{=}0$ or whose pose collapses to a degenerate configuration.

Gaussian noise. We sample a noise scale $\sigma_g \sim \text{Unif}\{10^{-2}, 2{\times}10^{-2}, 5{\times}10^{-2}\}$ and add zero-mean isotropic Gaussian noise of standard deviation σ_g to all 54 channels of $\bar{\mathbf{y}}_t^h$. Both wrist and finger

errors increase, attenuating q_W^h and q_F^h via Eq. 12. This type imitates the residual prediction noise that MANO regression heads incur on textured backgrounds.

Depth scaling. We sample a scale factor $s \sim \mathrm{Unif}(0.5, 2.0)$ and rescale the wrist translation as $\bar{\mathbf{p}}^{h\prime} = s\,\bar{\mathbf{p}}^h$, leaving wrist rotation and finger articulation unchanged. Only the wrist error increases, attenuating q_W^h via Eq. 12. This type imitates the metric-depth ambiguity intrinsic to monocular geometry estimators on egocentric video. On our deployable pipeline this manifests as DA3 [18] drift on novel scenes that shifts the wrist by hundreds of millimeters.

Wrist-rotation jitter. We add zero-mean Gaussian noise of standard deviation $\sigma_R = 0.05$ rad to the rot6d representation of the wrist rotation $\bar{\boldsymbol{\theta}}^h_{\mathrm{wrist}}$, leaving wrist translation and finger articulation unchanged. Since the wrist error of Eq. 1 is defined on translation alone, we attenuate q_W^h via the angular variant $q_W^{h\prime} = q_W^h \cdot \exp(-\theta_R/\sigma_R)$, where θ_R is the geodesic distance between the perturbed and unperturbed wrist rotations. This type imitates tracker confusion under fast head rotation.

Finger lock-in. We sample a contiguous window of $L \sim \mathrm{Unif}(5, 30)$ frames and clamp the 45 finger axis-angle channels of $\bar{\boldsymbol{\theta}}^h_{\mathrm{fingers}}$ to zero across the window. The finger quality channel is clamped to $q_F^{h\prime} = 0$ via Eq. 13. This type imitates short-window articulation failures during occluded grasps.

Per-hand dropout. We sample a contiguous window of $L \sim \mathrm{Unif}(5, 30)$ frames, zero all 54 channels of $\bar{\mathbf{y}}_t^h$, and reset the hand-confidence flag to $\delta_t^h = 0$ across the window. Both quality channels are clamped to $q_W^{h\prime} = q_F^{h\prime} = 0$ via Eq. 13. This type imitates extended out-of-frame events when one hand leaves the egocentric camera frustum during bimanual manipulation.

Effect on the training distribution. Without augmentation, the calibration of Eq. 3 places only $\sim 20\%$ of training frames in the low-quality region $q < 0.1$ on each q component, by construction. Under the augmentation operator $\mathcal{A}$, this fraction rises to over 50% on each of q_W^h and q_F^h. The per-channel forward process of Sec. 3.3 therefore receives substantially broader exposure to the low-quality regime under augmentation. The increase requires no additional annotated frames: every perturbed sample inherits its ground-truth motion target from the original corpus.

F Training Dataset Details

The quality network is pretrained on eight egocentric bimanual hand-pose corpora totalling 10M frames, split into two tiers based on annotation type. **Tier 1** (native MANO ground truth) comprises five datasets, on which the quality label is computed directly from MANO joint-regression error. **Tier 2** (3D joint annotations only) comprises three datasets, on which the quality label is computed from 3D joint MPJPE without any MANO fitting. All eight corpora are restricted to egocentric viewpoints: for ARCTIC and Ego-Exo4D we use the egocentric stream only, and for Re:InterHand we use the released egocentric camera split. Tab. 10 summarizes scale, viewpoint, and annotation type for each corpus.

Table 10: **Training corpora for quality-network pretraining.** All eight datasets are restricted to egocentric viewpoints. Tier 1 datasets provide native MANO parameters. Tier 2 datasets provide 3D joint positions only and supervise the quality label through joint MPJPE without MANO fitting.

Tier	Dataset	Frames	Subjects	View	GT type
1	HOT3D [2]	1.5M	19	ego	MANO
	ARCTIC [6]	0.24M	10	ego (stream)	MANO
	Re:InterHand [28]	0.15M	26	ego (synth.)	MANO
	H2O [15]	2.0M	4	ego	MANO
	HOI4D [23]	0.7M	9	ego	MANO
2	EgoDex [12]	2.3M	–	ego	25 joints
	EgoVerse [36]	1.5M	2087	ego	21 joints
	Ego-Exo4D [9]	1.6M	740	ego (stream)	21 joints
	Total	**10M**			

Tier 1 datasets.

HOT3D [2] is the primary benchmark: egocentric clips captured by Project Aria and Quest 3 with MANO annotations obtained from optical-marker motion capture. The dataset exhibits frequent hand out-of-view events and hand-object occlusion under dynamic egocentric camera motion.

ARCTIC [6] captures bimanual manipulation of articulated objects (scissors, laptops) with one egocentric and eight allocentric cameras. We use the egocentric stream only. Both hands are frequently mutually occluded during dexterous manipulation, producing large per-hand quality asymmetries. We adopt the official ARCTIC P2 protocol, which partitions the ten recording subjects (s01–s10) into eight training subjects (s01, s02, s04, s06, s07, s08, s09, s10), one validation subject (s05), and one test subject (s03) whose ground truth is withheld for the private leaderboard. Because the s03 ground truth is not publicly available, we report all ARCTIC numbers on the validation subject s05 (34 sequences sliced into 155 clips of $T{=}150$ frames at 30 fps). The same s05 is held out from the multi-corpus quality-network pretraining.

Re:InterHand [28] provides photorealistically re-rendered InterHand2.6M [27] two-hand interaction motions under randomized illumination and camera placement. We take only the released egocentric camera split ($\sim$147K frames). Native MANO parameters are inherited from the source motion capture, and the corpus supplies dense bimanual hand-hand interaction supervision under ego viewpoints, a regime under-represented in the rest of the mixture.

H2O [15] captures egocentric two-hand object manipulation across 4 subjects, 8 actions, and 36 objects with native MANO annotations and per-frame object 6DoF poses.

HOI4D [23] captures 4000 egocentric sequences of category-level human-object interaction across 9 indoor scenes and 16 object categories with a head-mounted Kinect Azure, providing native MANO parameters and 3D joint positions under natural ego head motion. HOI4D contains a strong right-hand bias (only $\sim 10\%$ of valid hand-frames are left-hand), which primarily contributes right-hand egocentric supervision to the quality network.

Tier 2 datasets.

EgoDex [12] provides 829 hours of egocentric video from Apple Vision Pro with high-precision 25-joint hand tracking across 194 tabletop tasks.

EgoVerse [36] aggregates 1,362 hours of egocentric demonstrations from 2,087 participants across 240 scenes, offering the largest demographic and environmental diversity.

Ego-Exo4D [9] provides 21M automatic 3D hand-joint annotations from synchronized ego and exo cameras across 13 cities. We use the egocentric stream only.

Tier 2 datasets are used for quality-network pretraining only: the quality label $q{=}\exp(-e/\sigma)$ is computed from the 3D joint MPJPE between the visual-encoder prediction and the dataset ground truth, requiring no MANO parameter fitting.

G Evaluation Metrics

This section provides formal definitions of the evaluation metrics reported in the main paper. Let $\hat{\mathbf{J}} \in \mathbb{R}^{T\times K\times 3}$ and $\mathbf{J} \in \mathbb{R}^{T\times K\times 3}$ denote the predicted and ground-truth joint positions of a single hand over T frames with K joints. All metrics are computed per hand on frames with valid ground-truth annotations, then averaged across hands and clips.

Procrustes-Aligned MPJPE (PA-MPJPE). PA-MPJPE isolates articulation quality from global trajectory drift by applying a per-frame Procrustes alignment (rotation, translation, and uniform scale) before measuring joint error. At each frame t, the optimal similarity transform $s_t, \mathbf{R}_t, \mathbf{t}_t = \mathrm{Procrustes}(\hat{\mathbf{J}}_t, \mathbf{J}_t)$ is computed, and the metric is

$$\text{PA-MPJPE} \;=\; \frac{1}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\|s_t\mathbf{R}_t\hat{\mathbf{J}}_{t,k} + \mathbf{t}_t - \mathbf{J}_{t,k}\|_2. \tag{14}$$

World MPJPE (W-MPJPE). Following [51, 49], W-MPJPE is the world space MPJPE after a similarity alignment $(s^*, \mathbf{R}^*, \mathbf{t}^*)$ on the first frame, capturing trajectory drift accumulated from the

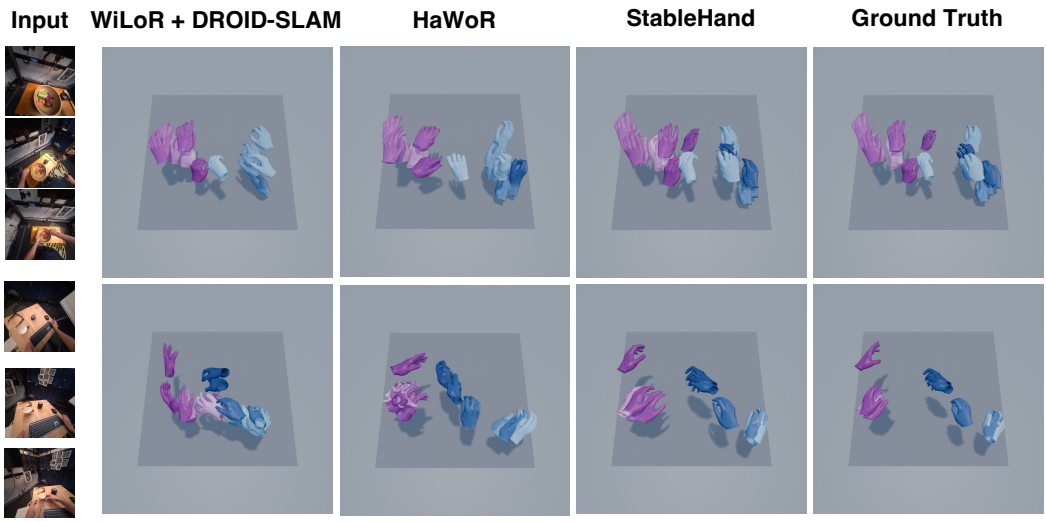


Figure 9: **Additional qualitative results on HOT3D [2].** Two further HOT3D clips with long missing-hand spans, comparing WiLoR [34]+DROID-SLAM [43], HaWoR [51], our StableHand, and the Ground Truth. Layout, color, and temporal-shading conventions follow Fig. 3.

shared anchor:

$$\text{W-MPJPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\|s^*\mathbf{R}^*\hat{\mathbf{J}}_{t,k}+\mathbf{t}^*-\mathbf{J}_{t,k}\|_2. \tag{15}$$

World-Aligned MPJPE (WA-MPJPE). WA-MPJPE uses the same expression but with $(s^*, \mathbf{R}^*, \mathbf{t}^*)$ a single global similarity transform that minimizes the total MPJPE over all T frames jointly, capturing the residual per-joint pose error after the global trajectory alignment.

Acceleration Error (AccEr). AccEr measures the mean per-joint acceleration discrepancy between the predicted and ground-truth trajectories at frame rate $f = 30$ fps:

$$\mathbf{a}_t = \mathbf{J}_{t+1} - 2\mathbf{J}_t + \mathbf{J}_{t-1}, \qquad \text{AccEr} = \frac{1}{(T-2)K}\sum_{t=2}^{T-1}\sum_{k=1}^{K}\frac{\|\hat{\mathbf{a}}_{t,k}-\mathbf{a}_{t,k}\|_2}{(1/f)^2}. \tag{16}$$

Mean Relative Root Position Error (MRRPE). MRRPE measures the spatial consistency between the two hands by comparing their predicted and ground-truth wrist offset, following [6, 27]. Let $\hat{\mathbf{j}}^L_{t,0}, \hat{\mathbf{j}}^R_{t,0} \in \mathbb{R}^3$ and $\mathbf{j}^L_{t,0}, \mathbf{j}^R_{t,0} \in \mathbb{R}^3$ denote the predicted and ground-truth left and right wrist positions at frame t. With $\mathcal{B} \subseteq \{1, \ldots, T\}$ the set of frames at which both hands have valid ground-truth annotations,

$$\text{MRRPE} = \frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}}\left\|(\hat{\mathbf{j}}^L_{t,0}-\hat{\mathbf{j}}^R_{t,0})-(\mathbf{j}^L_{t,0}-\mathbf{j}^R_{t,0})\right\|_2. \tag{17}$$

MRRPE is reported on ARCTIC only, where dexterous bimanual manipulation makes the relative spatial relationship between the two hands a primary axis of evaluation.

H Additional Qualitative Results

This section presents qualitative comparisons that complement Fig. 3 of the main paper. Fig. 9 and Fig. 10 report additional clips on HOT3D and ARCTIC respectively, following the same layout as the main qualitative figure. Fig. 11 reports two representative failure modes of our method on HOT3D, in which one hand is unobserved by the upstream visual stream across most or all of the clip and the corresponding channel is synthesized entirely from the prior. Fig. 12 reports two analogous failure modes on ARCTIC, in which persistent bimanual hand-object occlusion simultaneously degrades both hands' observations during long contact phases.

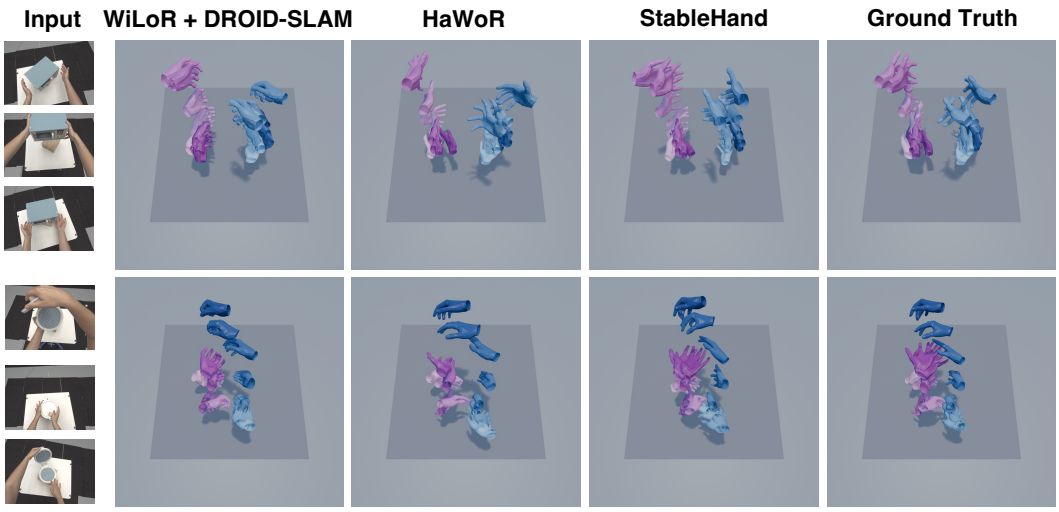


Figure 10: **Additional qualitative results on ARCTIC [6].** Two further ARCTIC clips with persistent hand-object occlusion, comparing WiLoR [34]+DROID-SLAM [43], HaWoR [51], our StableHand, and the Ground Truth. Layout, color, and temporal-shading conventions follow Fig. 3.

I Zero-Shot In-the-Wild Inference

Beyond the in-distribution evaluations on HOT3D and ARCTIC, we run StableHand zero-shot on HD-EPIC [33], a recent in-the-wild egocentric corpus excluded from both the generative-model training splits and the eight-corpus quality-network pretraining mixture (Sec. F). Fig. 13 reports a representative clip in which the egocentric camera traverses dim corridors and varied indoor scenes never seen at training time. The recovered world-space dual-hand mesh trajectory remains spatially coherent, suggesting that the per-component quality conditioning transfers to upstream observations whose error distribution differs from any single training corpus.

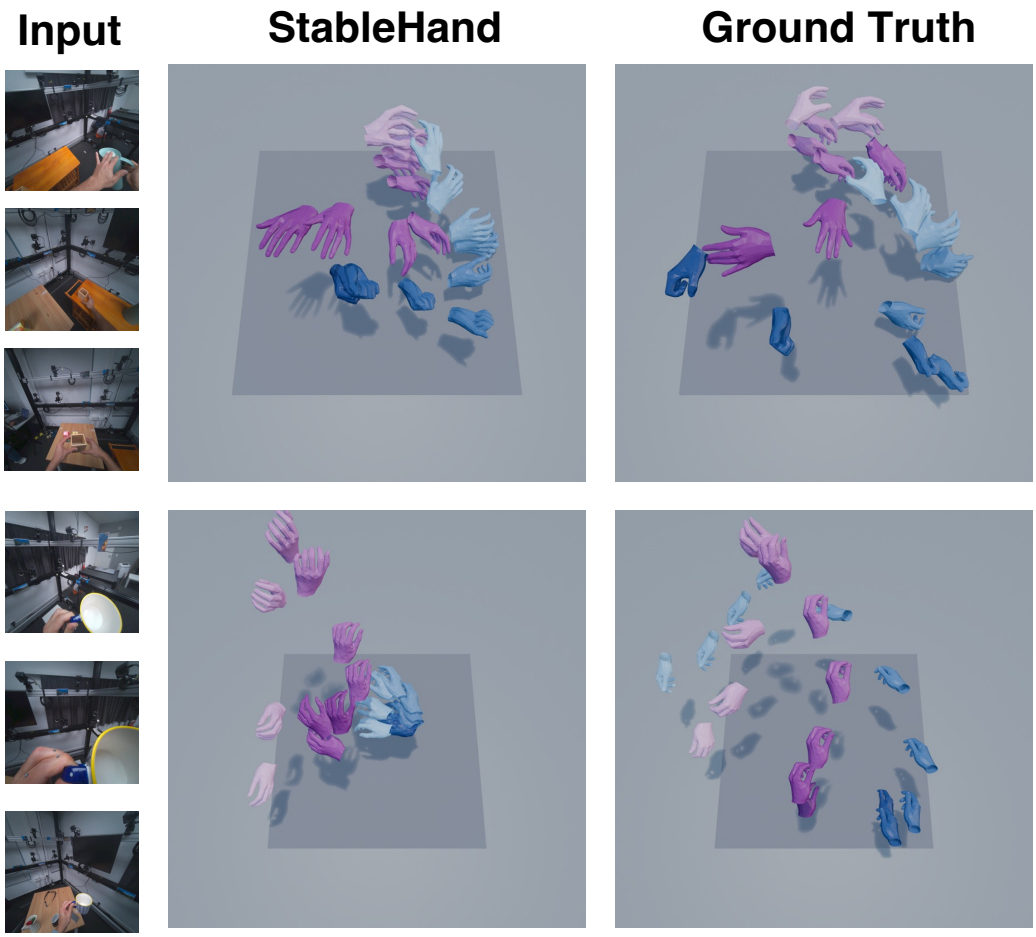


Figure 11: **Failure cases of StableHand on HOT3D [2].** Each row shows three input frames (left) together with our prediction and the Ground Truth, with color and temporal-shading conventions following Fig. 3. Top: the left hand leaves the egocentric view for an extended span, leaving its channel unobserved for tens of consecutive frames. Bottom: the right hand never enters the camera frustum across the entire clip, leaving its channel unobserved at every frame. In both cases the generative process synthesizes a plausible but incorrect trajectory from the prior alone, illustrating the limit of our method when the upstream visual stream provides no observation to anchor a channel.

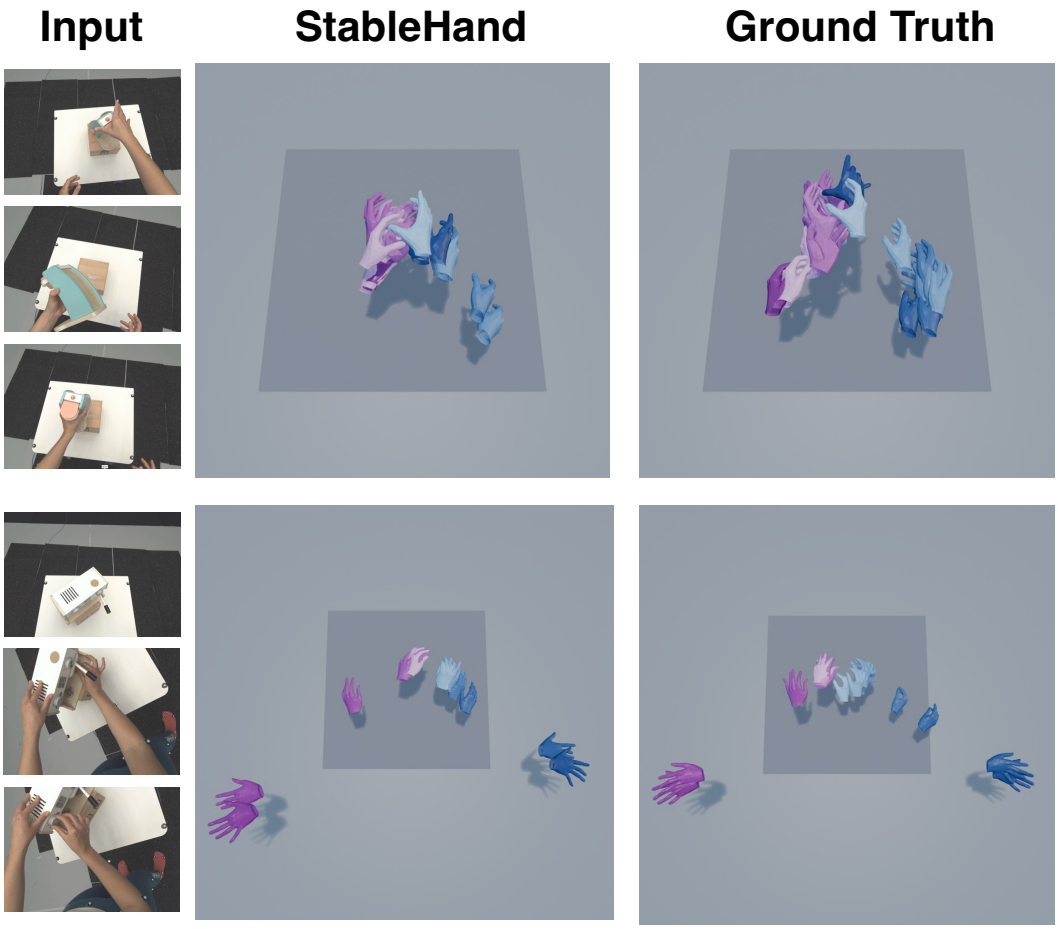


Figure 12: **Failure cases of StableHand on ARCTIC [6].** Each row shows three input frames (left) together with our prediction and the Ground Truth, with color and temporal-shading conventions following Fig. 3. Top: a bimanual contact phase in which the manipulated object simultaneously occludes both hands' fingers, leaving both finger channels with degraded observations across the contact span. Bottom: a bimanual manipulation phase in which the object covers both hands intermittently, leaving the two-hand spatial relationship without a reliable observation to anchor against. In both cases simultaneous degradation of both per-hand quality channels forces the generative process to recover from the prior, illustrating the structural limit of per-component quality conditioning when no hand provides a reliable observation across the contact phase.

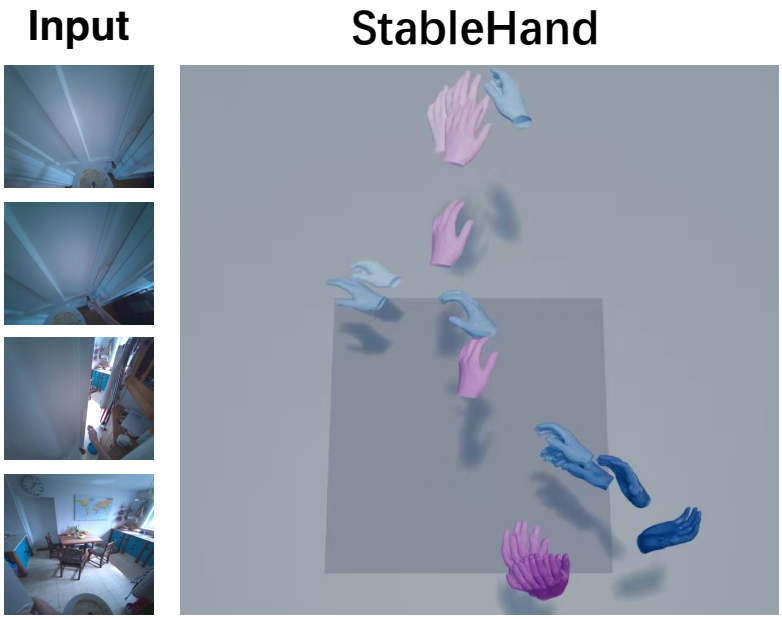


Figure 13: **Zero-shot in-the-wild inference on HD-EPIC [33].** Four input frames sampled from a single HD-EPIC clip (left) span dim corridors, kitchens, and dining rooms outside our training distribution. The right panel shows the recovered world-space dual-hand mesh trajectory (left hand magenta, right hand blue, mesh shading dark→light encoding temporal order). Neither the generative model (trained on HOT3D and ARCTIC) nor the quality network (pretrained on the eight-corpus mixture of Sec. F) was exposed to HD-EPIC during training, yet the recovered trajectory remains spatially coherent.