DYNAMICHOI: COUPLED DYNAMICS FOR PHYSICS-AWARE HOI RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study hand-object interaction (HOI) reconstruction from monocular RGB videos, where partial observations can produce visually plausible yet mechanically inconsistent trajectories. Existing methods mainly enforce visual and geometric agreement, leaving the underlying interaction dynamics insufficiently constrained. We propose DYNAMICHOI, a physics-aware HOI reconstruction framework combining geometry-grounded diffusion refinement with coupled hand-object dynamics. Geometry spatially grounds visual evidence for trajectory refinement, while articulated inverse dynamics and Newton-Euler dynamics derive hand generalized forces and object wrenches for dynamics-level supervision. We further couple hand and object dynamics through contact-force transfer and recover active hand actuation as an interaction-level physical quantity. We formulate its empirical magnitude distribution into a probabilistic prior that penalizes unlikely actuation and suppresses mechanically implausible reconstructed motion. Experiments on three HOI datasets show consistent improvements in both hand and object reconstruction. The reconstructed trajectories further benefit downstream applications including hand world-model generation and robotic manipulation learning, demonstrating the value of physics-aware HOI modeling beyond reconstruction.

1 INTRODUCTION

Hands are the primary interface through which humans perform everyday hand-object interactions (HOIs). Reconstructing these interactions from monocular videos into temporal 3D hand-object mesh trajectories is essential for immersive technologies (Zhou et al., 2023; Feng et al., 2020), embodied learning (Qin et al., 2022; Chen et al., 2025), and world modeling (Li et al., 2026). Monocular videos provide only partial observations under occlusion and viewpoint ambiguity, allowing multiple 3D trajectories to explain similar visual evidence and making reliable reconstruction challenging.

Recent work has substantially advanced HOI reconstruction. Optimization-based methods jointly fit hand, object, and scene representations (Huang et al., 2022; Ye et al., 2023; Fan et al., 2024), while feed-forward approaches exploit contact, occlusion, and learned interaction priors (Chen et al., 2022; Wang et al., 2025; Aboukhadra et al., 2026; Wang et al., 2026). Recent methods have also extended HOI reconstruction to egocentric videos (Zhang et al., 2025a; Fu et al., 2026; Ye et al., 2026). Despite this progress, two challenges remain. First, geometry is not fully exploited to ground visual representations, limiting ambiguity resolution under occlusion and restricted viewpoints (Ye et al., 2026). More importantly, reconstruction is typically supervised at the trajectory level, which encourages geometric accuracy but does not explicitly constrain the dynamics implied by the estimated motion. For example, a reconstructed object may suddenly translate or rotate during manipulation without a corresponding force being exerted by the hand. Dynamics provides complementary supervision by constraining the forces and torques required to produce reconstructed motion.

Introducing such dynamical constraints remains difficult because physical quantities such as forces are not directly observable from videos. Zhang et al. (2024) and Ismayilzada et al. (2026) pioneered articulated hand dynamics by deriving generalized forces from reconstructed trajectories through inverse dynamics, providing physical supervision beyond geometric states. However, these formulations characterize only the hand and do not account for the dynamics of the manipulated object. In HOI, object motion is governed by its own rigid-body dynamics, while the hand and object are mechanically coupled through contact: active hand actuation produces hand motion and generates

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.

interaction forces that drive the object. Therefore, modeling hand dynamics alone or treating the two bodies independently leaves this mechanical dependency unconstrained. Physics-aware HOI reconstruction thus requires both hand and object dynamics as well as their physical coupling through contact to form a unified dynamically-constrained system.

To address these challenges, we propose DYNAMICHOI, a physics-aware HOI reconstruction framework that integrates geometry-grounded visual refinement with coupled hand-object dynamics. To better resolve visual ambiguity, we encode the hand-object geometry and use it to spatially ground visual evidence within a diffusion-based iterative refinement, providing geometry-aligned guidance for trajectory estimation. To further constrain the motion itself, we derive hand generalized forces through articulated inverse dynamics (Zhang et al., 2025b; Ismayilzada et al., 2026) and rigid-object wrenches through Newton-Euler dynamics (Luh et al., 1980; Khalil, 2011; Hu et al., 2022; Lutter, 2023). Aligning these motion-implied physical quantities between reconstructed and ground-truth trajectories introduces dynamics-level supervision beyond trajectory alignment.

We further couple the hand and object through contact-force transfer (Hu et al., 2022; Nakajima et al., 2022). Active hand actuation jointly accounts for the generalized force required by hand motion and the contact reaction induced by driving the object, providing a shared physical explanation for both trajectories. Our key insight is that active actuation during everyday manipulation follows a characteristic statistical distribution (Tanghe et al., 2019). We estimate this empirical distribution and impose its negative log-likelihood as a probabilistic regularizer, penalizing reconstructions that induce low-probability actuation. Together, the hand dynamics, object dynamics, and actuation prior form a coupled dynamical system that provides physically grounded supervision for HOI reconstruction.

Experiments across three HOI reconstruction datasets (Banerjee et al., 2025; Chao et al., 2021; Hampali et al., 2020) demonstrate that DYNAMICHOI consistently improves both hand and object reconstruction. We further use the reconstructed trajectories to condition hand world-model video generation (Li et al., 2026) and train dexterous robotic manipulation (Chen et al., 2025). Improvements in both applications show that physics-aware reconstruction provides more useful motion signals beyond reconstruction accuracy. In summary, our contributions are threefold:

- We present a geometry-grounded HOI reconstruction framework that uses 3D geometry to spatially ground visual evidence during diffusion refinement for more accurate trajectory estimation.
- We introduce dynamics-level supervision by aligning hand generalized forces and object wrenches between reconstructed and ground-truth trajectories to encourage physically consistent motion.
- We couple hand-object dynamics through contact-force transfer and regularize active hand actuation with a probabilistic prior to suppress reconstructions that induce implausible actuation.

2 RELATED WORK

Hand and object reconstruction. Image-based hand models such as HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025) learn strong visual priors for mesh regression, while video-based methods such as HaWoR (Zhang et al., 2025a) and Dyn-HaMR (Yu et al., 2025) further exploit temporal context and camera motion to recover global hand trajectories. For objects, FoundationPose (Wen et al., 2024) conditions 6D pose estimation on object geometry, whereas ForeHOI (Chen et al., 2026) combines visual evidence with 3D shape priors to reconstruct objects under hand occlusion. However, these methods do not explicitly leverage joint 3D hand-object geometry to organize visual evidence. In contrast, our framework grounds visual cues in the 3D HOI geometry within diffusion refinement, improving reconstruction under occlusion and monocular ambiguity.

Hand-object interaction reconstruction. Joint HOI reconstruction reduces monocular ambiguity by exploiting dependencies between the hand and object. Early methods jointly estimate or refine their geometry and poses using contact and non-penetration constraints (Hasson et al., 2019; Grady et al., 2021; Chen et al., 2022). Video-based approaches further incorporate differentiable rendering, implicit representations, and learned interaction priors (Ye et al., 2023; Huang et al., 2022; Fan et al., 2024), while recent methods improve generalization to limited viewpoints, unseen categories, articulated objects, and egocentric settings (Wang et al., 2025; Aboukhadra et al., 2026; Wang et al., 2026; Fu et al., 2026; Ye et al., 2026). These methods primarily enforce visual and geometric consistency of the motion, while our proposed framework focuses on modeling articulated-hand and rigid-object dynamics to constrain the physical requirements underlying their joint trajectories.

Physics-aware HOI modeling. Physics-aware methods improve mechanical plausibility through explicit dynamics or learned physical constraints. Physical Interaction estimates contact forces that satisfy rigid-object dynamics from RGB-D observations (Hu et al., 2022), while Physics-Aware HOI Denoising refines hand motion given an accurate object trajectory (Luo et al., 2024). Other hand-centric approaches introduce phase-dependent physical constraints (Zhang et al., 2025b) or probabilistic articulated dynamics (Ismayilzada et al., 2026). Nevertheless, existing methods do not jointly model hand dynamics, object dynamics, and their coupling through contact. We address this gap by coupling the two dynamical systems through contact forces and regularizing the resulting active hand actuation with a probabilistic prior.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Given an RGB video $I_{1:T} = \{I_t\}_{t=1}^{T}$ and the canonical mesh $\mathcal{M}^o$ of the interacting object [1], HOI reconstruction aims to recover the 3D hand and object mesh trajectories over T frames. Following prior work (Ismayilzada et al., 2026; Ye et al., 2026), we represent the reconstruction using compact parametric states and then deterministically decode them into meshes. Specifically, a reconstruction model $\mathcal{G}_\Theta$ with learnable parameters Θ predicts the joint hand-object trajectory during inference

$$\mathbf{x}_{1:T} = \left\{(\mathbf{x}_t^h, \mathbf{x}_t^o)\right\}_{t=1}^{T} = \mathcal{G}_\Theta\left(I_{1:T}, \mathcal{M}^o\right), \tag{1}$$

where $\mathbf{x}_t^h$ and $\mathbf{x}_t^o$ denote the hand and object parametric states at frame t, respectively. The hand state $\mathbf{x}_t^h = (\boldsymbol{\rho}_t^h, \mathbf{t}_t^h, \boldsymbol{\beta}^h)$ is represented by MANO parameters (Romero et al., 2017), where $\boldsymbol{\rho}_t^h \in \mathbb{R}^{16\times 6}$ contains the 6D rotations of the global hand orientation and the 15 articulated MANO joints, $\mathbf{t}_t^h \in \mathbb{R}^3$ denotes the global translation, and $\boldsymbol{\beta}^h \in \mathbb{R}^{10}$ represents the hand shape and remains fixed throughout the video. The estimated states $\mathbf{x}_{1:T}^h$ are decoded through the MANO layer (Romero et al., 2017) to obtain the hand mesh trajectory $\mathcal{M}_{1:T}^h$. The object state $\mathbf{x}_t^o = (\mathbf{R}_t^o, \mathbf{t}_t^o)$ is represented by its rigid pose, where $\mathbf{R}_t^o \in SO(3)$ and $\mathbf{t}_t^o \in \mathbb{R}^3$ denote its rotation and translation, respectively. The object mesh at frame t is obtained through the rigid transformation $\mathcal{M}_t^o = \mathbf{R}_t^o \mathcal{M}^o + \mathbf{t}_t^o$.

Trajectory-level supervision alone does not guarantee physically plausible motion. We therefore further constrain the hand and object dynamics individually, as well as their physical coupling during interaction. Let $\hat{\mathbf{x}}_{1:T}$ denote the ground-truth trajectory. We formulate the reconstruction objective as

$$\mathbf{x}_{1:T}^* = \arg\min_{\mathbf{x}_{1:T}\in\mathcal{X}} \left[\mathcal{C}_t\left(\mathbf{x}_{1:T}, \hat{\mathbf{x}}_{1:T}\right) + \lambda_d \mathcal{C}_d\left(\mathbf{x}_{1:T}, \hat{\mathbf{x}}_{1:T}\right) + \lambda_a \mathcal{C}_a\left(\mathbf{x}_{1:T}^h, \mathbf{x}_{1:T}^o\right)\right], \tag{2}$$

where $\mathcal{X}$ denotes the admissible trajectory space. The three constraints capture complementary aspects of reconstruction: $\mathcal{C}_t$ enforces trajectory alignment, $\mathcal{C}_d$ independently constrains the articulated-hand and rigid-object dynamics, and $\mathcal{C}_a$ constrains their coupled dynamics during interaction. The weights λ_d and λ_a balance the corresponding physical constraints. Eq. 2 only provides a conceptual decomposition of constraints; in practice, they are instantiated as training objectives for the model.

3.2 PHYSICS-AWARE HOI RECONSTRUCTION

Figure 1 shows the overall framework of DYNAMICHOI, consisting of frozen reconstruction estimators and a learnable trajectory refinement network we design. Specifically, a frozen hand estimator $\mathcal{E}^h$ (Potamias et al., 2025) takes the frame I_t as input and predicts hand parameters $\mathbf{y}_t^h$, while a frozen object estimator $\mathcal{E}^o$ (Wen et al., 2024) takes I_t and the object template $\mathcal{M}^o$ as input and predicts the object parameters $\mathbf{y}_t^o$, resulting an initial estimate of the hand and object trajectory $\mathbf{y}_{1:T} = \left(\mathbf{y}_{1:T}^h, \mathbf{y}_{1:T}^o\right)$. Because these estimations rely on framewise visual evidence, their predictions can suffer from temporal and physical inconsistency, particularly under monocular ambiguity and occlusion. We refine the initial estimates (Zhang et al., 2025b;a; Ismayilzada et al., 2026) by modeling visual-geometric evidence and physical characteristics through an iterative diffusion process.

3.2.1 REFINEMENT NETWORK

[1] Canonical mesh is the object 3D shape in a reference coordinate frame before any rotation or translation.

We formulate the refinement within a diffusion process (Ho et al., 2020; Peebles & Xie, 2023) that enables progressive correction of temporal incoherence through iterative denoising. Specifically, given the initial reconstruction $\mathbf{y}_{1:T}$, randomly-sampled diffusion steps n and other inputs, the refinement network f_Θ estimates the temporally consistent trajectory by

$$\mathbf{x}_{1:T}^{(n)} = f_\Theta\left(\mathbf{y}_{1:T}; n, I_{1:T}, \mathcal{M}^o\right). \tag{3}$$

f_Θ contains a forward noise-injection process and a learnable denoising process, which progressively refines the initial trajectory by visual-evidence grounding with 3D geometry and hand-object dependency modeling. During training, we first sample a diffusion step n and construct the corresponding noisy trajectory $\mathbf{z}_{n,1:T}$ through the forward noise-injection process:

$$\mathbf{z}_{n,1:T} = \hat{\mathbf{x}}_{1:T} + \eta_n^2\left(\mathbf{y}_{1:T} - \hat{\mathbf{x}}_{1:T}\right) + \eta_n \kappa \boldsymbol{\epsilon}, \tag{4}$$

where η_n controls the diffusion level, κ determines the stochastic noise magnitude, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes Gaussian noise. Similar to $\mathbf{y}_{1:T}$, we represent the noisy trajectory as $\mathbf{z}_{n,1:T} = \{(\mathbf{z}^h_{n,t}, \mathbf{z}^o_{n,t})\}_{t=1}^T$, comprising the noisy hand trajectory $\mathbf{z}^h_{n,1:T}$ and the noisy object trajectory $\mathbf{z}^o_{n,1:T}$. The denoising process is performed through the following modules:

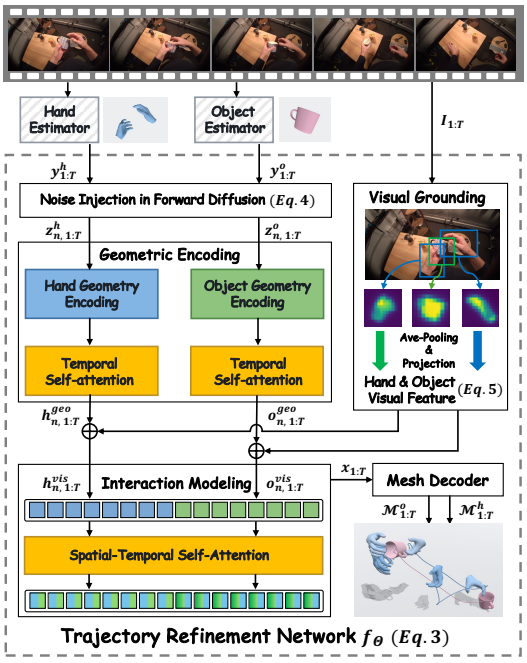


Figure 1: Framework overview. Framewise reconstruction models first estimate the initial trajectory in parametric states, which is then refined progressively by a diffusion-based refinement network, and finally decoded to HOI mesh trajectory.

Geometric encoding. We perform geometric encoding in mesh space to expose the explicit 3D structure hidden in the compact state parameters, enabling the denoising process to reason about hand-object spatial relationships and local surface geometry. The intermediate noisy states $\mathbf{z}_{n,1:T}$ are first converted into meshes by applying MANO layers to hand parameters and the rigid transformation to object parameters, which are subsequently encoded into latent features through separate geometry encoders. We further employ temporal self-attention layers to aggregate temporal information across the frame sequence, which produces the geometric feature sequences $\mathbf{h}^{\text{geo}}_{n,1:T}$ and $\mathbf{o}^{\text{geo}}_{n,1:T}$.

Visual grounding. The geometric features capture explicit 3D structure but lack appearance information from the observed video. We therefore ground them with spatially corresponding visual evidence. Specifically, a frozen visual encoder (Oquab et al., 2024) extracts a patch-level feature map $\mathbf{F}_t$ from each frame I_t. To establish spatial correspondence, we derive 3D geometric anchors from the initial hand and object estimates $\mathbf{y}^h_t$ and $\mathbf{y}^o_t$, using MANO joints for the hand and pose-transformed canonical keypoints for the object. These anchors are projected onto the image to retrieve and aggregate their corresponding visual features ζ^h_t and ζ^o_t:

$$\zeta^m_t = \frac{1}{|\mathcal{Q}^m_t|} \sum_{\mathbf{p} \in \mathcal{Q}^m_t} \mathbf{F}_t(\Pi_t(\mathbf{p})), \qquad m \in \{h, o\}, \tag{5}$$

where $\mathcal{Q}^h_t$ and $\mathcal{Q}^o_t$ denote the sets of visible hand and object anchors, respectively, and Π_t denotes the camera projection function. The resulting geometry-aligned visual features are fused with their corresponding geometric features through residual projections: $\mathbf{h}^{\text{vis}}_{n,t} = \mathbf{h}^{\text{geo}}_{n,t} + \phi^h(\zeta^h_t)$ and $\mathbf{o}^{\text{vis}}_{n,t} = \mathbf{o}^{\text{geo}}_{n,t} + \phi^o(\zeta^o_t)$, where ϕ^h and ϕ^o are learnable projection layers. This produces the visually grounded feature sequences $\mathbf{h}^{\text{vis}}_{n,1:T}$ and $\mathbf{o}^{\text{vis}}_{n,1:T}$.

Interaction modeling. Although the hand and object features are visually grounded, they are still encoded separately, which does not explicitly capture their mutual dependencies. We therefore concatenate $\mathbf{h}^{\text{vis}}_{n,1:T}$ and $\mathbf{o}^{\text{vis}}_{n,1:T}$ and jointly process them with spatial-temporal self-attention layers (Dosovitskiy et al., 2021) to model HOI across time. The output tokens are then separated into the interaction-aware hand and object features according to the original token positions, with the

first T tokens forming $\tilde{\mathbf{h}}_{n,1:T}$ and the remaining T tokens forming $\tilde{\mathbf{o}}_{n,1:T}$. Separate prediction heads regress residual updates to the initial hand and object states as $\mathbf{x}_t^{h,(n)} = \mathbf{y}_t^h + g^h(\tilde{\mathbf{h}}_{n,t})$ and $\mathbf{x}_t^{o,(n)} = \mathbf{y}_t^o + g^o(\tilde{\mathbf{o}}_{n,t})$, where g^h and g^o are corresponding MLP projectors. We concatenate them as $\mathbf{x}_t^{(n)} = [\mathbf{x}_t^{h,(n)}, \mathbf{x}_t^{o,(n)}]$ and collect over frames to obtain $\mathbf{x}_{1:T}^{(n)} = \{\mathbf{x}_t^{(n)}\}_{t=1}^T$ as the output of f_Θ.

Diffusion objective. We instantiate the trajectory-level constraint $\mathcal{C}_t$ in Eq. 2 with the diffusion objective $\mathcal{L}_{\text{diff}}$. Specifically, we train the diffusion-based refinement network f_Θ to recover the ground-truth trajectory from noisy trajectories sampled at different diffusion levels. We align the estimated trajectory $\mathbf{x}_{1:T}^{(n)}$ with the ground-truth $\hat{\mathbf{x}}_{1:T}$ by

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{n,\boldsymbol{\epsilon}}\left[d_x\left(\hat{\mathbf{x}}_{1:T}, \mathbf{x}_{1:T}^{(n)}\right)\right], \tag{6}$$

where d_x measures the trajectory discrepancy in rotation and translation spaces. This objective enables f_Θ to recover accurate HOI trajectories through joint visual-geometric representation learning.

3.2.2 Hand and Object Dynamics

Although the diffusion objective enables trajectory reconstruction, it does not explicitly account for the underlying dynamics that produce the hand and object motion, which may lead to physically implausible behaviors such as abnormal accelerations or abrupt rotations. To promote physical consistency beyond visual plausibility, we derive the hand and object dynamics from their trajectories and align the resulting physical quantities between the reconstructed and ground-truth motions.

Hand dynamics. Following previous work (Zhang et al., 2024; Ismayilzada et al., 2026), we regard the human hand as an articulated body and derive the generalized force required to produce its motion. The 16 hand rotations $\boldsymbol{\rho}_t^h$ are converted into Euler-ZXY angles $\boldsymbol{\phi}_t^h \in \mathbb{R}^{48}$, corresponding to the global orientation and 15 articulated MANO joints. Together with the global translation $\mathbf{t}_t^h$, we define the generalized coordinates as $\mathbf{q}_t^h = [(\mathbf{t}_t^h)^\top, (\boldsymbol{\phi}_t^h)^\top]^\top \in \mathbb{R}^{51}$. We use temporal finite differences between adjacent frames to estimate the generalized velocity $\dot{\mathbf{q}}_t^h$ and acceleration $\ddot{\mathbf{q}}_t^h$. Given these kinematic quantities, inverse dynamics is used to recover the resultant generalized force:

$$\boldsymbol{\tau}_t^h = M^h(\mathbf{q}_t^h)\ddot{\mathbf{q}}_t^h + C^h(\mathbf{q}_t^h, \dot{\mathbf{q}}_t^h)\dot{\mathbf{q}}_t^h + \mathbf{g}_{\text{grav}}^h(\mathbf{q}_t^h), \tag{7}$$

where M^h, C^h, and $\mathbf{g}_{\text{grav}}^h$ denote the hand mass matrix, Coriolis and centrifugal terms, and gravitational term, respectively. Applying Eq. 7 over the frame sequence yields the generalized-force trajectory $\{\boldsymbol{\tau}_t^h\}_{t=1}^T$, which reveals the dynamic requirements for driving hand movement.

Object dynamics. We introduce the Newton–Euler formalism (Luh et al., 1980; Khalil, 2011; Hu et al., 2022; Lutter, 2023), a fundamental framework for inverse dynamics in robotics and rigid-body control, into HOI reconstruction to relate object motion to its underlying physical requirements. Specifically, as discussed in Sect. 3.1, we model the object as a rigid body with pose $\mathbf{x}_t^o = (\mathbf{R}_t^o, \mathbf{t}_t^o)$. We compute its center-of-mass linear acceleration $\mathbf{a}_t^o$ from the translation trajectory $\{\mathbf{t}_t^o\}_{t=1}^T$, and derive its angular velocity $\boldsymbol{\omega}_t^o$ and angular acceleration $\dot{\boldsymbol{\omega}}_t^o$ from the rotation trajectory $\{\mathbf{R}_t^o\}_{t=1}^T$ using temporal finite differences between adjacent frames. The physical effort required to produce this motion is thus represented by its object wrench:

$$\mathbf{w}_t^o = \begin{bmatrix}\mathbf{f}_t^o \\ \boldsymbol{\tau}_t^o\end{bmatrix} = \begin{bmatrix} m^o(\mathbf{a}_t^o - \mathbf{g}) \\ \mathbf{I}_t^o\dot{\boldsymbol{\omega}}_t^o + \boldsymbol{\omega}_t^o \times (\mathbf{I}_t^o\boldsymbol{\omega}_t^o)\end{bmatrix}, \tag{8}$$

where $\mathbf{f}_t^o$ and $\boldsymbol{\tau}_t^o$ denote the non-gravitational external force and torque required to drive the translational and rotational motion of the object, respectively. $\mathbf{g}$ is the gravitational acceleration, $\mathbf{I}_t^o$ is the inertia tensor in the world frame, and m^o denotes the object mass. We use the object mass provided by the dataset when available; otherwise, we employ a large language model (OpenAI, 2026a) to estimate it. This formulation maps the object-pose trajectory to its wrench profile $\{\mathbf{w}_t^o\}_{t=1}^T$, providing an explicit dynamics-based characterization of object motion. Additional theoretical and implementation details are provided in the supplementary material.

Dynamics objective. The hand generalized forces and object wrenches derived above provide a dynamics-level description of the motion. We instantiate the dynamics constraint $\mathcal{C}_d$ in Eq. 2 by aligning the mechanical quantities implied by the reconstructed trajectory with those derived from the ground-truth trajectory. For frame t, we compute the hand generalized force and object wrench

$(\boldsymbol{\tau}_t^{h,(n)}, \mathbf{w}_t^{o,(n)})$ at diffusion step n from the predicted trajectory, and $(\hat{\boldsymbol{\tau}}_t^h, \hat{\mathbf{w}}_t^o)$ from the ground-truth trajectory. The resulting dynamics objective is

$$\mathcal{L}_{\text{dyn}} = \sum_{t=1}^{T} \left\| \boldsymbol{\tau}_t^{h,(n)} - \hat{\boldsymbol{\tau}}_t^h \right\|_2^2 + \sum_{t=1}^{T} \left\| \mathbf{w}_t^{o,(n)} - \hat{\mathbf{w}}_t^o \right\|_2^2. \tag{9}$$

This objective complements the diffusion supervision by enforcing consistency in the mechanical quantities underlying the motion, thereby discouraging dynamically inconsistent trajectories.

3.2.3 Coupled HOI Dynamics

Even if individually plausible, hand and object dynamics may still be mutually inconsistent, since physically valid HOI requires the same interaction to explain both motions (Hu et al., 2022; Nakajima et al., 2022). Active hand actuation provides such a joint physical description by accounting for both the generalized force required to produce hand motion and the interaction force exerted to drive object motion. Therefore, we use it to couple dynamics and regularize it with a probabilistic prior (Tanghe et al., 2019) to encourage physically plausible joint motion.

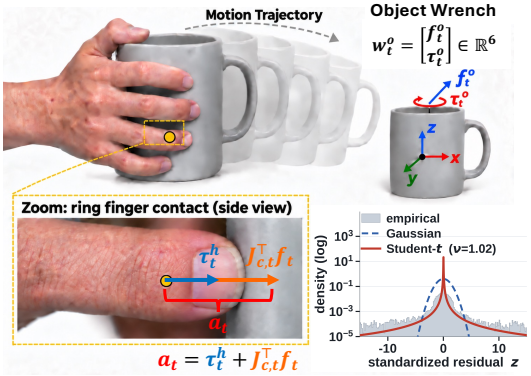


Figure 2: Coupled HOI dynamics. The bottom-right plot shows the empirical distribution of standardized active actuation $z = (a_{t,d} - \mu_d)/\sigma_d$ for an example DoF. The logarithmic density axis highlights the heavy-tailed behavior, which is empirically better captured by the Student-t distribution than the Gaussian distribution.

To recover active hand actuation, we first infer the contact forces exerted by the hand to drive the object motion. Since we focus on in-air manipulation, where the object is primarily driven by hand contact, we do not explicitly model external supports such as table contact. We approximate distributed contact using $K = 6$ anchors on the hand mesh, including the five fingertips and palm center. Given the contact geometry at frame t, we infer $\mathbf{f}_t \in \mathbb{R}^{3K}$ whose combined effect explains the object wrench $\mathbf{w}_t^o$. These contact forces are mapped to the hand generalized coordinates through $J_{c,t}^\top \mathbf{f}_t$ and combined with the generalized force $\boldsymbol{\tau}_t^h$ required by the hand motion, yielding the active hand actuation $\mathbf{a}_t \in \mathbb{R}^{51}$:

$$\mathbf{a}_t = \boldsymbol{\tau}_t^h + J_{c,t}^\top \mathbf{f}_t, \tag{10}$$

which jointly accounts for the physical effort underlying the hand and object motion, thereby coupling their dynamics. More details of contact-force inference are provided in the supplementary material.

We assume that hand forces during everyday object manipulation typically remain within a characteristic range, and thus active actuation should follow a structured distribution rather than vary arbitrarily. We estimate this distribution from framewise actuation computed on ground-truth training trajectories. As shown in Figure 2, most samples concentrate within a narrow range, while larger values occasionally occur during dynamic interactions such as lifting or releasing objects, yielding a heavy-tailed distribution well captured by a Student-t distribution empirically. Accordingly, we instantiate the coupled-dynamics constraint $\mathcal{C}_a$ in Eq. 2 by fitting a separate Student-t distribution to each DoF and using its negative log-likelihood to penalize unlikely actuation. For coordinate $d \in \{1, \ldots, 51\}$ with fitted parameters (μ_d, σ_d, ν_d), the active-actuation objective is

$$\mathcal{L}_{\text{act}} = \sum_{t,d} \alpha_t \hat{c}_t \left[\log \sigma_d + \frac{\nu_d + 1}{2} \log \left(1 + \frac{(a_{t,d} - \mu_d)^2}{\nu_d \sigma_d^2} \right) \right], \tag{11}$$

where $a_{t,d}$ is the d-th DoF of $\mathbf{a}_t$, $\hat{c}_t \in \{0, 1\}$ indicates ground-truth hand-object contact, and α_t emphasizes dynamically active frames. Since $\mathbf{a}_t$ jointly depends on hand motion, object motion, and contact geometry, this objective provides a learned prior over their coupled dynamics while remaining tolerant to legitimate high-actuation events.

Table 1: Comparison on HOT3D, HO3D-v2, and DexYCB datasets. **Bold** / underline denote the best / second-best reconstruction methods. † denotes our reproduction, and unmarked values are their paper-reported. "–" denotes metrics that are either inapplicable to the method or unavailable due to missing reports and unreleased code.

(a) Hand-object reconstruction on HOT3D dataset.

	Hand					Object						
Method	W↓	WA↓	PA↓	ACCEL↓	RTE↓	ADD@.1↑	ADD@.3↑	ADD-S@.1↑	ADD-S@.3↑	0.1d↑	ACCEL↓	5°5cm↑
*Input estimates (**y**)*												
WiLoR†	11.74	3.70	6.52	16.22	75.16	–	–	–	–	–	–	–
FoundationPose†	–	–	–	–	–	12.8	38.9	25.3	52.8	6.5	70.15	6.7
Reconstruction methods												
HaWoR†	11.73	3.92	8.38	12.90	73.69	–	–	–	–	–	–	–
WHOLE	10.41	3.26	6.67	–	–	–	51.1	–	69.9	–	–	–
DYNAMICHOI	**5.26**	**1.83**	**4.15**	**1.68**	**73.49**	**25.8**	**65.4**	**48.2**	**78.4**	**7.1**	**5.57**	**9.8**

(b) Hand reconstruction on HO3D-v2 dataset.

Method	PA↓	AUC↑	PA-V↓	AUC-V↑	F@5↑	F@15↑	ACCEL↓
*Input estimates (**y**)*							
WiLoR†	7.67	0.847	7.66	0.846	0.647	0.984	4.12
Reconstruction methods							
AMVUR	8.30	0.835	8.20	0.836	0.608	0.965	–
HaMeR	7.70	0.846	7.90	0.841	0.635	0.980	–
Hamba	7.50	0.850	7.70	0.846	0.648	0.982	–
DYNAMICHOI	**7.45**	**0.851**	**7.49**	**0.850**	**0.656**	**0.985**	**2.77**

(c) Hand reconstruction on DexYCB dataset.

Method	MPJPE↓	PA↓	AUC↑	ACCEL↓
*Input estimates (**y**)*				
WiLoR†	11.52	5.33	0.893	6.88
Reconstruction methods				
Deformer	13.64	5.22	–	6.77
HaWoR†	11.09	5.32	0.894	3.88
PAD-Hand	10.56	**4.63**	–	3.34
DYNAMICHOI	**9.51**	4.68	**0.906**	**3.27**

3.3 TRAINING OBJECTIVE AND INFERENCE

We jointly optimize three objectives to train the model in an end-to-end fashion, which is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{d}}\mathcal{L}_{\text{dyn}} + \lambda_{\text{a}}\mathcal{L}_{\text{act}}, \tag{12}$$

where $\mathcal{L}_{\text{diff}}$ supervises trajectory reconstruction, $\mathcal{L}_{\text{dyn}}$ constrains the individual hand and object dynamics, and $\mathcal{L}_{\text{act}}$ further enforces their coupled mechanics through contact. Together, these objectives realize our formulated problem in Eq. 2, which encourages reconstructed trajectories that are both geometrically accurate and physically consistent. The physical quantities are only calculated during training, thus no additional computational cost is introduced at inference.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Datasets and protocols. We conduct experiments on an egocentric video dataset HOT3D (Banerjee et al., 2025), and two allocentric video datasets HO3D-v2 (Hampali et al., 2020) and DexYCB (Chao et al., 2021). For HOT3D, we follow the Aria video split (Ye et al., 2026), using 1,311 training and 147 evaluation clips, each with 150 frames. For HO3D-v2, we use the official split with 55 training sequences and 13 evaluation sequences (11,524 frames). For DexYCB, we follow the official protocol and treat each camera view as a clip, yielding 6,400/320/1,280 clips for training/validation/testing.

Evaluation metrics. For hand reconstruction, HOT3D reports W-MPJPE (W; cm), WA-MPJPE (WA; cm), PA-MPJPE (PA; mm), RTE (%), and ACCEL (mm/frame^2), where W and WA align trajectories using the first two frames and the full sequence, respectively. HO3D-v2 reports PA-MPJPE for both joints and vertices (PA/PA-V; mm), their corresponding AUC/AUC-V (in $[0, 1]$), F@5/F@15, and ACCEL (mm/frame^2). DexYCB reports wrist-relative MPJPE, PA-MPJPE (PA; mm), AUC (in $[0, 1]$), and ACCEL (mm/frame^2). Object pose is evaluated using ADD(-S) AUC (%) at thresholds of 0.1/0.3 m, 0.1d recall (%), ACCEL (mm/frame^2) and $5°$/5 cm success (%). More evaluation and implementation details are provided in the supplementary material.

4.2 EXPERIMENTAL RESULTS ANALYSIS

Model Comparison. We compare DYNAMICHOI with existing reconstruction methods, including WiLoR (Potamias et al., 2025), FoundationPose (Wen et al., 2024), HaWoR (Zhang et al., 2025a),

Table 2: Module ablation. **Bold** / underline denote the best / second-best configurations.

(a) Ablations for hand-object reconstruction on the HOT3D dataset.

	Hand				Object			
Model	W↓	WA↓	PA↓	ACCEL↓	ADD@0.1↑	ADD-S@0.1↑	0.1d↑	5°5cm↑
w/o Geometric	5.83	1.92	4.33	1.71	24.4	46.0	**7.1**	9.5
w/o Grounding	6.65	2.40	4.45	1.74	21.5	43.0	5.3	8.1
w/o Interaction	5.72	1.90	4.24	1.73	21.2	40.0	5.5	8.0
Full model	**5.26**	**1.83**	**4.15**	**1.68**	**25.8**	**48.2**	**7.1**	**9.8**

(b) Ablations for hand reconstruction on the HO3D dataset.

Model	PA↓	PA-V↓	F@5↑
w/o Geometric	7.52	7.51	0.655
w/o Grounding	7.55	7.51	0.654
w/o Interaction	7.59	7.52	0.654
Full model	**7.45**	**7.49**	**0.656**

Table 3: Loss ablation. **Bold** / underline denote the best / second-best configurations.

(a) Ablations for hand-object reconstruction on the HOT3D dataset.

Training objective			Hand				Object			
$\mathcal{L}_{\text{diff}}$	$\mathcal{L}_{\text{dyn}}$	$\mathcal{L}_{\text{act}}$	W↓	WA↓	PA↓	ACCEL↓	ADD@0.1↑	ADD-S@0.1↑	0.1d↑	5°5cm↑
✓			8.27	2.82	5.16	6.10	17.8	33.6	4.7	7.4
✓		✓	7.77	2.31	5.06	5.01	20.0	37.9	5.3	8.5
✓	✓		6.12	1.91	4.22	2.78	24.9	47.4	6.7	**9.8**
✓	✓	✓	**5.26**	**1.83**	**4.15**	**1.68**	**25.8**	**48.2**	**7.1**	**9.8**

(b) Ablations for hand on HO3D.

Training objective			Hand		
$\mathcal{L}_{\text{diff}}$	$\mathcal{L}_{\text{dyn}}$	$\mathcal{L}_{\text{act}}$	PA↓	PA-V↓	F@5↑
✓			7.55	7.59	0.650
✓		✓	7.48	7.50	0.651
✓	✓		7.46	7.50	0.651
✓	✓	✓	**7.45**	**7.49**	**0.656**

WHOLE (Ye et al., 2026), AMVUR (Jiang et al., 2023), HaMeR (Pavlakos et al., 2024), Hamba (Dong et al., 2024), Deformer (Fu et al., 2023), and PAD-Hand (Ismayilzada et al., 2026). Table 1 shows that DYNAMICHOI achieves the best performance on most metrics while consistently improving the input hand and object estimates. The gains are particularly large on the egocentric HOT3D dataset, indicating that our model remains effective under moving viewpoints. Importantly, modeling object dynamics and coupling them with hand dynamics improves not only object reconstruction but also hand motion. On the DexYCB dataset, DYNAMICHOI outperforms PAD-Hand, which explicitly models hand dynamics, suggesting that object dynamics provide complementary constraints on hand reconstruction through physical interaction. This benefit is further reflected in the reduction in acceleration error (ACCEL) across all datasets for both hand and object, indicating that our proposed physical constraints improve not only trajectory accuracy but also the dynamical consistency of hand-object motion.

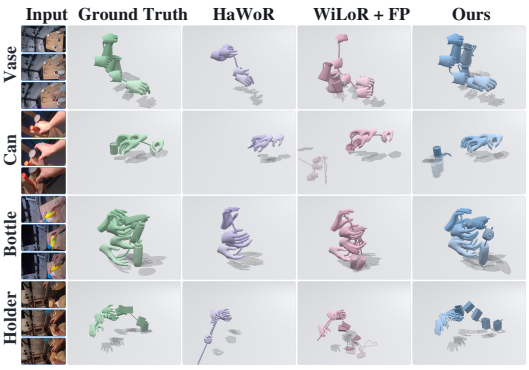


Figure 3: Visualization for reconstruction.

Module Ablation. As shown in Table 2, for hand reconstruction, removing visual grounding causes the largest degradation on egocentric HOT3D, indicating that geometry-aligned visual evidence is particularly important for correcting trajectories in a moving camera. In contrast, the interaction module brings a minor effect, because dense hand-object co-motion can already be captured by temporal refinement. On fixed-view HO3D, removing interaction causes the largest degradation, as persistent occlusion and partial visibility make explicit hand-object motion coupling more important, while geometric encoding is less critical under a stable camera. For object reconstruction on HOT3D, interaction is the most influential module, showing that hand motion provides strong constraints for recovering object pose. Geometric encoding has a smaller effect, because the rigid object structure is already well preserved, leaving the other two modules as main cues for resolving its pose changes.

Loss Ablation. As shown in Table 3, adding $\mathcal{L}_{\text{dyn}}$ (2nd vs. 4th rows) substantially improves both hand and object reconstruction on HOT3D, confirming the benefit of explicitly constraining their dynamics. Comparing the last two rows further shows that $\mathcal{L}_{\text{act}}$ provides a smaller but consistent additional gain. We attribute this limited improvement partly to the large variability of video content and to errors in the forces estimated by the Euler-Lagrange and Newton-Euler equations, which may propagate into the inferred actuation distribution. More accurate force estimation and actuation modeling remain promising directions for future work. On fixed-view HO3D, both physical losses yield modest gains, suggesting that dynamics supervision is particularly beneficial for egocentric reconstruction, where camera motion and trajectory ambiguity make purely visual constraints insufficient.

Actuation Distribution. As shown in Table 4, we compare actuation priors using the average NLL per valid training coordinate and PA/ADD@0.1 reconstruction metrics. We can observe that an inappropriate prior can degrade reconstruction: the Gaussian prior underperforms removing $\mathcal{L}_{\text{act}}$, supporting the heavy-tailed nature of the inferred actuation. Fitting ν_d separately for each DoF consistently outperforms fixed-ν Student-t priors and achieves the best overall performance. Notably, better likelihood fit does not necessarily imply better reconstruction: although $\nu = 2$ yields lower NLL than $\nu = 10$, it performs worse on object reconstruction, suggesting that likelihood fit alone does not determine the effectiveness of the prior as a reconstruction constraint.

Table 4: Ablation of different distribution choices for modeling active hand actuation on HOT3D dataset.

Distribution	NLL ↓	PA ↓	ADD@.1 ↑
w/o $\mathcal{L}_{\text{act}}$	–	4.22	24.9
Gaussian	-0.39	4.54	24.2
Laplace	-0.11	4.17	24.9
Cauchy (ν=1)	-1.45	4.20	24.7
Student-t distributions			
Fixed ν=10	-0.70	4.17	25.6
Fixed ν=5	-0.96	4.16	25.2
Fixed ν=2	-1.15	4.16	25.3
Fitted ν_d	**-1.59**	**4.15**	**25.8**

4.3 Downstream Applications

Beyond reconstruction, we evaluate whether physics-aware hand trajectories from DynamicHOI benefit downstream embodied tasks. For dexterous manipulation, we follow ViViDex (Chen et al., 2025) and retarget trajectories reconstructed on the DexYCB dataset for robot policy learning. For egocentric world modeling, we follow EgoHOI (Li et al., 2026) and use trajectories reconstructed on the HOT3D dataset as hand-action conditions. In both tasks, only the input hand trajectories vary across reconstruction methods, while all other components remain fixed. We use reconstructed trajectories from models trained on the corresponding dataset without task-specific retraining.



(a) Dexterous manipulation. (b) Egocentric video generation.

Figure 4: Downstream applications with different hand-trajectory sources.

Dexterous manipulation. We evaluate eight held-out DexYCB relocate tasks, training one PPO policy (Schulman et al., 2017) for each object and trajectory source and evaluating 100 episodes per task. As shown in Figure 4a, downstream success provides a task-level measure of reconstruction quality. Methods with comparable reconstruction accuracy, such as WiLoR and HaWoR (see Table 1c), also achieve similar success rates. In contrast, DynamicHOI reduces MPJPE by 17.4% over WiLoR, and this advantage is amplified in embodied learning, increasing the success rate from 25.0% to 74.9%. This suggests that improvements in physically coherent hand reconstruction can translate into substantially larger gains when the trajectories are used for robotic learning.

Egocentric world modeling. We evaluate both action fidelity and generated hand appearance. mLPIPS measures perceptual similarity within the GT hand region, while MPJPE measures the kinematic accuracy of the generated hand. Figure 4b shows that DynamicHOI performs best among reconstructed action sources on all metrics. Since EgoHOI conditions generation on framewise MANO renderings aligned with image pixels, more accurate trajectories provide better spatial guidance for where and how the hand should appear. Thus, improved reconstruction not only yields more faithful generated motion but also translates into better visual quality in the hand region.

5 Conclusion

We presented DynamicHOI, a physics-aware HOI reconstruction framework that combines geometry-grounded trajectory refinement with coupled hand-object dynamics. We model articulated-hand and rigid-object dynamics and couple them through contact, with a learned probabilistic prior over active hand actuation. Experiments on egocentric and allocentric datasets demonstrate improved reconstruction, while downstream gains in dexterous manipulation and egocentric world modeling show the broader value of physically coherent HOI trajectories. Future work will explore more accurate force estimation, richer actuation priors, and unified object representations that generalize beyond known template meshes to unseen objects.

AI USE STATEMENT

In this work, we used generative AI tools only to assist with manuscript writing and figure illustration. Specifically, large language models were used for translation and language polishing to improve the clarity and presentation of the manuscript. We also used ChatGPT Images 2.5 (OpenAI, 2026b) to assist in creating the illustrative elements in Figure 2, except for the distribution plot in the bottom-right corner, to provide a clearer visual explanation of the force decomposition. Generative AI tools were not used to formulate the research ideas or methodology, conduct experiments, analyze data, generate experimental results, or draw scientific conclusions. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final content of this work.

REFERENCES

Ahmed Tawfik Aboukhadra, Marcel Rogge, Nadia Robertini, Abdalla Arafa, Jameel Malik, Ahmed Elhayek, and Didier Stricker. GHOST: Fast category-agnostic hand-object interaction reconstruction from RGB videos using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. ForeHOI: Feed-forward 3D object reconstruction from daily hand-object interaction videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8868–8879, 2026.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2022.

Zerui Chen, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. ViViDex: Learning vision-based dexterous manipulation from human videos. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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

Roy Featherstone. *Rigid Body Dynamics Algorithms*. Springer, 2008.

Qi Feng, Hubert PH Shum, and Shigeo Morishima. Resolving hand-object occlusion for mixed reality with joint deep learning and model optimization. *Computer Animation and Virtual Worlds*, 31(4-5):e1956, 2020.

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, 2026.

Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23543–23554. IEEE, 2023.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193–3203. IEEE, 2020.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11807–11816, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia 2022 Conference Papers*. Association for Computing Machinery, 2022. doi: 10.1145/3550469.3555421.

Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *ACM SIGGRAPH Asia Conference Papers*, 2022.

Elkhan Ismayilzada, Yufei Zhang, and Zijun Cui. PAD-Hand: Physics-aware diffusion for hand motion recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 758–768. IEEE, 2023.

Wisama Khalil. Dynamic modeling of robots using newton-euler formulation. In *Informatics in Control, Automation and Robotics: Revised and Selected Papers from the International Conference on Informatics in Control, Automation and Robotics 2010*, pp. 3–20. Springer, 2011.

Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

Dayou Li, Lulin Liu, Bangya Liu, Shijie Zhou, Jiu Feng, Ziqi Lu, Minghui Zheng, Chenyu You, and Zhiwen Fan. Egocentric world model for photorealistic hand-object interaction synthesis, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

John YS Luh, Michael W Walker, and Richard PC Paul. On-line computational scheme for mechanical manipulators. *Journal of Dynamic Systems, Measurement, and Control*, 102(2):69–76, 06 1980. ISSN 0022-0434. doi: 10.1115/1.3149599. URL https://doi.org/10.1115/1.3149599.

Haowen Luo, Yunze Liu, and Li Yi. Physics-Aware Hand-Object Interaction Denoising . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2341–2350, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00227. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00227.

Michael Lutter. A differentiable newton–euler algorithm for real-world robotics. In *Inductive Biases in Machine Learning for Robotics and Control*, pp. 9–34. Springer, 2023.

Takayuki Nakajima, Yuki Asami, Yui Endo, Mitsunori Tada, and Naomichi Ogihara. Prediction of anatomically and biomechanically feasible precision grip posture of the human hand based on minimization of muscle effort. *Scientific Reports*, 12(1):13247, 2022.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026a. Model family released July 9, 2026. Accessed: 2026-08-17.

OpenAI. Introducing chatgpt images 2.5. https://openai.com/index/introducing-chatgpt-images-2-5/, September 2026b. Accessed: 2026-09-17.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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

Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision (ECCV)*, 2022.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Kevin Tanghe, Maarten Afschrift, Ilse Jonkers, Friedl De Groote, Joris De Schutter, and Erwin Aertbeliën. A probabilistic method to estimate gait kinetics in the absence of ground reaction force measurements. *Journal of biomechanics*, 96:109327, 2019.

Fulvio Tonon. Explicit exact formulas for the 3-D tetrahedron inertia tensor in terms of its vertex coordinates. *Journal of Mathematics and Statistics*, 1(1):8–11, 2005.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. Magichoi: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. ArtHOI: Taming foundation models for monocular 4D reconstruction of hand-articulated-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yufei Ye, Jiaman Li, Ryan Rong, and C. Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *CVPR Findings*, 2026.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-HaMR: Recovering 4D interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ZIyAHaLlsn.

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)*, 2025a.

Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2305–2317. IEEE, 2024.

Yufei Zhang, Zijun Cui, Jeffrey O Kephart, and Qiang Ji. Diffusion-based 3d hand motion recovery with intuitive physics. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7306–7317. IEEE, 2025b.

Kanglei Zhou, Chen Chen, Yue Ma, Zhiying Leng, Hubert PH Shum, Frederick WB Li, and Xiaohui Liang. A mixed reality training system for hand-object interaction in simulated microgravity environments. In *2023 IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 167–176. IEEE, 2023.

Appendix: Supplementary Material

This supplementary material provides the theoretical details of the hand and object dynamics and of the contact-force inference (Appendix A), implementation and evaluation details (Appendix B), additional visualizations (Appendix C), and a discussion of limitations (Appendix D). Figures, tables, and equations in this supplementary material are numbered with the prefix "S"; numbers without this prefix refer to the main paper.

A Theoretical Details

The dynamics of Sect. 3.2.2 and Sect. 3.2.3 are computed with temporal finite differences at $\Delta t = 1/30\,\mathrm{s}$, on both the predicted and the ground-truth trajectories; the ground-truth quantities are treated as constants. During training, Eq. 9 and Eq. 11 are evaluated on 16-frame training windows (Appendix B.1).

A.1 Articulated Hand Inverse Dynamics

Generalized coordinates and kinematics. The generalized coordinates are $\mathbf{q}^h = [\mathbf{t}^h; \boldsymbol{\phi}_0; \ldots; \boldsymbol{\phi}_{15}] \in \mathbb{R}^{51}$, where $\boldsymbol{\phi}_j = (\phi_{j,x}, \phi_{j,y}, \phi_{j,z})$ are the Euler angles of the global orientation ($j = 0$) and of the 15 finger joints. Following PhysPT (Zhang et al., 2024), the Euler-ZXY convention denotes extrinsic rotations about the z, x, and y axes, in this order, i.e. the rotation matrix $\mathbf{R}_y(\phi_{j,y})\mathbf{R}_x(\phi_{j,x})\mathbf{R}_z(\phi_{j,z})$. The world rotation of joint j with parent $\mathrm{pa}(j)$ is therefore

$$\mathbf{W}_j = \mathbf{W}_{\mathrm{pa}(j)}\,\mathbf{R}_y(\phi_{j,y})\,\mathbf{R}_x(\phi_{j,x})\,\mathbf{R}_z(\phi_{j,z}), \qquad \mathbf{W}_{\mathrm{pa}(0)} = \mathrm{Id}, \tag{S1}$$

and the three rotation axes of joint j in world coordinates are $\mathbf{e}_{j,y} = \mathbf{W}_{\mathrm{pa}(j)}\mathbf{e}_y$, $\mathbf{e}_{j,x} = \mathbf{W}_{\mathrm{pa}(j)}\mathbf{R}_y\mathbf{e}_x$, and $\mathbf{e}_{j,z} = \mathbf{W}_{\mathrm{pa}(j)}\mathbf{R}_y\mathbf{R}_x\mathbf{e}_z$, where $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ are the coordinate axes. With the rest-pose joint positions $\bar{\mathbf{u}}_j$, the joint positions are $\mathbf{u}_j = \mathbf{u}_{\mathrm{pa}(j)} + \mathbf{W}_{\mathrm{pa}(j)}(\bar{\mathbf{u}}_j - \bar{\mathbf{u}}_{\mathrm{pa}(j)})$, and the centre of mass (CoM) of part k is $\boldsymbol{\xi}_k = \mathbf{u}_k + \mathbf{W}_k\bar{\boldsymbol{\xi}}_k$.

Rigid parts and mass properties. Each MANO vertex is assigned to the joint with the largest linear-blend-skinning weight, and each face to the majority part of its vertices, which yields the 16 rigid parts of Fig. S1(a). For every part with triangle set $\mathcal{F}_k$ and vertex centroid $\mathbf{b}_k$, we approximate its volume by the tetrahedra spanned by its triangles and $\mathbf{b}_k$,

$$V_k \approx \sum_{f \in \mathcal{F}_k} \tfrac{1}{6}\big|\det[\mathbf{v}_{f,1} - \mathbf{b}_k,\ \mathbf{v}_{f,2} - \mathbf{b}_k,\ \mathbf{v}_{f,3} - \mathbf{b}_k]\big|, \tag{S2}$$

and accumulate the closed-form second moments of these tetrahedra (Tonon, 2005) into the body inertia $\bar{\mathbf{I}}_k$. The part masses m_k and inertias follow from a constant tissue density of $1000\,\mathrm{kg/m^3}$. These constants are computed once on the canonical MANO hand (mean shape, zero pose).

Terms of the manipulator equation. Let $\mathcal{A}(k)$ be the set of rotational DoFs of the joints on the kinematic path from part k to the root. For a rotational DoF i with world axis $\mathbf{e}_i$ located at joint position $\mathbf{u}_i$, the Jacobian columns of part k are

$$J^p_{k,i} = \mathbb{1}[i \in \mathcal{A}(k)]\ \mathbf{e}_i \times (\boldsymbol{\xi}_k - \mathbf{u}_i), \qquad J^r_{k,i} = \mathbb{1}[i \in \mathcal{A}(k)]\ \mathbf{e}_i, \tag{S3}$$

and the three translational DoFs contribute the columns Id_3 to J^p_k and $\mathbf{0}$ to J^r_k. With the world-frame part inertia $\mathbf{I}^w_k = \mathbf{W}_k\bar{\mathbf{I}}_k\mathbf{W}_k^\top$ and the part angular velocity $\boldsymbol{\omega}_k = J^r_k\dot{\mathbf{q}}^h$, the terms of Eq. 7 are

$$M^h(\mathbf{q}^h) = \textstyle\sum_k \big(m_k\, J_k^{p\top} J_k^p + J_k^{r\top}\mathbf{I}^w_k J^r_k\big), \tag{S4}$$

$$C^h(\mathbf{q}^h, \dot{\mathbf{q}}^h)\dot{\mathbf{q}}^h = \textstyle\sum_k \big(m_k\, J_k^{p\top}\dot{J}^p_k\dot{\mathbf{q}}^h + J_k^{r\top}\mathbf{I}^w_k\dot{J}^r_k\dot{\mathbf{q}}^h + J_k^{r\top}(\boldsymbol{\omega}_k \times \mathbf{I}^w_k\boldsymbol{\omega}_k)\big), \tag{S5}$$

$$\mathbf{g}^h_{\mathrm{grav}}(\mathbf{q}^h) = -\textstyle\sum_k m_k\, J_k^{p\top}\mathbf{g}, \tag{S6}$$

which is the standard Lagrangian form of rigid multibody dynamics (Featherstone, 2008; Khalil, 2011). Velocities and accelerations use central differences between adjacent frames, $\dot{\mathbf{q}}_t = (\mathbf{q}_{t+1} - \mathbf{q}_{t-1})/(2\Delta t)$ and $\ddot{\mathbf{q}}_t = (\mathbf{q}_{t+1} - 2\mathbf{q}_t + \mathbf{q}_{t-1})/\Delta t^2$, and the Jacobian derivative is approximated by $\dot{J}_t \approx (J_{t+1} - J_{t-1})/(2\Delta t)$.

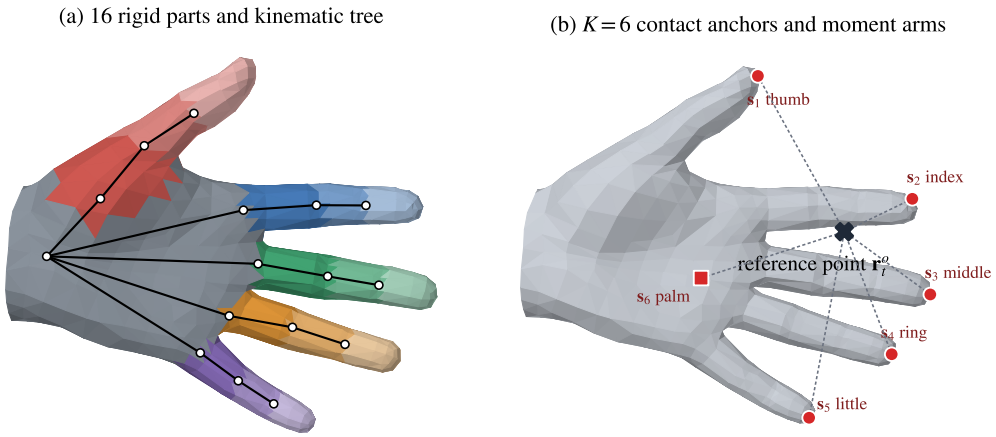


Figure S1: **Hand model used by the dynamics.** (a) The 16 rigid parts obtained from the MANO skinning weights (one colour family per finger, lighter = more distal) and the kinematic tree of the inverse dynamics. (b) The $K = 6$ contact anchors (the five fingertips and the palm centre) and their moment arms to the reference point $\mathbf{r}_t^o$ of the object wrench, which form the grasp matrix of Eq. S8. The object position in (b) is illustrative.

Why generalized forces. Eq. 7 is linear in $\ddot{\mathbf{q}}^h$, but the mass matrix re-weights each DoF by the inertia it moves: an error of the global rotation or of the wrist, which displaces the whole hand, is penalized far more than the same angular error of a distal phalanx. The Coriolis and gravity terms further couple the DoFs along each kinematic chain.

A.2 Rigid-Object Newton–Euler Dynamics

Differentiation on SO(3). Rotations are differentiated on the manifold rather than component-wise. With $\mathrm{Log} : SO(3) \to \mathbb{R}^3$ the rotation-vector map,

$$\boldsymbol{\omega}_t^o = \frac{\mathrm{Log}\big(\mathbf{R}_{t+1}^o \mathbf{R}_{t-1}^{o\top}\big)}{2\Delta t}, \qquad \dot{\boldsymbol{\omega}}_t^o = \frac{\boldsymbol{\omega}_{t+1}^o - \boldsymbol{\omega}_{t-1}^o}{2\Delta t}, \qquad \mathbf{a}_t^o = \frac{\mathbf{t}_{t+1}^o - 2\mathbf{t}_t^o + \mathbf{t}_{t-1}^o}{\Delta t^2}. \tag{S7}$$

Because $\mathbf{R}_{t+1}\mathbf{R}_{t-1}^\top$ is a left (spatial) increment, $\boldsymbol{\omega}_t^o$ is expressed in the same frame as the pose, consistent with the world-frame inertia $\mathbf{I}_t^o = \mathbf{R}_t^o \bar{\mathbf{I}}^o \mathbf{R}_t^{o\top}$ of Eq. 8. (Here $\mathbf{a}_t^o$ denotes the object's linear acceleration, not the hand actuation $\mathbf{a}_t$ of Eq. 10.)

Inertia. $\bar{\mathbf{I}}^o$ is the inertia tensor of the canonical mesh $\mathcal{M}^o$, assuming a uniform mass distribution with total mass m^o (Sect. 3.2.2).

Role of gravity. For a given object, $m^o\mathbf{g}$ does not depend on the trajectory, so it cancels in the object term of $\mathcal{L}_{\mathrm{dyn}}$ (Eq. 9). This term therefore supervises the inertial wrench $[m^o\mathbf{a}_t^o;\ \mathbf{I}_t^o\dot{\boldsymbol{\omega}}_t^o + \boldsymbol{\omega}_t^o \times \mathbf{I}_t^o\boldsymbol{\omega}_t^o]$, in which translational errors are weighted by the mass and rotational errors by the inertia of the object. Gravity does not cancel in the actuation of Eq. 10: during quasi-static holding it is the dominant part of the wrench that the grasp must supply.

A.3 Contact-Force Inference and Active Actuation

Grasp matrix. Let $\mathbf{s}_{t,l}$ $(l = 1, \ldots, K)$ be the anchor positions on the reconstructed hand (Fig. S1(b)) and $\mathbf{r}_t^o$ the point about which the object wrench of Eq. 8 is expressed. A force $\mathbf{f}_{t,l}$ applied at $\mathbf{s}_{t,l}$ produces the wrench $[\mathbf{f}_{t,l};\ (\mathbf{s}_{t,l} - \mathbf{r}_t^o) \times \mathbf{f}_{t,l}]$ about $\mathbf{r}_t^o$, hence

$$\mathbf{w}_t^o \approx \boldsymbol{\Gamma}_t \mathbf{f}_t, \qquad \boldsymbol{\Gamma}_t = \begin{bmatrix} \mathrm{Id}_3 & \cdots & \mathrm{Id}_3 \\ [\mathbf{s}_{t,1} - \mathbf{r}_t^o]_\times & \cdots & [\mathbf{s}_{t,K} - \mathbf{r}_t^o]_\times \end{bmatrix} \in \mathbb{R}^{6\times 3K}, \tag{S8}$$

where $[\cdot]_\times$ is the skew-symmetric cross-product matrix.

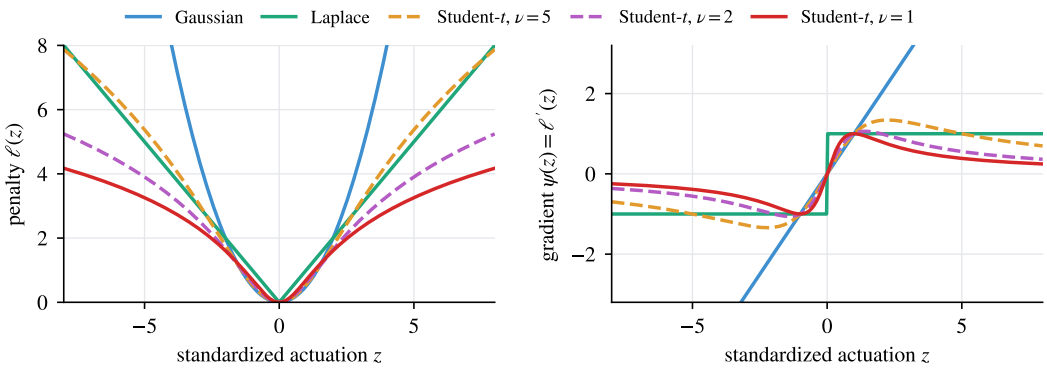


Figure S2: **Penalty and gradient of the candidate actuation priors** in standardized units. The Student-t gradient is bounded and redescending, so heavy-tailed but physically legitimate actuation spikes are tolerated while moderate deviations are corrected.

Minimum-norm inverse grasp mapping. With $3K = 18$ unknowns and 6 equations the inverse is under-determined. Among the anchor-force distributions that reproduce the object wrench, we select the one of minimum magnitude, regularized by a small ridge term:

$$\mathbf{f}_t = \arg\min_{\mathbf{f}} \ \|\mathbf{\Gamma}_t \mathbf{f} - \mathbf{w}_t^o\|_2^2 + \gamma \|\mathbf{f}\|_2^2 = \mathbf{\Gamma}_t^\top \left(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top + \gamma \, \mathrm{Id}_6\right)^{-1} \mathbf{w}_t^o, \qquad \gamma = 10^{-4}. \tag{S9}$$

On frames without ground-truth hand–object contact ($\hat{c}_t = 0$) the anchor forces are set to zero, so that a hand that does not touch the object cannot carry force; these frames do not contribute to $\mathcal{L}_{\text{act}}$ (Eq. 11). The closed form only inverts a 6×6 matrix and is batched over frames. Its solution is the least-effort (minimum-ℓ_2) grasp that explains the object motion.

Contact Jacobian and actuation. Each anchor l is attached to its nearest MANO part $k(l)$, and a force $\mathbf{f}_{t,l}$ at $\mathbf{s}_{t,l}$ maps to the generalized coordinates along the kinematic path of Eq. S3:

$$\left(J_{c,t}^\top \mathbf{f}_t\right)_{\text{trans}} = \textstyle\sum_l \mathbf{f}_{t,l}, \qquad \left(J_{c,t}^\top \mathbf{f}_t\right)_i = \textstyle\sum_l \mathbb{1}[i \in \mathcal{A}(k(l))] \ \mathbf{e}_i \cdot \left((\mathbf{s}_{t,l} - \mathbf{u}_i) \times \mathbf{f}_{t,l}\right). \tag{S10}$$

By Newton's third law the hand receives $-\mathbf{f}_t$, whose generalized load is $J_{c,t}^\top(-\mathbf{f}_t)$. The actuation that the hand must produce itself is the part of the motion-implied generalized force that the contact reaction does not explain, $\mathbf{a}_t = \boldsymbol{\tau}_t^h - J_{c,t}^\top(-\mathbf{f}_t) = \boldsymbol{\tau}_t^h + J_{c,t}^\top \mathbf{f}_t$, which is Eq. 10. Through $\mathbf{f}_t$, the actuation depends on the object motion (via $\mathbf{w}_t^o$) and on the hand–object geometry (via $\mathbf{\Gamma}_t$).

A.4 The Student-t Actuation Prior

For one DoF with standardized actuation $z = (a_{t,d} - \mu_d)/\sigma_d$, the Student-$t$ penalty of Eq. 11 (without the constant $\log \sigma_d$) and its derivative, which is the influence function of the corresponding M-estimator (Hampel et al., 1986), are

$$\ell_\nu(z) = \tfrac{\nu+1}{2} \log\!\left(1 + \tfrac{z^2}{\nu}\right), \qquad \psi_\nu(z) = \ell_\nu'(z) = \frac{(\nu+1)\, z}{\nu + z^2}. \tag{S11}$$

For $|z| \ll \sqrt{\nu}$, $\ell_\nu(z) \approx \frac{\nu+1}{2\nu} z^2$ and $\psi_\nu(z) \approx \frac{\nu+1}{\nu} z$, i.e. the prior acts as a quadratic penalty; for $|z| \gg \sqrt{\nu}$ its gradient decays as $(\nu+1)/z$ (Fig. S2). Moderate implausible actuation is therefore pushed back, while the rare but legitimate high-actuation events of lifting, placing, or impact, which populate the tails, receive a bounded gradient and do not dominate training (Lange et al., 1989). The Gaussian prior has an unbounded gradient and over-penalizes exactly these events, which is consistent with it performing worse than removing $\mathcal{L}_{\text{act}}$ in Table 4; the Laplace prior has a constant gradient magnitude. Since ψ_ν is maximal at $|z| = \sqrt{\nu}$, the degrees of freedom set the scale beyond which the prior progressively stops responding to a sample, which motivates fitting them per DoF.

B Implementation Details

B.1 Datasets and Protocols

Table S1 summarizes the three datasets. During training, sequences are cut into windows of 16 frames with a stride of 4 frames. At test time, each sequence is processed in 16-frame windows whose

Table S1: **Datasets.** The training and evaluation splits are described in Sect. 4.

	HOT3D (Banerjee et al., 2025)	HO3D-v2 (Hampali et al., 2020)	DexYCB (Chao et al., 2021)
Viewpoint	egocentric, moving (Aria, fisheye)	allocentric, static	allocentric, static (8 views)
Objects	33	10 YCB objects	20 YCB objects
Frame rate	30 fps	30 fps	30 fps

Table S2: **Network and optimization hyperparameters.** Training times are wall-clock on one NVIDIA H100.

	HOT3D	HO3D-v2	DexYCB
feature width / attention heads	512 / 8	512 / 8	768 / 8
temporal self-attention layers	6	6	12
geometry encoders	two-layer MLPs		
visual encoder	DINOv2 ViT-B/14, $16 \times 16 \times 768$ patch tokens, frozen		
trainable parameters	33.1 M	33.1 M	59.6 M
optimizer	AdamW (Loshchilov & Hutter, 2019), weight decay 0.01, 5% warm-up		
learning rate / schedule	2×10^{-4} / constant	2.8×10^{-4} / constant	1.4×10^{-4} / cosine
batch size (windows)	48	48	384
training epochs	40	5	6
weight averaging	exponential moving average, decay 0.999 (evaluated model)		
diffusion	ResShift, $N = 4$, $\kappa = \pi/12$		
training time	$\approx$39 h	$\approx$5 h	$\approx$2 h

predictions are assembled into full-length trajectories before evaluation. The ground-truth contact labels $\hat{c}_t$ of Eq. 11 are derived from the ground-truth meshes, from the minimum distance between the hand vertices and the object surface, with a hysteresis (contact starts below 1.5 cm and ends above 2.5 cm) followed by a 3-frame median filter.

B.2 Visual Features and Anchor Projection

The frozen visual encoder is DINOv2 ViT-B/14 (Oquab et al., 2024), applied to every frame resized to 224×224; the resulting 16×16 grid of 768-d patch tokens is computed once and cached. The projection Π_t of Eq. 5 uses the pinhole intrinsics on HO3D-v2 and DexYCB and the Aria FISHEYE624 camera model on HOT3D. The hand anchors are the 21 MANO joints and the object anchors are 32 keypoints on the canonical object surface, and the patch features are sampled bilinearly at the projected anchor locations.

B.3 Diffusion Schedule and Trajectory Discrepancy

Diffusion schedule. The forward process of Eq. 4 is the residual-shift diffusion of ResShift (Yue et al., 2023), which progressively shifts the ground-truth trajectory toward the front-end estimate. We use $N = 4$ steps, $\kappa = \pi/12$, and the geometric schedule of ResShift between $\eta_1^2 = \min(10^{-3}, (0.04/\kappa)^2)$ and $\eta_N^2 = 0.999$ with exponent 0.1:

step n	1	2	3	4
η_n (the noise standard deviation is $\eta_n \kappa$)	0.032	0.698	0.871	0.999
η_n^2 (fraction of the residual $\mathbf{y} - \hat{\mathbf{x}}$ retained)	0.001	0.487	0.759	0.999

Trajectory discrepancy. The discrepancy d_x of Eq. 6 is a weighted sum of ℓ_2 distances between the predicted and ground-truth hand rotations (6D representation, with an additional weight on the global orientation), hand translations, and object rotations (6D) and translations.

B.4 Network and Optimization Hyperparameters

Table S2 lists the network and optimization hyperparameters. The physical quantities of Appendix A are computed only in the training losses, so they add no cost at inference.

B.5 Evaluation Metrics

Table S3 summarizes the metric definitions. The HOT3D hand metrics follow the definitions used by HaWoR. The HO3D-v2 metrics are computed with the official evaluation code of the dataset against

Table S3: **Evaluation metrics.** Hand metrics use the 21 MANO joints (or the 778 vertices where indicated).

Metric	Definition
W-MPJPE (cm)	World-frame joint error after a similarity alignment (Umeyama, 1991) estimated on the first two frames and applied to the whole sequence.
WA-MPJPE (cm)	World-frame joint error after one similarity alignment estimated on the whole sequence.
PA-MPJPE / PA-MPVPE (mm)	Joint / vertex error after per-frame Procrustes (similarity) alignment.
MPJPE (mm)	Wrist-relative joint error (DexYCB).
AUC / AUC-V	Area under the PCK curve of the PA-aligned joints / vertices for thresholds in $[0, 50]$ mm, in $[0, 1]$.
F@5 / F@15	F-score between the PA-aligned predicted and ground-truth vertices at 5 / 15 mm.
ACCEL (mm/frame^2)	Mean norm of the difference between predicted and ground-truth accelerations (second temporal differences) of the hand joints or of the object model points.
ADD@δ / ADD-S@δ (%)	Area under the accuracy–threshold curve of the ADD error (mean distance of corresponding model points) / ADD-S error (mean closest-point distance) for thresholds in $[0, \delta]$, $\delta \in \{0.1, 0.3\}$ m, $\times 100$.
0.1d (%)	Fraction of frames whose ADD is below 10% of the object diameter.
5°5cm (%)	Fraction of frames with rotation error below 5° and translation error below 5 cm.

its released evaluation annotations, and ACCEL is computed on root-relative joints. Baselines marked † in Table 1 were re-run with their public code and scored with the same evaluation scripts as our method.

B.6 Downstream Applications

Dexterous manipulation. We use the state-based relocate task of ViViDex (Chen et al., 2025). The eight tasks correspond to eight DexYCB sequences, one per object, all of which belong to the DexYCB test split and are therefore unseen by the reconstruction model. For each task and hand-trajectory source, the reconstructed hand trajectory is retargeted to the robot hand with the retargeting of ViViDex, while the object reference trajectory and all other components are kept fixed. Each policy is evaluated on 100 episodes with randomized initial object pose, and an episode is successful if the object ends within 3 cm of the goal while the hand is in contact with it.

Egocentric world modeling. We follow EgoHOI (Li et al., 2026) on the eight HOT3D sequences showcased by EgoHOI and replace only the hand-action condition (per-frame MANO renderings aligned to the image) by the trajectory of each source; all generation settings are unchanged. None of these eight sequences is part of the training data of the reconstruction model used in this experiment. MPJPE is measured by running the same frozen WiLoR (Potamias et al., 2025) on the generated frames and comparing the estimated joints with the HOT3D ground-truth MANO joints. mLPIPS is the LPIPS (AlexNet) distance between generated and ground-truth frames, restricted to the ground-truth hand region.

C Additional Visualizations

Motion-implied physical quantities. Fig. S3 plots the mean hand-joint acceleration and the object force and torque of Eq. 8, computed from the input, refined, and ground-truth trajectories of three HOT3D clips. The frame-wise front ends produce frequent spikes in all three quantities, up to about three orders of magnitude above the ground truth. After refinement, the hand acceleration closely follows the ground truth, and the object force and torque remain much closer to the ground truth than those of the input. Deviations remain: isolated dips of the object force (e.g. Vase, $t \approx 1.1$ s), torque overshoots during fast motion (e.g. Mug, $t \approx 2.1$ s), and elevated torque while the object is static (Vase, $t < 1$ s).

Temporal comparison on HOT3D. Fig. S4 shows the first 2 s of a HOT3D clip, rendered onto the input frames after whole-trajectory alignment. HaWoR reconstructs hands only. After alignment, the frame-wise FoundationPose object of the input lies outside the crop and appears only at the image border. The object refined by DynamicHOI is offset from the ground truth while it rests on the table ($t = 0.1$ s), then follows it as the hand approaches and grasps it; residual orientation errors remain during the fast rotation ($t = 1.4$ s).

Downstream applications. Fig. S5 shows one example per application. In EgoHOI, the WiLoR action condition misplaces the hand and the generated vase follows it, whereas the generation

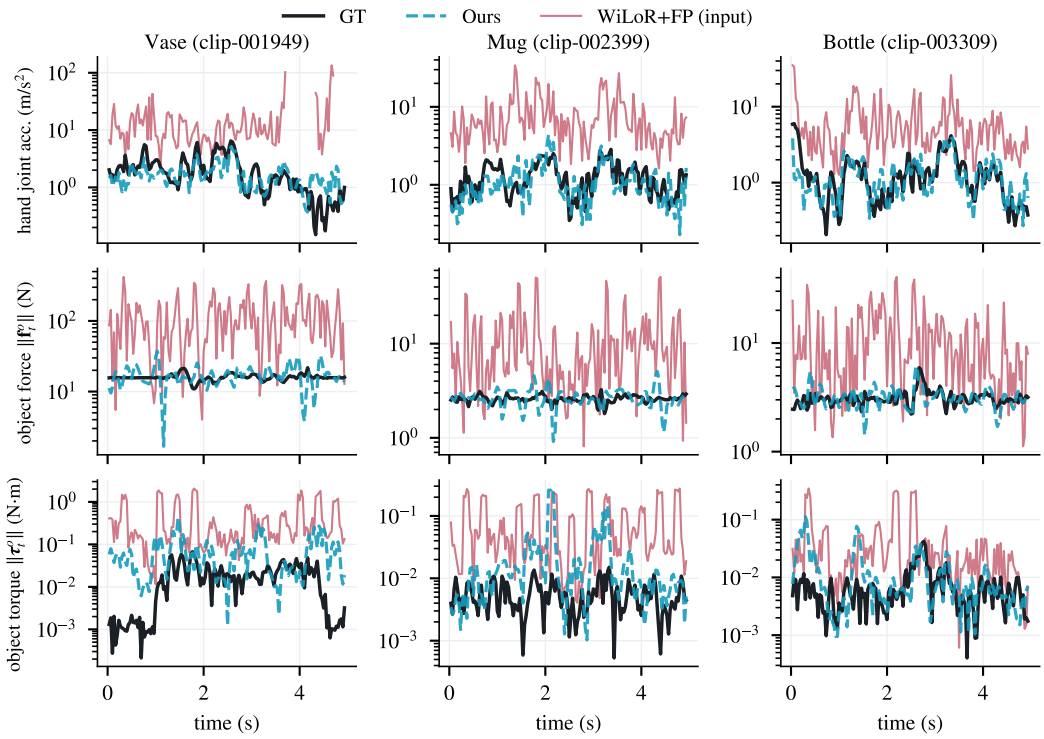


Figure S3: **Motion-implied physical quantities** on three HOT3D clips (log scale): mean hand-joint acceleration (top), object force $\|m^o(\mathbf{a}_t^o - \mathbf{g})\|$ (middle), and object torque $\|\boldsymbol{\tau}_t^o\|$ (bottom), for the ground truth, the front-end input (WiLoR + FoundationPose), and DYNAMICHOI. When the object is static, the force equals the object's weight.

conditioned on our trajectory is much closer to the grasp of the ground-truth video. In ViViDex, the WiLoR hand trajectory used as retargeting input grasps the can off-centre and the resulting policy drops it, while the policy obtained from our trajectory places the can at the goal, as the policy obtained from ground-truth hands does.

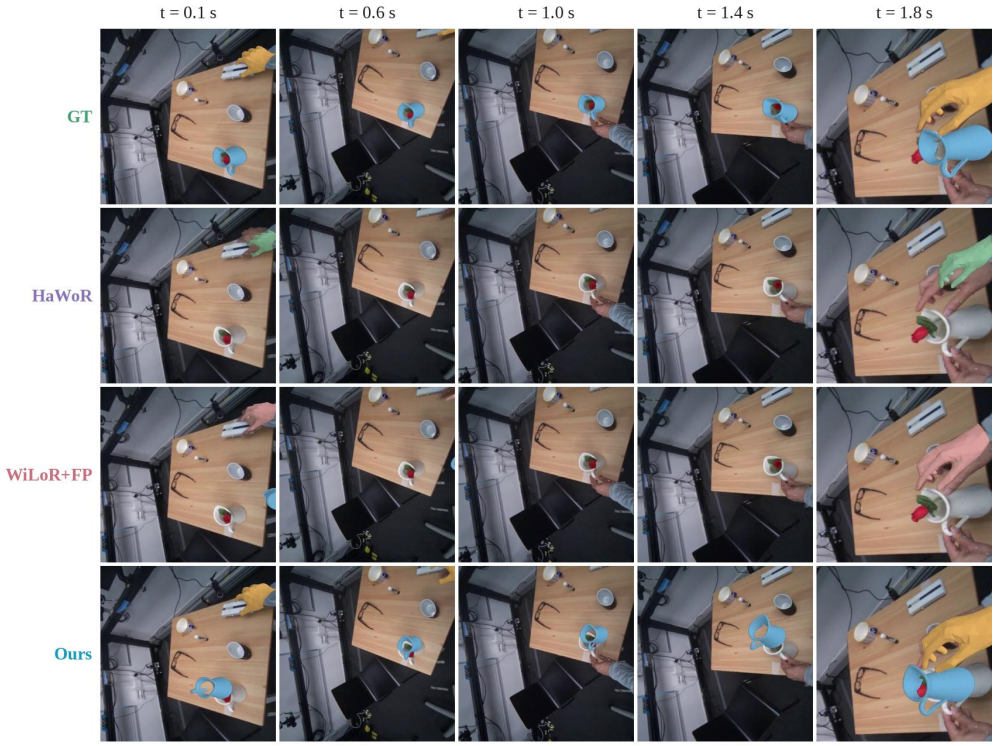
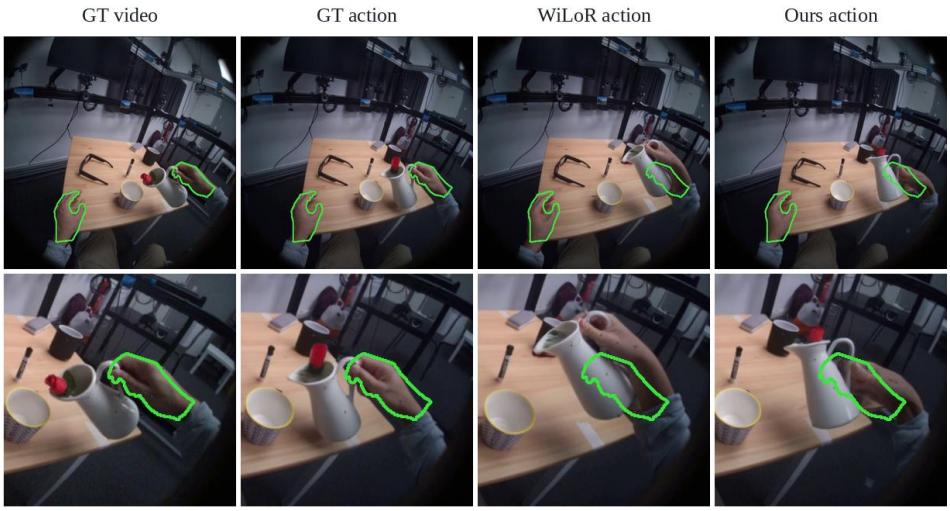
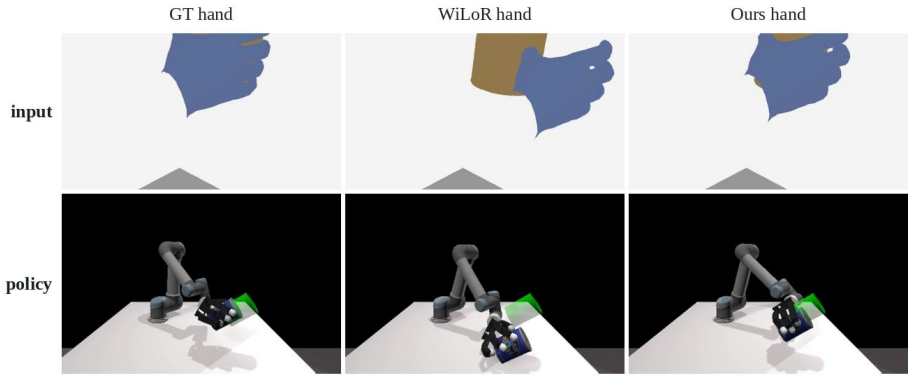


Figure S4: **Qualitative comparison over time on HOT3D** (Vase, first 2 s of the clip). Rows: ground truth, HaWoR (hands only), WiLoR + FoundationPose (the input $\mathbf{y}$), and DYNAMICHOI. Hands and objects are rendered after whole-trajectory (WA) alignment.



(a) Egocentric world modeling (EgoHOI). Green contour: ground-truth hand region used by mLPIPS.



(b) Dexterous manipulation (ViViDex, master chef can): retargeting input (top) and final state of the trained policy (bottom).

Figure S5: **Downstream applications** with different hand-trajectory sources.

D LIMITATIONS

Contact model. The minimum-norm inverse grasp mapping (Eq. S9) selects the least-effort force distribution over six fixed anchors and contains no friction-cone or unilateral (pushing-only) constraint. Its solution is one plausible grasp rather than the true one, and $\mathcal{L}_{\mathrm{act}}$ constrains a statistic of this estimate. Validating the inferred forces would require datasets with calibrated, localized hand–object forces.

External supports. We model in-air manipulation. When the object rests on or slides along a surface, part of its wrench is supplied by the support and is attributed to the hand.

Physical parameters. The hand uses a constant density, and an estimated object mass can deviate from the true one; errors of m^o scale the object wrench and hence the actuation.

Finite differences. All dynamics use second temporal differences of reconstructed poses at 30 fps, which amplify noise. This is why the dynamics serve as training-time supervision against ground-truth-derived quantities rather than as a test-time objective; residual high-frequency errors of the reconstructed object force and torque remain visible in Fig. S3.

Front-end dependence and object templates. The refiner corrects a bounded residual around $\mathbf{y}$. When the front end is already close to the ground truth, the room for improvement is small, and when it loses the object for many frames the refined object can remain off. The method also requires a template mesh of the manipulated object.