GENERATIVE HAND MODELING: A UNIFIED MODEL FOR EGOCENTRIC HAND MOTION RECONSTRUCTION, INFILLING AND GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: **Generative Hand Modeling (GHM).** A common hand-motion modeling framework supports reconstruction from visual observations, motion infilling, and scene- and language-conditioned generation across diverse hand–object interactions.

ABSTRACT

Modeling 4D hand motion is important for understanding human manipulation and learning from human demonstrations. Existing approaches, however, typically treat reconstruction, motion infilling, and generation as separate problems, despite their shared need to model hand articulation and temporal dynamics. We present Generative Hand Modeling (GHM), a unified framework that formulates these tasks as conditional hand motion modeling under varying levels of visual observability. GHM uses a shared motion representation and backbone to learn from different forms of visual and semantic supervision. It jointly models both hands within a common motion space. Experiments across these settings demonstrate strong reconstruction and temporal infilling performance, while retaining the ability to synthesize hand motion from scene and language conditions. Our results show the potential of unified generative modeling for learning hand motion from diverse sources of egocentric interaction data.

1 INTRODUCTION

Learning from human manipulation is a promising direction for embodied intelligence (Zheng et al., 2026; Hoque et al., 2026). Egocentric videos paired with 3D hand motion offer a scalable source of such demonstrations. Recent advances have enabled hand motion reconstruction from video (Zhang et al., 2025; Xu et al., 2026), alongside interaction synthesis from scene and semantic cues (Prakash et al., 2025). *Reconstruction* seeks motion consistent with visual observations, while *generation* synthesizes plausible motion from scene and semantic cues. Temporal *infilling* switches between the two: reconstructing on observed frames and generating within missing intervals. Although their conditions and objectives differ, these tasks share the need to model hand articulation and temporal dynamics.

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

Recent work has begun to unify motion estimation and generation through conditional synthesis (Li et al., 2025a; Sun et al., 2026). Beyond supporting multiple tasks, unification offers a way to jointly use reconstruction supervision from egocentric videos and interaction-specific supervision for generation. Rather than constructing separate models for different amounts of available visual evidence, we ask *whether reconstruction, infilling, and generation can be learned within a shared generative model*, such that **supervision from different observation regimes** contributes to a **common model of hand motion**.

To this end, we introduce Generative Hand Modeling (GHM), a unified generative model for 4D hand motion under varying levels of visual observability. GHM unifies the three tasks by varying the conditioning information while sharing the motion backbone and output representation. For reconstruction and infilling, the model is conditioned on visual observations throughout the sequence or outside missing intervals, respectively. For generation, it uses an initial scene condition, including the initial MANO hand states, and a language instruction. The shared representation jointly describes both hands, allowing their temporal and inter-hand dependencies to be modeled without hand-flipping preprocessing. To connect frame-wise visual evidence with motion in a common reference frame, we represent motion in both camera and world coordinates and additionally link the predictions through a consistency objective.

Concretely, GHM learns a conditional distribution over complete dual-hand motion sequences using a shared diffusion model. Explicit condition masks allow the model to accommodate missing visual or semantic annotations across training datasets. During training, we additionally mask available conditions while retaining supervision on all valid motion targets, exposing the model to varying levels of motion observability. This training strategy allows the shared generative model to use available observations for reconstruction and infer plausible motion where visual evidence is missing.

We evaluate GHM across hand motion reconstruction, temporal infilling, and text-conditioned generation. Under our evaluation protocol on HOT3D, GHM reduces reconstruction MPJPE by 25.8% relative to the strongest evaluated baseline on this metric. Ablations further show that combining diffusion and regression training improves reconstruction pose accuracy over regression-only training, supporting the compatibility of generative motion modeling and accurate reconstruction.

Our contributions are:

- We formulate reconstruction, infilling, and generation as conditional 4D hand motion modeling under varying visual observability. Explicit condition masking enables heterogeneous visual and semantic supervision to contribute to a generalist motion model.
- We introduce GHM, which jointly models both hands and combines diffusion and single-pass regression training with a camera–world consistency objective. Shared model parameters support direct prediction for reconstruction and infilling, as well as iterative sampling for text-conditioned generation.

2 Related Work

Hand Motion Reconstruction Image-based hand reconstruction commonly builds on the MANO representation (Romero et al., 2017), with transformer-based methods such as HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025) estimating hand pose and shape from individual images. Beyond individual frames, temporal modeling and camera-motion estimation enable world-space hand motion reconstruction (Yu et al., 2025; Zhang et al., 2025; Zhu et al., 2026).

Occlusions and out-of-view intervals further motivate learned motion priors to complement incomplete visual evidence. Such priors have been incorporated through latent-space fitting (Duran et al., 2024), dedicated motion infilling (Zhang et al., 2025), and flow-based recovery (Zeng et al., 2026; Xu et al., 2026). A complementary line repurposes pretrained video diffusion representations for hand motion reconstruction (Wang et al., 2026; Liu et al., 2026). More broadly, Li et al. (2025a) and Sun et al. (2026) explore unified estimation and generation for human-body and hand motion, respectively. Our model extends this perspective to language-directed interactions, connecting visual reconstruction and infilling with motion generation conditioned on an initial scene and a task instruction.

Hand Motion Generation Diffusion models have advanced conditional human motion synthesis, from text-conditioned generation (Tevet et al., 2023) to autoregressive synthesis with spatial control (Zhao et al., 2025). GigaHands (Fu et al., 2025) and HandX (Zhang et al., 2026) provide annotated bimanual motion data and benchmarks for text-conditioned generation. CLUTCH (Thambiraja et al., 2026) explores this task through language modeling over discrete hand-motion tokens. Complementary work also studies body-conditioned hand motion completion (Shi et al., 2026). For hand–object interaction, Text2HOI (Cha et al., 2024) and DiffH2O (Christen et al., 2024) exploit explicit object geometry together with contact or grasp modeling, while LatentHOI (Li et al., 2025b) explores grasp-aware latent diffusion for interaction generation.

LatentAct (Prakash et al., 2025) anchors language-guided hand interactions in an observed scene through a specified 3D contact point. MEgoHand (Zhou et al., 2026) explores egocentric motion generation conditioned on visual context, language, and an initial hand pose. Our model follows this scene-conditioned formulation, with optional 2D wrist-trajectory guidance and without requiring object meshes or predefined 3D contact targets. Rather than training a standalone generator, we integrate this setting with visual reconstruction and infilling in a shared hand-motion model.

3 METHOD

Problem formulation. GHM formulates three hand-motion tasks as conditional generation problems under different conditions. All three tasks share an initial scene condition $\mathbf{S}_0$. For hand motion reconstruction, GHM models $p_\theta(\mathbf{X} \mid \mathcal{C}_{\text{rec}})$, where $\mathcal{C}_{\text{rec}} = \{\mathbf{S}_0, \mathbf{O}_{1:T}, \mathbf{E}_{1:T}\}$ contains the scene, per-frame hand observations, and camera motion, respectively. For hand motion generation, GHM models $p_\theta(\mathbf{X} \mid \mathcal{C}_{\text{gen}})$, where $\mathcal{C}_{\text{gen}} = \{\mathbf{S}_0, \mathbf{d}\}$ contains the scene and motion description. In this setting, $\mathbf{S}_0$ also includes the initial MANO states of both hands; an optional 2D wrist trajectory $\mathbf{W}_{1:T}$ may provide additional control. Hand motion infilling is formulated similarly to reconstruction but uses incomplete hand observations: $p_\theta(\mathbf{X} \mid \mathcal{C}_{\text{inf}})$, where $\mathcal{C}_{\text{inf}} = \{\mathbf{S}_0, \mathbf{O}_\Omega, \mathbf{E}_{1:T}\}$ and $\Omega \subset \{1, \ldots, T\}$ indexes the visible frames. GHM completes the missing motion from the visible observations. Despite their different inputs, all three tasks are handled by a single model with shared parameters.

Figure 2 gives an overview of the conditioning pathways, motion prediction, and cross-representation consistency.

3.1 DUAL-HAND WORLD–CAM MOTION REPRESENTATION

Usually, reconstruction methods like HaWoR recover hand motions in detected frames' camera space, while generation methods like DiffH2O model hand–object interaction in an object-centered canonical space. To support both reconstruction and generation, GHM models hand motion in both camera and world space, with the world frame defined by the first camera frame c_0 of each episode.

Each hand $h \in \{\text{L}, \text{R}\}$ is represented at frame i by the 115-dimensional vector

$$\mathbf{x}_{h,i} = \big[\underbrace{\boldsymbol{\theta}_{h,i}}_{90}, \underbrace{\boldsymbol{\beta}_h}_{10}, \underbrace{R_{c_i,h}}_{6}, \underbrace{\Delta R_{c_0,h,i}}_{6}, \underbrace{\mathbf{p}_{c_0,h,i}}_{3}\big], \tag{1}$$

where $\boldsymbol{\theta}_{h,i} \in \mathbb{R}^{90}$ contains the 15 MANO hand-joint rotations, $\boldsymbol{\beta}_h \in \mathbb{R}^{10}$ contains the shape parameters, $R_{c_i,h}$ is the absolute root orientation in the current camera frame c_i, and $\mathbf{p}_{c_0,h,i}$ is the offset-aware absolute translation in world space. Both root-orientation fields are encoded as 6D rotations.

The two orientation channels serve complementary roles. The direct channel $R_{c_i,h}$ aligns each prediction with the frame-local image evidence. The incremental channel is $\Delta R_{c_0,h,i} = R_{c_0,h,i-1}^\top R_{c_0,h,i}$, with $\Delta R_{c_0,h,0} = I$, where $R_{c_0,h,i}$ denotes the root orientation expressed in world space. Absolute $R_{c_0,h,i}$ is not a separate output field. It is recovered by composing the adjacent rotations from an orientation anchor provided by a visible frame of the same hand, or by the initial hand state for generation. The consistency objective below aligns the predicted increments with direct camera-space predictions.

Explicit dual-hand modeling. As real-world hand manipulation often requires bimanual coordination, GHM directly takes both left- and right-hand observations as input and jointly predicts their motions. This formulation preserves hand identity without flipping-based canonicalization.

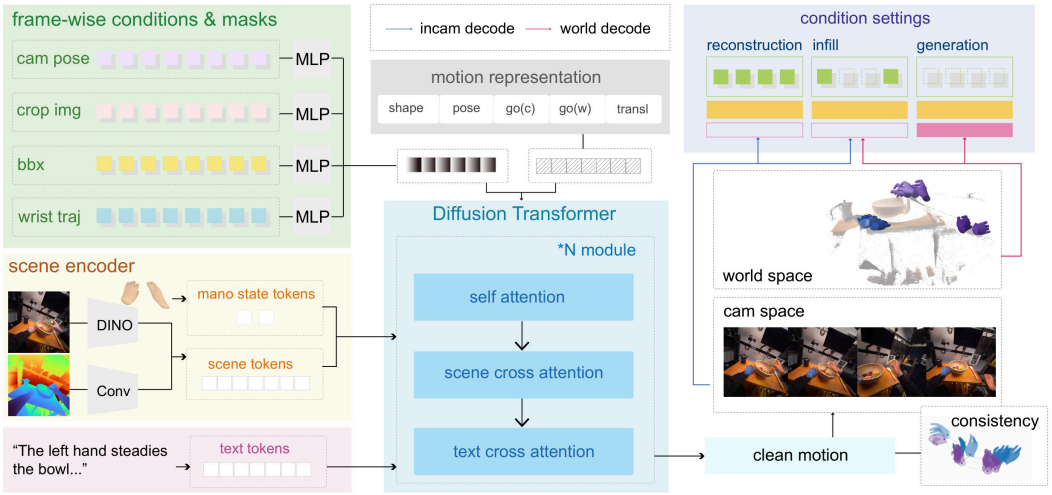


Figure 2: **Overview of GHM.** Frame-wise conditions and availability masks are combined with noisy dual-hand motion. The Transformer predicts clean motion using scene and text context through cross-attention. Cross-representation consistency connects the camera- and world-space predictions.

3.2 Unified Model Architecture

GHM builds on GENMO (Li et al., 2025a), adapting its clean-motion-predicting diffusion framework to dual-hand motion using the representation introduced in Section 3.1. To condition hand motion on the surrounding environment, GHM extends the backbone with a dedicated cross-attention pathway for scene features. Given a clean motion $\mathbf{X}$ and the task-specific conditions $\mathcal{C}_\tau$, the forward process forms a noisy sample $\mathbf{X}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{X} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$. The corresponding DDPM objective is

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{X},t,\boldsymbol{\epsilon}}\left[\|D_\theta(\mathbf{X}_t, t, \mathcal{C}_\tau) - \mathbf{X}\|_2^2\right], \tag{2}$$

Here, t is the diffusion timestep, $\bar{\alpha}_t$ is the cumulative noise schedule, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$ is Gaussian noise, and D_θ is the conditional denoiser.

GHM adopts a RoPE-based Transformer denoiser with bidirectional self-attention for temporal motion modeling. Frame-aligned conditions are fused with the noisy motion and timestep embeddings to form per-frame tokens, while text and scene context are incorporated through separate cross-attention pathways. The denoiser predicts clean dual-hand motion $\widehat{\mathbf{X}}$, with an auxiliary camera head providing weak-perspective crop-camera parameters for geometric supervision. Task identity is specified solely by condition visibility, without a task token, task embedding, or task-specific prediction head.

Condition inputs. Per-frame conditions are encoded using modality-specific embeddings and fused with the noisy motion, while scene and text conditions are encoded as sequence-level context tokens and incorporated through cross-attention. Hand-specific features are aggregated separately for the left and right hands and concatenated in a fixed order before fusion with the noisy motion, preserving handedness in the conditioning representation. For generation, tokens encoding the initial MANO states of both hands are appended to the visual scene tokens as part of the scene condition.

GHM uses condition masks to accommodate task-dependent inputs and incomplete annotations within a unified conditioning interface. Each condition is paired with a binary availability mask, set to zero when the corresponding annotation is absent from the source dataset. During training, available conditions are additionally masked out at random. This masking affects only the conditioning inputs; all valid ground-truth targets remain supervised.

The conditioning interface also accommodates additional geometric inputs, such as 2D hand keypoints; we evaluate this extension in Appendix D.1.

3.3 TRAINING OBJECTIVES

Diffusion and single-pass regression. Following GENMO's dual-mode training framework, GHM trains a shared denoiser in both diffusion and regression modes on every batch. The diffusion mode optimizes Eq. equation 2 with a cosine schedule and uniformly sampled timesteps. In regression mode, the denoiser predicts $\mathbf{X}$ from an initial noise sample in a single forward pass using the same supervision objectives. GHM additionally supervises the absolute world-space root orientations recovered by composing the predicted rotation increments against their ground-truth counterparts.

Cross-representation consistency. In addition to the denoising objective and geometric supervision, GHM uses a cross-representation consistency loss to ensure that its world- and camera-space predictions describe the same hand motion. The direct camera-space predictions serve as stop-gradient targets. For orientation, the direct predictions are transformed to a common world frame and converted into adjacent relative rotations, which supervise the predicted world-space increments through a chordal loss. For translation, the absolute world-space prediction is decoded into a physical wrist position and transformed to c_i using the calibrated camera extrinsics, where it is aligned with the detached camera-space wrist prediction. The two terms are combined as $\mathcal{L}_{\text{cons}} = \lambda_R \mathcal{L}_R + \lambda_p \mathcal{L}_p$, where

$$\mathcal{L}_R = \underset{i,h}{\text{mean}}\, d_{\text{chord}}\left(\Delta \widehat{R}_{c_0,h,i}, \text{sg}[\widetilde{\Delta R}_{h,i}]\right), \quad \mathcal{L}_p = \underset{i,h}{\text{mean}}\, \left\| \widehat{\mathbf{q}}^{(0)}_{c_i,h} - \text{sg}[\widetilde{\mathbf{q}}_{c_i,h}] \right\|_1 . \tag{3}$$

Here, $\widetilde{\Delta R}_{h,i}$ is the relative rotation between adjacent direct orientation predictions after conversion to world coordinates. $\widehat{\mathbf{q}}^{(0)}_{c_i,h}$ and $\widetilde{\mathbf{q}}_{c_i,h}$ denote wrist positions derived from the world-space and direct camera-space predictions, respectively, both expressed in c_i. The function d_{chord} is the rotation chordal distance, sg denotes stop-gradient, and λ_R and λ_p weight the two losses. The orientation term requires a valid adjacent frame.

4 EXPERIMENTS

We evaluate GHM as a unified model for hand-motion modeling under three conditioning settings: reconstruction from complete visual observations, generation from scene and language conditions, and infilling from incomplete observations.

Datasets. We train our model on a diverse collection of egocentric hand–object interaction datasets, including HOT3D (Banerjee et al., 2025), OakInk2 (Zhan et al., 2024), TACO (Liu et al., 2024), HOI4D (Liu et al., 2022), ARCTIC (Fan et al., 2023), EgoDex (Hoque et al., 2026), Xperience (Ropedia, 2026), and HoloAssist (Wang et al., 2023). In total, the training corpus contains 155,775 motion sequences and approximately 43.4 million frames. We evaluate reconstruction on HOT3D (in-domain) and H2O (Kwon et al., 2021) (unseen, OOD). For HOT3D, we follow the official `HOT3D-Clips` partition and evaluate on the same 70 test clips as HandFlow. We evaluate generation on HoloAssist, with an additional baseline comparison on TACO, and infilling on HOT3D and ARCTIC. Dataset statistics, splits, and evaluation settings are provided in Appendix B.

Metrics. **i) Reconstruction and infilling.** For camera-space reconstruction and infilling, we report **MPJPE**, wrist root-aligned **RA-**MPJPE , and Procrustes-aligned **PA-**MPJPE. World-space metrics use similarity alignment fitted to the first two valid frames (**W-**) or all valid frames (**WA-**). We also report camera-space wrist trajectory error (**RTE**) and joint acceleration error (**Accel.**). **ii) Generation.** We report MPJPE and **Global PA-MPJPE** to facilitate comparison with LatentAct; the latter uses a single similarity transform for the full joint trajectory. Since multiple motions can satisfy the same conditions, agreement with a single reference does not fully characterize generation quality. We therefore also assess wrist distance and approach-aligned orientation relative to a fixed object reference point (**Min-Dist.**, **GO**), wrist-trajectory similarity (**DTW**), and sample diversity. Position and distance metrics are in millimeters; GO is in degrees, RTE in percent, and Accel. in m/s^2. Appendix C.2 provides detailed definitions.

Table 1: **Camera-space hand-motion reconstruction.** Results are reported for both hands and separately for the left and right hands.

	Method	MPJPE			RA-MPJPE			PA-MPJPE			RTE			Accel.		
		Both	Left	Right	Both	Left	Right	Both	Left	Right	Both	Left	Right	Both	Left	Right
HOT3D	HaMeR	67.77	70.03	65.91	59.58	60.16	59.10	9.74	9.69	9.79	3.21	4.05	2.52	12.89	13.16	12.66
	HaWoR	30.87	29.82	31.73	14.82	15.37	14.36	6.02	6.39	5.71	2.09	2.61	1.65	8.33	7.94	8.65
	WiLoR	55.45	60.47	51.34	20.78	22.13	19.67	6.95	7.11	6.82	2.79	3.45	2.25	12.53	11.96	13.01
	HandFlow	24.51	28.49	21.25	12.75	15.22	**10.73**	**5.45**	6.05	**4.96**	1.62	2.27	**1.09**	**3.36**	**3.31**	**3.41**
	MINT†	54.83	48.46	60.04	20.49	19.50	21.30	8.02	8.12	7.94	2.60	3.01	2.26	9.62	8.89	10.21
	GHM	**18.19**	**20.14**	**16.59**	**12.70**	**13.65**	11.93	5.56	**5.85**	5.33	**1.28**	**1.49**	1.11	4.68	4.62	4.72
H2O	HaMeR	48.68	39.01	58.33	22.02	22.69	21.35	7.46	7.08	7.84	4.13	4.21	4.05	6.73	6.57	6.89
	HaWoR	36.84	39.97	**33.69**	14.84	15.46	14.21	7.27	7.49	7.05	5.06	5.81	4.30	8.28	8.23	8.32
	WiLoR	37.94	34.41	41.46	19.39	19.65	19.12	**6.47**	**6.09**	**6.85**	3.53	3.78	3.29	5.93	5.66	6.19
	HandFlow	89.48	82.81	96.14	20.82	22.88	18.77	7.54	7.76	7.33	4.46	4.37	4.56	**4.02**	**3.89**	**4.15**
	MINT†	250.56	261.71	239.41	16.70	17.98	15.42	7.38	7.44	7.31	8.43	8.83	8.03	9.80	9.50	10.10
	GHM	**35.79**	**34.09**	37.50	**14.07**	**14.46**	**13.67**	6.89	6.79	6.98	**2.58**	**2.75**	**2.40**	4.10	4.03	4.16

†MINT retains its native detector-free inference and is evaluated on the same valid frames as all other methods.

4.1 Hand-Motion Reconstruction

We first evaluate the hand-pose reconstruction performance of the unified generative model given complete visual observations. We compare with HaMeR (Pavlakos et al., 2024), HaWoR (Zhang et al., 2025), WiLoR (Potamias et al., 2025), HandFlow (Xu et al., 2026), and MINT (Zhu et al., 2026), re-evaluating all baselines from their official checkpoints under our protocol. For camera-space evaluation, all methods, including GHM, share the same HaWoR-YOLO detections, except MINT†, which retains its native detector-free inference. All methods are scored on the same valid frames (i.e., frames with valid annotations and successful HaWoR-YOLO detections).

Tab. 1 shows strong MPJPE performance for GHM, alongside competitive RA-MPJPE and PA-MPJPE. Joint training with the generation objective may contribute to this accuracy, as our dual-mode ablation (Tab. 6) shows improved reconstruction over regression alone. GHM also maintains strong performance on H2O, which is outside the training domains of all compared methods, supporting transfer to an unseen dataset.

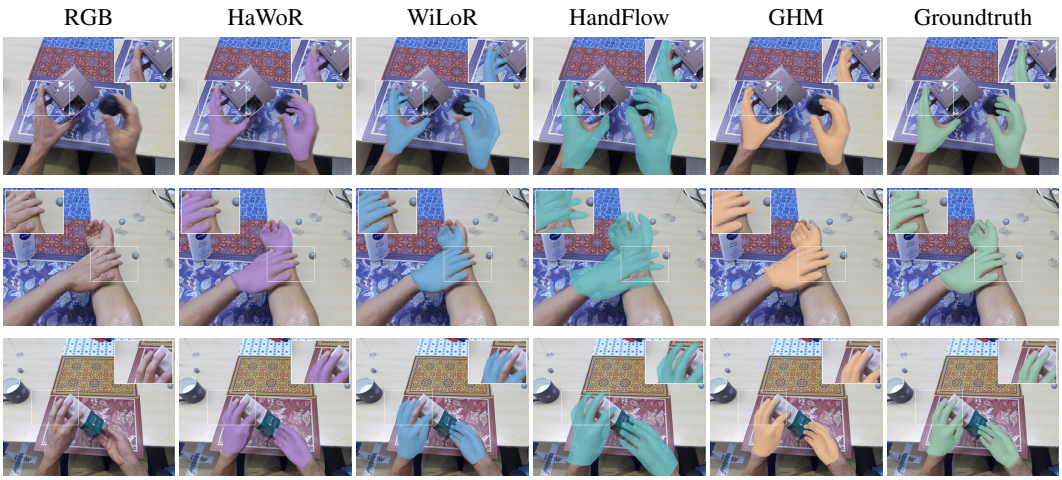


Figure 3: **Camera-space hand reconstruction on H2O.** Each row shows a different example, with columns showing the input RGB image, HaWoR, WiLoR, HandFlow, GHM, and Groundtruth. Insets highlight local hand-pose details.

World-space reconstruction. Our unified model supports both camera-space hand-pose estimation and world-space hand-trajectory rollout, with the world coordinate system aligned to the first

frame (C_0). We therefore also evaluate its directly decoded world-space trajectories, conditioning on ViPE-estimated (Huang et al., 2025) camera motion. For comparison, we also transfer the camera-space predictions from Sec. 4.1 to the frame-zero world coordinate system using ViPE-estimated camera extrinsics (Cam2World).

We compare with the published results of UniHand (Sun et al., 2026), which also represents hand motion in the first-frame camera coordinate system. As shown in Tab. 2, with estimated camera extrinsics, Cam2World consistently outperforms C_0 rollout on both W-MPJPE and WA-MPJPE. C_0 rollout still achieves better W-MPJPE than HaWoR's camera-space predictions transformed into C_0. We retain the C_0 rollout branch primarily to support conditional generation and motion infilling.

Table 2: **World-space hand-motion reconstruction on HOT3D.** Results aggregate both hands. HaWoR and GHM use shared valid frames.

Method	W-MPJPE	WA-MPJPE
UniHand (2026)	63.97	25.24
HaWoR (Est. Ext.) (2025)	43.49	<u>16.24</u>
GHM C_0 rollout (Est. Ext.)	<u>41.49</u>	19.23
GHM Cam2World (Est. Ext.)	**38.05**	**14.17**

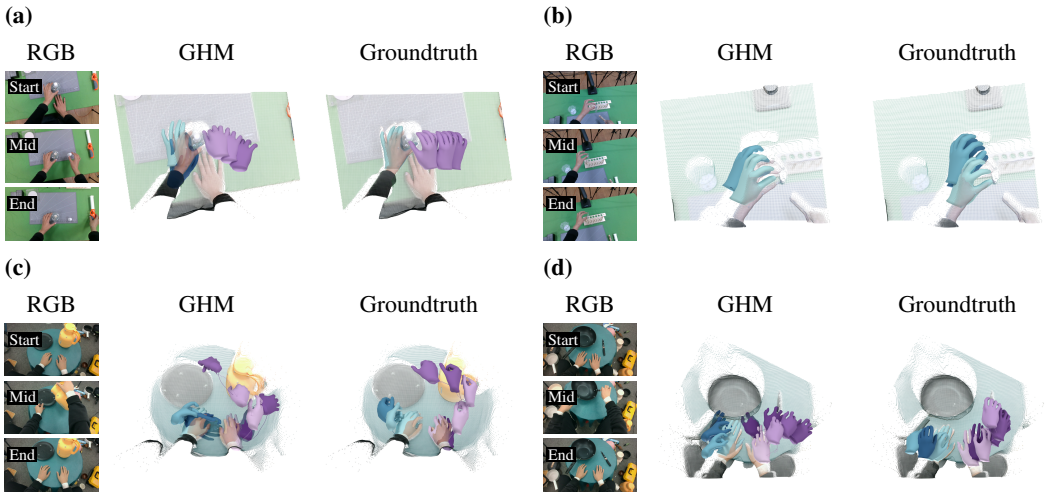


Figure 4: **Qualitative hand-motion generation on OakInk2 and TACO.** Panels (a)–(b) show OakInk2 examples, and (c)–(d) show TACO examples. Each example shows RGB frames at the start, middle, and end of the action, followed by wrist-conditioned GHM generation and Groundtruth in world coordinates. Blue and purple denote left and right hands; darker shades indicate later times. RGB frames provide visual context, not additional generation inputs. Hand poses are sampled independently for GHM and Groundtruth; the point cloud depicts only the initial scene.

4.2 Conditional Hand-Motion Generation

We compare GHM with LatentAct on long action sequences from HoloAssist. We evaluate GHM both with and without the complete GT 2D wrist trajectory as an additional input, while LatentAct predicts in 30-frame windows conditioned on each window's starting RGB image and released 3D contact centroid. We assemble these window predictions into full-action sequences for evaluation. Under these respective inputs, GHM achieves lower MPJPE and Global PA-MPJPE than the chunked LatentAct baseline (Tab. 3), along with lower minimum-distance and wrist-trajectory DTW errors. Adding wrist conditioning further improves these errors. Wrist conditioning also re-

Table 3: **Generation accuracy and diversity on HoloAssist.** We report average right-hand results.

Metric	LatentAct	GHM w/o wrist	GHM w/ wrist
MPJPE	103.13	<u>71.40</u>	**63.56**
PA-MPJPE*	42.69	<u>35.77</u>	**31.48**
Min-Dist.	31.56	<u>16.23</u>	**12.47**
GO	45.27	<u>25.60</u>	**23.78**
Abs. DTW	84.98	<u>52.31</u>	**45.79**
Wrist Div.	13.14	27.81	32.24
Pose Div.	12.56	10.33	9.11

*PA-MPJPE uses one full-sequence similarity alignment. Abs. denotes absolute. Div. denotes diversity; Pose Div. is wrist-relative.

duces GO error, suggesting that positional guidance can help resolve hand orientation during the approach, even without explicitly specifying it.

Diversity. Both GHM variants exhibit greater wrist-trajectory diversity than LatentAct, but lower root-relative pose diversity. Wrist conditioning improves prediction accuracy while retaining trajectory variation; in this evaluation, wrist diversity increases and root-relative diversity decreases relative to generation without wrist guidance. Protocol details are provided in Appendix C.1.

Wrist-following baseline. To assess whether the gains go beyond directly applying the wrist trajectory to a fixed hand pose, we compare GHM with a baseline that fixes the initial GT hand pose and global orientation and follows the input 2D wrist trajectory using a constant depth taken from the first GT frame. On egocentric TACO sequences, GHM achieves lower per-frame and sequence-level PA-MPJPE and Min-Dist. (Tab. 4). The improvement after per-frame alignment suggests more accurate hand articulation.

Table 4: **Comparison with a wrist-following baseline on TACO.**

Metric	Fixed pose + wrist	GHM
PA-MPJPE (mm)	14.92	**7.47**
PA-MPJPE* (mm)	53.87	**47.23**
Min-Dist.† (mm)	27.77	**20.75**

PA-MPJPE uses per-frame alignment; *uses one similarity transform for the full sequence. †Evaluated on the right hand.

4.3 Hand-Motion Infilling

We compare GHM with HaWoR and Dyn-HaMR on hand-motion infilling in HOT3D and ARCTIC, covering off-image frames and missing detections. All methods share an *observation mask* that identifies observed and unobserved hand frames from the same HaWoR-YOLO detections. Consecutive unobserved frames of the same hand form a *gap*; all target gaps constitute the *requested gaps*.

Gap coverage is the percentage of requested gaps for which a method produces at least one valid motion prediction. The methods differ in their observation requirements: Dyn-HaMR requires an observed frame of the same hand before and after each gap (*two-sided anchors*), while HaWoR requires at least one frame with both hands observed within a 120-frame window.

For a fair comparison, we evaluate GHM separately on the same gaps successfully handled by each baseline, denoted as GHM^{H} and GHM^{D}, respectively, in Tab. 5. We also evaluate *GHM (all)* on all requested gaps to measure its overall coverage. For each matched comparison, errors are computed on the intersection of the baseline's and GHM's finite predictions and the fixed target mask, then averaged equally across evaluable gaps. The all-gap results use all valid predictions from GHM with the same equal-gap averaging.

Table 5: **Hand-motion infilling on HOT3D and ARCTIC.** Results aggregate both hands with equal weight per evaluable gap.

	Method	Gap cov.	MPJPE	PA	Gap-W	Gap-WA
HOT3D	HaWoR	92.9%	155.61	8.50	64.54	31.58
	GHM^{H}	92.9%	**102.73**	**6.76**	**49.97**	**25.24**
	Dyn-HaMR full	45.1%	814.63	10.73	84.48	34.80
	GHM^{D}	45.1%	**91.11**	**6.68**	**49.26**	**25.92**
	GHM (all)	99.1%	104.09	6.88	50.30	25.39
ARCTIC	HaWoR	95.8%	228.69	15.97	94.65	50.61
	GHM^{H}	95.8%	**79.78**	**8.85**	**49.63**	**22.33**
	Dyn-HaMR full	62.6%	98.56	13.22	44.63	24.45
	GHM^{D}	62.6%	**64.11**	**8.75**	**32.83**	**17.93**
	GHM (all)	100.0%	81.45	8.88	50.89	22.74

Gap cov.: gap coverage; PA: PA-MPJPE; $^{\text{H}}/^{\text{D}}$: matched to HaWoR/Dyn-HaMR. all: all valid predictions.

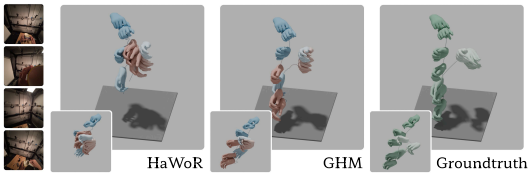


Figure 5: **Qualitative comparison of hand-motion infilling on HOT3D.** Selected input video frames are shown on the left, followed by HaWoR, GHM, and the ground-truth motion. Insets show close-up views of the hand motion.

We use world-space predictions for infilling because the camera-space branch lacks direct visual observations at missing frames. Motion is expressed in the frame-zero coordinate system, with global orientations recovered by accumulating predicted relative rotations. Gap-W and Gap-WA measure world-space joint errors after aligning each predicted gap to ground truth with a single

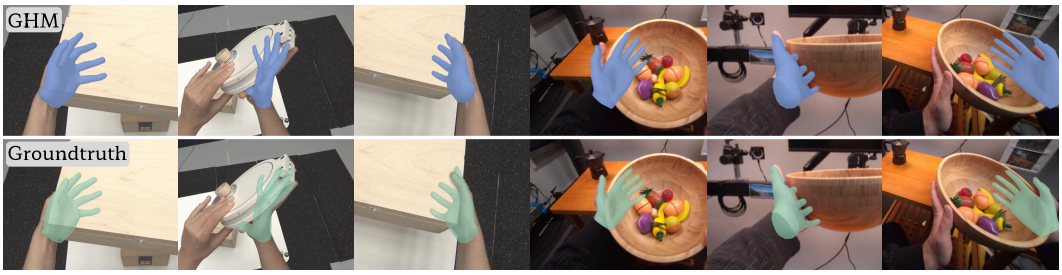


Figure 6: **Hand infilling results under heavy object occlusion or self-occlusion, where YOLO detection fails.** The first row shows GHM, and the second row shows Groundtruth.

similarity transform (scale, rotation, and translation). This transform is fitted using the first two valid frames for Gap-W and all valid frames for Gap-WA, using the common hand-frames within each gap for matched comparisons.

On matched hand-frames, GHM outperforms both HaWoR and Dyn-HaMR across all four error metrics. The HaWoR comparison covers 92.9% of HOT3D gaps and 95.8% of ARCTIC gaps. Across the full requested population, GHM returns valid predictions for 99.1% of HOT3D gaps and every ARCTIC gap. The matched comparisons show that the accuracy gains persist on the same evaluation frames, while the broader gap coverage extends the range of missing-observation patterns that the model can handle. Additional infilling details are provided in Appendix C.3.

4.4 Ablation Studies

Table 6: **Dual-mode training ablation on the HOT3D test set.**

Training	MPJPE	PA-MPJPE	RTE
Regression only	22.69	5.99	1.58
Mixed (dual-mode)	**20.58**	**5.79**	**1.40**

Table 7: **Cross-representation consistency ablation on the HOT3D test set.**

Setting	MPJPE	W-MPJPE	WA-MPJPE
w/o consistency	19.70	53.51	25.15
w/ consistency	**19.27**	**43.40**	**18.82**

Dual-mode training. We compare dual-mode training with a regression baseline to evaluate the impact of the generation objective on reconstruction. As shown in Tab. 6, dual-mode training improves joint-position and wrist-trajectory accuracy compared with regression-only training. The reductions in both MPJPE and PA-MPJPE suggest that incorporating the generation objective benefits reconstruction beyond absolute positioning, with gains retained after Procrustes alignment.

Cross-representation consistency. We examine whether coupling the camera- and world-space representations improves the directly predicted world-space motion. We disable the rotation and translation consistency terms in Eq. 3.

As shown in Tab. 7, consistency improves world-space reconstruction under both W-MPJPE and WA-MPJPE, with larger gains than the slight improvement in camera-space MPJPE. In particular, the reduction in WA-MPJPE persists after full-sequence similarity alignment, supporting the benefit of consistency for world-space hand-motion reconstruction. World-space errors are computed from directly decoded world-space predictions.

5 Discussion

In this work, we presented Generative Hand Modeling (GHM), a unified generative model for egocentric hand motion reconstruction, infilling, and text-conditioned generation under different input conditions.

A remaining challenge is to learn rich motion priors from large collections of egocentric videos with limited contact information while ensuring precise hand–object interactions. Future work could, for example, explore joint modeling of hand and object motion to improve the precision of hand–object interactions. The goal is to benefit from diverse human motion patterns without compromising interaction accuracy.

AI USE STATEMENT

AI tools were used for two purposes: (1) to polish the language and improve the readability of the manuscript; and (2) to generate text annotations for datasets to construct text inputs for generation. They did not contribute to the research ideas or methodology. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses only publicly available datasets, in accordance with their respective licenses, terms of use, and access requirements.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide information about the training data in Appendix B and describe the model architecture in Section 3.2. We will release the code as open source in the future.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071. IEEE, 2025.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. Hmp: Hand motion priors for pose and shape estimation from video. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6341–6351, 2024. doi: 10.1109/WACV57701.2024.00623.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17461–17474, 2025. doi: 10.1109/CVPR52734.2025.01627.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

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 (ICCV)*, pp. 10138–10148, October 2021.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A GENeralist model for human MOtion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17416–17425, June 2025b.

Yufei Liu, Xixi Wang, Hao Li, Ganlong Zhao, Kaitong Cai, Chengkai Jin, Chunxiao Liu, Jianbo Liu, Siyuan Huang, Xingang Pan, and Hongsheng Li. Ace-ego-hand: Repurposing video diffusion models for occlusion-robust egocentric 3d hand motion recovery, 2026.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual Tool-ACtion-Object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751. IEEE, 2024.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21013–21022. IEEE, 2022.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, 2025.

Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7026–7036, June 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130883.

Ropedia. Xperience-10M: A large-scale egocentric multimodal dataset with structured 3D/4D annotations, 2026. URL `https://huggingface.co/datasets/ropedia-ai/xperience-10m`. Dataset.

Mingyi Shi, Xuelin Chen, and Taku Komura. Prior-first, condition-second: Scalable and controllable hand motion completion. *Computer Graphics Forum*, 2026. doi: 10.1111/cgf.70568. ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2026.

Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. UniHand: A unified model for diverse controlled 4D hand motion modeling. In *International Conference on Learning Representations*, 2026.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

Balamurugan Thambiraja, Omid Taheri, Radek Danecek, Giorgio Becherini, Gerard Pons-Moll, and Justus Thies. CLUTCH: Contextualized language model for unlocking text-conditioned hand motion modelling in the wild. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20270–20281, 2023.

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

Mingxi Xu, Bowen Duan, Yi Gu, Zhengyang Shen, Renjing Xu, and Yutao Yue. HandFlow: Fully generative 4D hand recovery with flow matching. *arXiv preprint arXiv:2607.11221*, 2026.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27716–27726. IEEE, 2025.

Huajian Zeng, Chaohua Yao, Yuantai Zhang, Jiaqi Yang, Rolandos Alexandros Potamias, and Xingxing Zuo. Stablehand: Quality-aware flow matching for world-space dual-hand motion estimation from egocentric video. In *Advances in Neural Information Processing Systems*, 2026.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456. IEEE, 2024.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815, 2025.

Zimu Zhang, Yucheng Zhang, Xiyan Xu, Ziyin Wang, Sirui Xu, Kai Zhou, Bing Zhou, Chuan Guo, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Handx: Scaling bimanual motion and interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2274–2284, June 2026.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. In *Conference on Robot Learning (CoRL)*, 2026.

Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. Megohand: Multimodal egocentric hand-object interaction motion generation. *Advances in Neural Information Processing Systems*, 38:49464–49490, 2026.

Zijie Zhu, Weiren Cai, Yizhou Wang, Zhenjie Yang, Yide Liu, Jiahao Chen, and Guanqi He. MINT: A unified model for world-space camera and hand motion estimation from scalable egocentric pipeline supervision. *arXiv preprint arXiv:2609.04958*, 2026.

Section A presents additional qualitative results, and Section B describes the data and implementation details. Section C specifies the evaluation protocols, metrics, and gap construction. Section D reports additional analyses of geometric conditioning, reconstruction, and infilling.

A Additional Qualitative Results

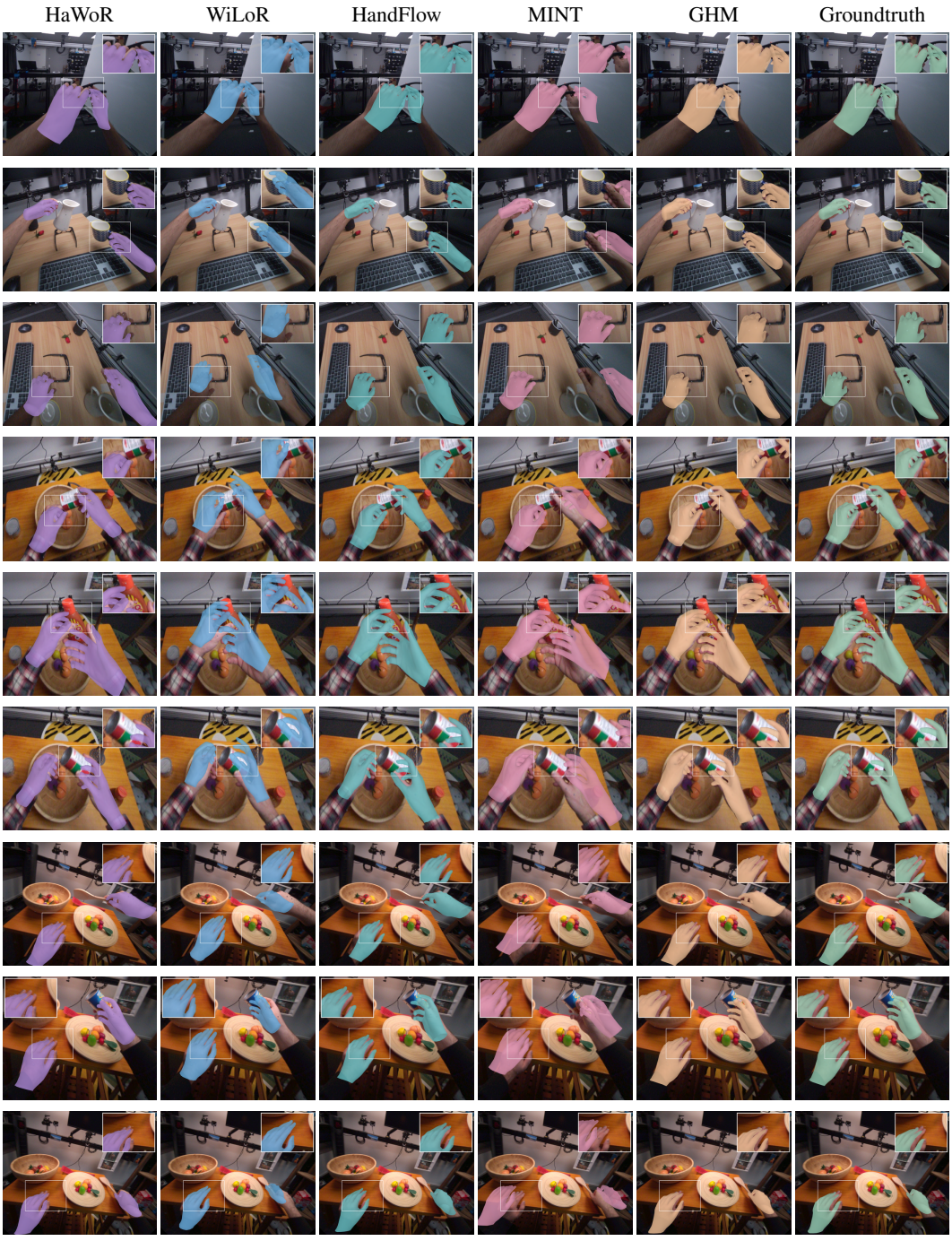
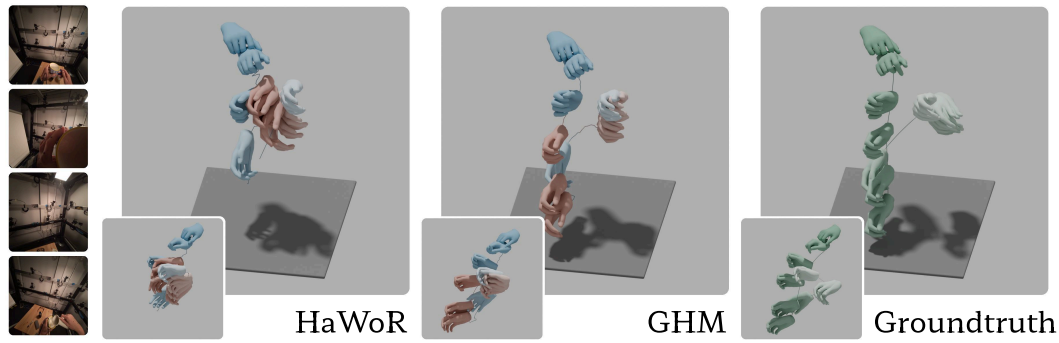
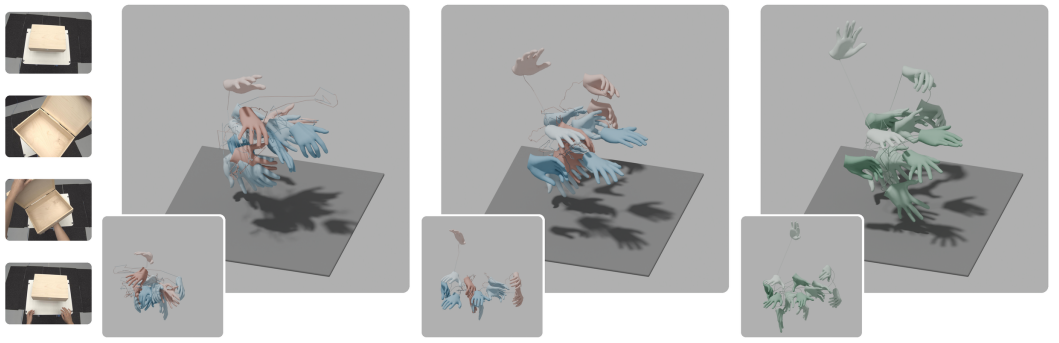


Figure 7: **Qualitative comparison of hand reconstruction on HOT3D.** Each row shows a different example. From left to right, the columns show HaWoR, WiLoR, HandFlow, MINT, GHM, and ground truth (GT). Hand meshes are overlaid on the input images, with insets highlighting local hand-pose details.



(a) HOT3D



(b) ARCTIC

Figure 8: **World-coordinate visualization of hand-motion infilling on HOT3D and ARCTIC.** Selected input video frames appear at the left of each row. The three motion panels show HaWoR, GHM, and ground truth (GT) from left to right in both rows. Insets provide additional views of the hand motion.



Figure 9: **Additional hand-motion infilling examples under occlusion on ARCTIC.** Eight examples are arranged in two groups of four columns. Within each group, the upper row shows GHM (blue) and the lower row shows ground truth (GT; green), with hand meshes overlaid on the corresponding input images. The examples include both hands and different manipulated objects.

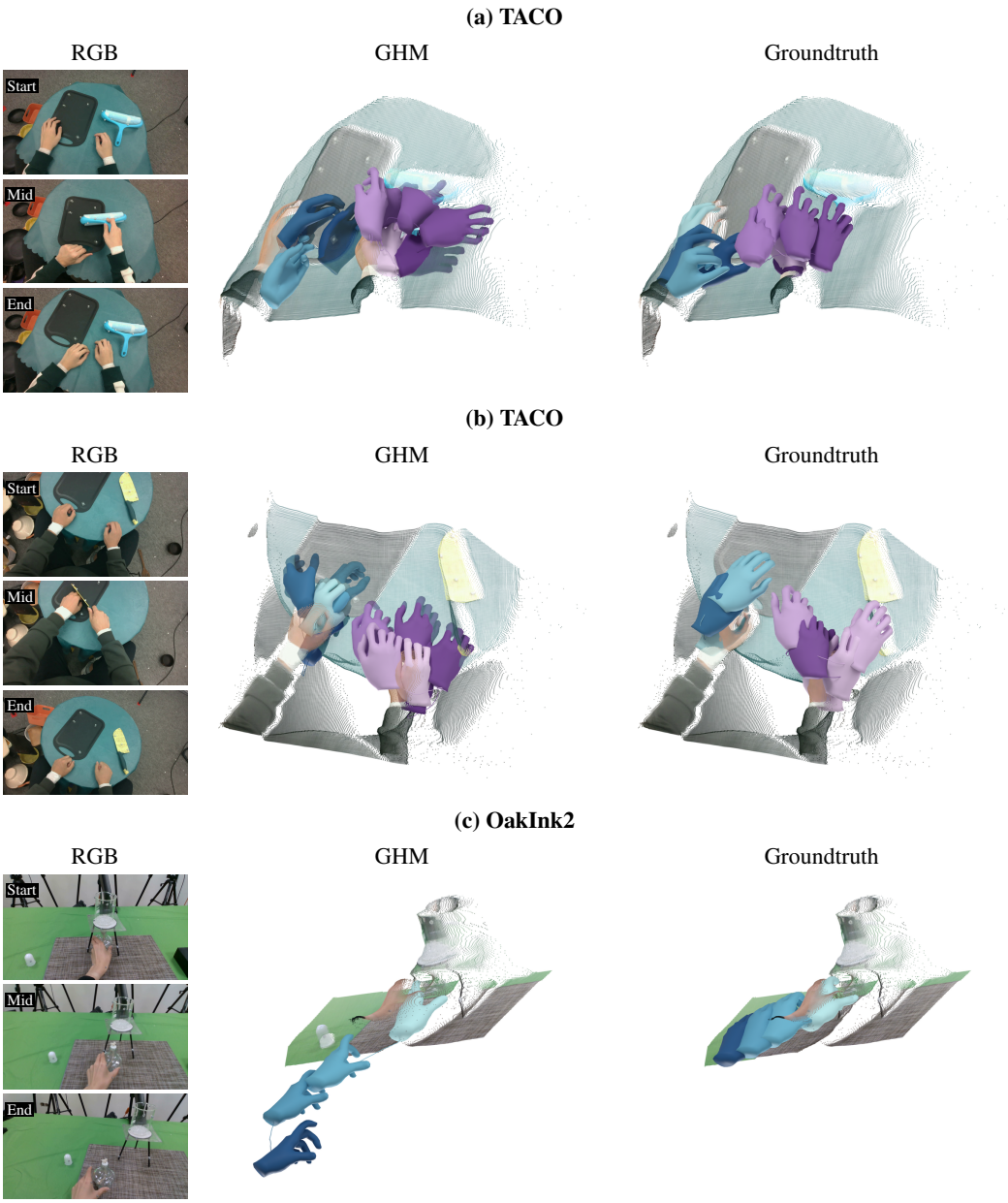


Figure 10: **Hand-motion generation examples on TACO and OakInk2.** Each example shows the Start, Mid, and End RGB frames on the left, GHM in the middle, and ground truth (GT) on the right. RGB frames provide visual context; the generated and GT motions are sampled independently. The point cloud depicts the initial scene. Blue and purple denote left and right hands, with darker shades indicating later times.

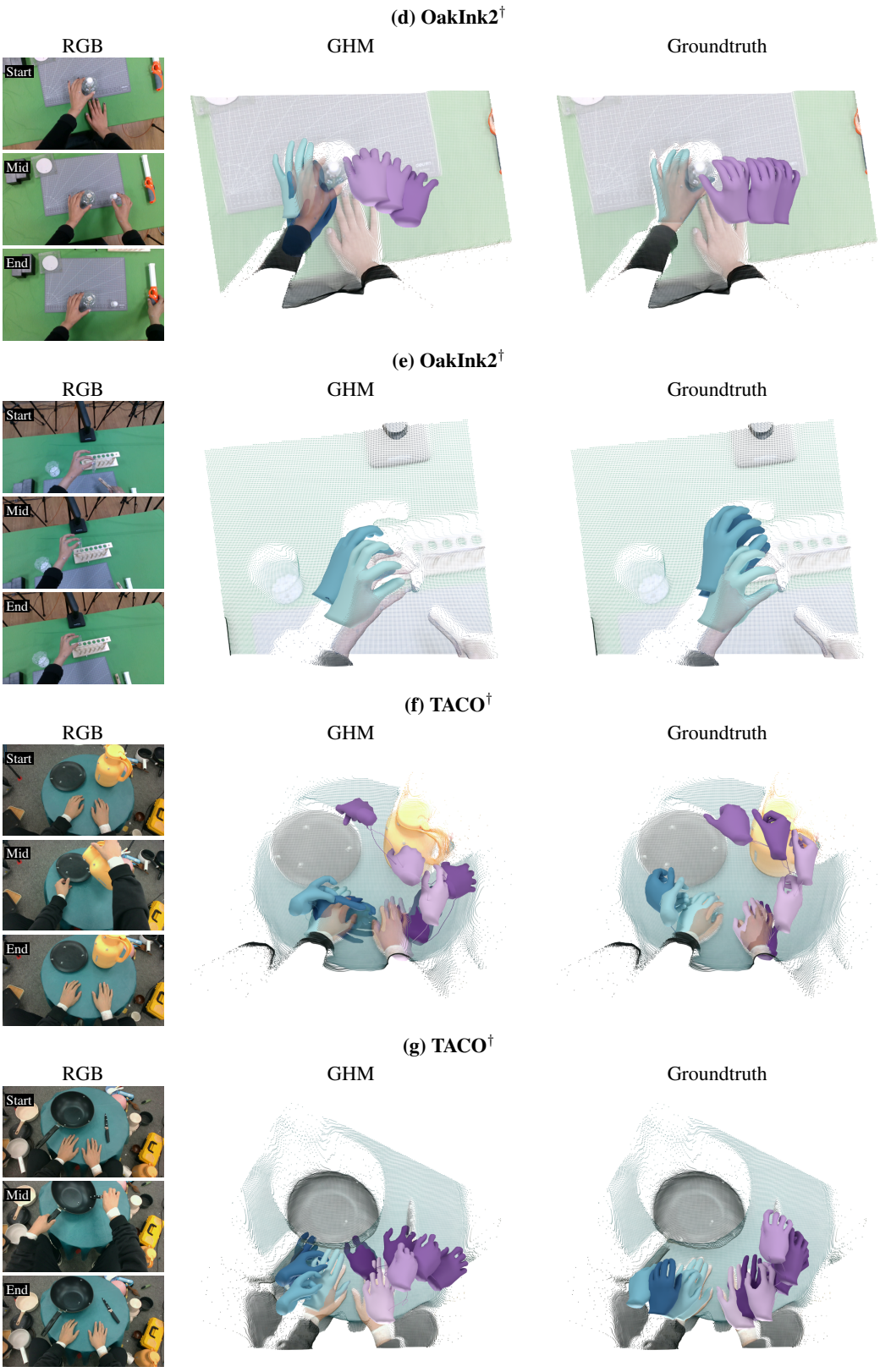


Figure 10: **Hand-motion generation examples on TACO and OakInk2 (continued).**

†: Enlarged examples from Figure 4 in the main paper.

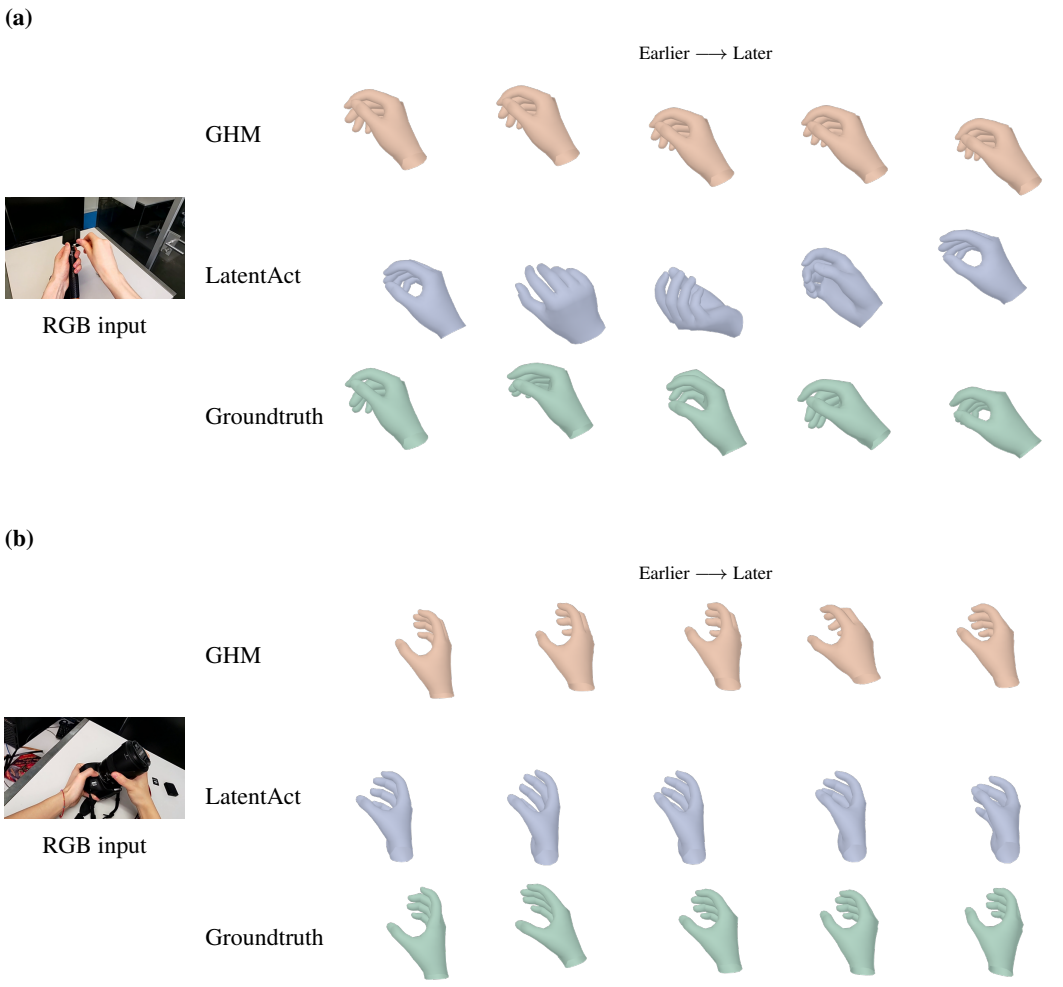


Figure 11: **Qualitative hand-motion generation comparison on HoloAssist.** Each example includes an input RGB image on the left and five selected right-hand poses per method on the right. The rows show GHM, LatentAct, and Groundtruth, respectively. Poses progress chronologically from left to right within each row. Keyframes are selected independently for each method; columns do not imply frame-wise temporal correspondence.

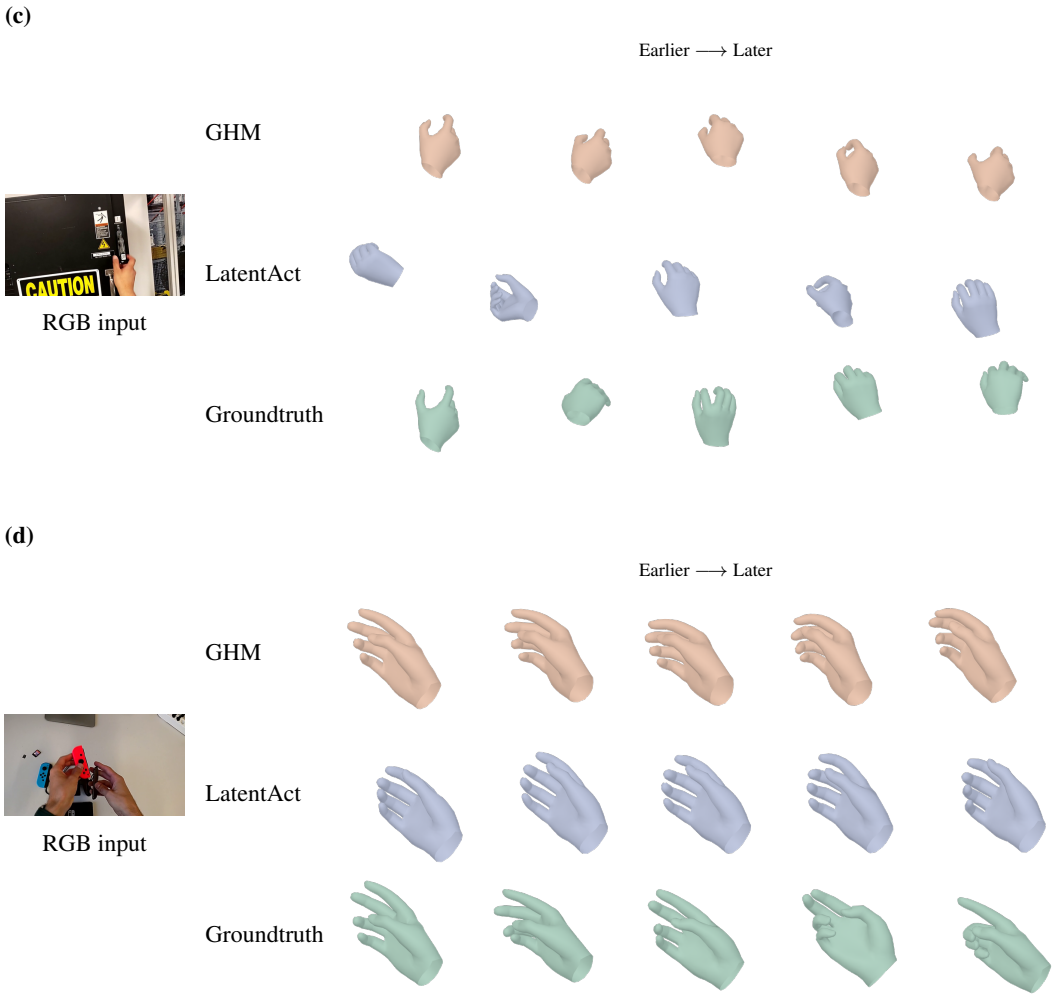


Figure 11: **Qualitative hand-motion generation comparison on HoloAssist (continued).** Rows show GHM, LatentAct, and Groundtruth. Each row progresses chronologically; columns are not temporally aligned across methods.

B Data and Implementation Details

B.1 Implementation Details

Model and optimization. Our main training configuration uses a 12-layer denoising transformer with a hidden dimension of 512, 16 attention heads, an MLP expansion ratio of 8, and dropout of 0.1. We implement the model in PyTorch and train it on eight GPUs with bfloat16 mixed precision. We use AdamW with a peak learning rate of 10^{-4}, weight decay of 0.01, and a global batch size of 128 clips.

Learning rate schedule. The learning rate increases linearly from 5×10^{-5} to 10^{-4} over the first 4,000 optimizer updates. It then follows a cosine decay to 5×10^{-5} at update 200,000 and remains constant thereafter.

Per-frame condition features. Per-hand appearance features are extracted from hand crops using a frozen, pretrained HaMeR image encoder. The 2D bounding-box features are normalized by focal length.

Scene and text features. Text is encoded by a frozen language encoder, while RGB and metric-depth XYZ patch features from the first frame provide the scene tokens. HaMeR appearance features, text embeddings, and DA3 depth estimates are precomputed offline.

B.2 Training Corpus Statistics

Table 8 reports the number of source motion sequences and frames contributed by each training dataset before temporal windowing, together with their cumulative duration. We estimate duration as the frame count divided by the frame rate and by 3,600, using 20 fps for Xperience and 30 fps for the other datasets. The corpus totals approximately 568.61 hours, excluding repeated frames introduced by sliding windows. TACO is included in training; H2O is used only for evaluation and is excluded from these statistics. We convert the sequences into clips of at most 120 frames using a sliding window with a stride of 10, while retaining shorter sequences as individual clips. HOT3D, OakInk2, TACO, HOI4D, ARCTIC, and EgoDex provide reconstruction supervision; their available text–motion pairs, together with Xperience and HoloAssist, provide generation supervision.

Table 8: **Training-corpus statistics before temporal windowing.**

Dataset	Entries	Frames	FPS	Hours
HOT3D	3,919	503,640	30	4.66
OakInk2	1,518	424,248	30	3.93
TACO	856	138,922	30	1.29
HOI4D	2,957	402,705	30	3.73
ARCTIC	270	180,531	30	1.67
EgoDex	14,386	4,286,120	30	39.69
Xperience	114,399	36,014,808	20	500.21
HoloAssist	17,470	1,451,597	30	13.44
Total	**155,775**	**43,402,571**	—	**568.61**

B.3 Dataset Splits

HOT3D. For HOT3D training, we adopt the same training split as HaWoR. For evaluation, we use the test split adopted by HandFlow and evaluate on HOT3D-Clips, enabling direct comparison on the same held-out clips.

OakInk2. We use the official OakInk2 training set and further partition it into training and validation subsets using an action-based split.

HoloAssist. We train our model on a filtered subset of the official HoloAssist training split, removing all matched action clips assigned to LatentAct's `tasks/test` split. We evaluate on the matched LatentAct `tasks/test` candidate subset, which contains 60 actions: 52 from HoloAssist training videos and eight from validation videos. The training and evaluation sets are therefore action-disjoint, but the former 52 evaluation actions are not video-disjoint from the training set. The full-sequence comparison uses the 59 actions with matched outputs from both methods.

C Evaluation Protocols

C.1 Generation Evaluation Protocol

GHM generates each full action in one rollout, with or without its complete GT 2D wrist trajectory as an additional condition. For LatentAct, each inference window receives its starting RGB image and the authors' released, preprocessed 3D contact centroid. Full-action predictions are assembled from 30-frame windows without contact registration or cross-fading. We score the right hand in world space. Table 3 uses five samples (seeds 0–4) for each of the 59 matched actions. Both GHM variants use the same checkpoint, CFG scale 1, and five DDIM steps. LatentAct uses stochastic codebook sampling in this experiment. We average errors over all samples within each action, then equally across actions, without best-of-K selection.

C.2 Evaluation Metric Definitions

Reconstruction. W-MPJPE fits a similarity transform using the first two valid frames of each hand trajectory; WA-MPJPE fits it using all valid frames. Infilling applies the same alignment rules separately within each gap. RTE measures camera-space wrist-trajectory error, and acceleration error compares predicted and GT joint accelerations.

Pose and interaction geometry. Global PA-MPJPE fits a single scale, rotation, and translation to the complete joint trajectory, unlike per-frame PA-MPJPE. For wrist positions w_t, predictions $\hat{w}_t$, and a fixed 3D object reference point q obtained from the first frame, we define

$$E_{\min} = \left| \min_t \|\hat{w}_t - q\|_2 - \min_t \|w_t - q\|_2 \right|. \tag{4}$$

Min-Dist. is the absolute difference between the predicted and GT minimum wrist distances to the object reference point. For GO, prediction and GT independently select their closest-approach frame. We take a contiguous 20-sample window around each, shifted at sequence boundaries if necessary, and average the $\mathrm{SO}(3)$ geodesic angle between global hand orientations at matched window progress. GO also permits different approach times.

Wrist-trajectory DTW. We report the mean 3D wrist distance along the temporal alignment found by dynamic time warping (DTW). Absolute DTW uses world-space wrist positions; relative DTW first subtracts each trajectory's own initial wrist position.

Diversity. We compare all ten pairs of the five samples at matching progress, without DTW alignment. Wrist diversity averages pairwise 3D wrist distances over frames. Root-relative joint diversity additionally subtracts each sample's own wrist at every frame before averaging distances over the 21 joints; it retains hand rotation and is not a finger-only measure. We average the ten pair distances within each action, then weight all actions equally, retaining all samples.

C.3 Evaluation Gap Construction

We form gaps from consecutive target frames of the same hand. Observed frames require both a shared-YOLO detection and a HaMeR feature; the nearest observed frames on either side provide temporal anchors. Errors are computed per gap and averaged equally across gaps. Per-hand results are reported in Section D.3.

Table 9: **HOT3D gaps with no same-hand observation.***

Sequence	Target interval	Frames
`seq_202`	$[0, 3)$	3
`seq_215`	$[93, 144)$	51
`seq_249`	$[0, 125)$	125
Total		**179**

*All three are left-hand gaps, and the corresponding 150-frame clips contain no valid left-hand visual observation.

Table 10: **Distribution of evaluation-gap lengths.** Length is the number of target hand-frames in a gap.

Dataset	1–15	16–30	31–60	61+
HOT3D	148	66	70	42
ARCTIC	379	132	64	21

HOT3D. We follow the off-image target selection rule used by HaWoR (Zhang et al., 2025). A target hand-frame must have a raw MANO annotation, valid and finite completed ground truth, and fewer than two of its 21 projected joints inside the image. We remove visual observations from all target frames and use the same target set for every method.

ARCTIC. We divide sequences into non-overlapping windows of at most 120 frames. A valid hand-frame is a target if either its reference hand observation or its shared YOLO–HaMeR observation is missing. This includes in-image occlusions that cause missing observations. Gaps are formed within each window, yielding 596 gaps and 9,540 target hand-frames from 37 sequences.

Evaluation subsets. Dyn-HaMR is evaluated on gaps with anchors on both sides: 147 in HOT3D and 373 in ARCTIC. For HaWoR, errors use target hand-frames with finite predictions from both HaWoR and GHM. Unsupported gaps remain in the all-gap coverage denominator. Table 10 summarizes the gap-length distribution.

Why HOT3D coverage is below 100%. On the all-target HOT3D population, GHM covers 323 of 326 gaps (99.1%) and 8,976 of 9,155 target hand-frames (98.0%). The three unsupported gaps comprise 179 frames and are all left-handed (Table 9). Although these frames have valid off-image ground truth for scoring, none of the corresponding clips contains a left-hand frame with both a shared-YOLO detection and a HaMeR feature.

Recovering absolute hand orientation from predicted frame-to-frame changes requires an observed frame of the same hand. The three HOT3D gaps without such an observation remain in the coverage denominator.

D Additional Analyses

D.1 Additional Geometric Conditioning

To test whether our condition fusion can accommodate an additional modality, we add 2D hand keypoints to the frame-aligned inputs.

Table 11 compares three input settings of the same fine-tuned checkpoint on HOT3D: no keypoints, GT-derived keypoints (oracle), and HaMeR-projected keypoints. Disabling keypoints is an inference-time input ablation, not a separately trained model. No synthetic keypoint corruption is applied at evaluation.

GT-derived keypoints improve reconstruction, demonstrating that GHM can exploit additional geometric constraints. With HaMeR-predicted keypoints, estimation errors may introduce misleading constraints that outweigh the benefit of the additional input.

Table 11: **Additional 2D keypoint conditioning on HOT3D.*** All rows use the same checkpoint and evaluation mask; the oracle row uses ground-truth-derived auxiliary inputs.

Keypoint input	MPJPE	PA-MPJPE
Disabled	20.14	5.73
GT-derived (oracle)	18.93	5.22
HaMeR-predicted	22.39	5.92

*Training noise magnitudes were set using HaMeR errors on this evaluation subset; these results constitute a development study.

D.2 Visible-Frame Reconstruction on ARCTIC

Table 12 reports supplementary reconstruction results on the ARCTIC validation subset.

Table 12: **Visible-frame reconstruction on ARCTIC with GT bounding boxes.** Results aggregate both hands on the common valid-prediction mask.

Method	MPJPE	Root MPJPE	PA-MPJPE	MPVPE	Root MPVPE
WiLoR	42.108	16.207	7.134	41.794	15.712
HaWoR	34.832	23.962	7.492	34.594	23.140
GHM	**31.704**	**15.618**	**7.110**	**31.599**	**15.197**

D.3 Per-Hand Infilling Results

Table 13 supplements Table 5 with separate left- and right-hand results for all-gap infilling.

Table 13: **Per-hand infilling results on HOT3D and ARCTIC.** We report equal-gap macro averages.

Dataset	Hand	Method	Gap coverage	MPJPE	PA-MPJPE	Gap-W	Gap-WA
HOT3D	Left	HaWoR	94.8%	142.03	8.58	65.45	31.60
		GHM (all)	98.6%	**101.85**	**6.68**	**51.56**	**25.31**
	Right	HaWoR	89.6%	181.97	8.36	62.70	31.53
		GHM (all)	100.0%	**108.14**	**7.24**	**47.96**	**25.53**
ARCTIC	Left	HaWoR	95.7%	221.31	15.99	109.48	55.32
		GHM (all)	100.0%	**91.77**	**9.23**	**60.67**	**24.68**
	Right	HaWoR	95.9%	235.08	15.95	82.05	46.53
		GHM (all)	100.0%	**72.49**	**8.58**	**42.58**	**21.07**