LEARNING WORLD DYNAMICS FROM DEBLURRING DISCREPANCY FOR HAND TRAJECTORY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting human hand trajectories in egocentric views is crucial for mixed reality, visual imitation learning, and human-robot collaboration. It enables autonomous systems to anticipate human intent and support timely interaction. However, this task remains challenging in dynamic interaction scenes, where moving objects and flexible hand adjustments introduce complex interaction dynamics. World models offer a promising way to address this challenge by modeling and predicting future world states beyond hand positions. Nonetheless, pronounced motion blur from fast hand-object motion and rapid viewpoint changes disturbs the observed interaction patterns and makes future visual states highly uncertain. In this paper, we propose to learn world dynamics from ***deblurring discrepancy***, a feature-level residual between blurred and deblurred visual representations. Our key insight is that motion blur manifests as visual ambiguity in image space, but originates from structured temporal integration during image formation. The feature discrepancy induced by deblurring thus preserves motion-related cues of hands, objects, and the camera. Based on this, we propose a deblurring-discrepancy-guided framework, dubbed **DYNAMICHAND**, for egocentric hand trajectory prediction (EHTP). **DYNAMICHAND** extracts blur-induced motion cues from deblurring discrepancy features and predicts their future evolution to facilitate world-state modeling and hand forecasting. We further design a discrepancy-aware distillation strategy, where discrepancy strength and variation adaptively weight the transfer of teacher latents to a lightweight student model for faster inference. Extensive experiments on public datasets and our new benchmark show that **DYNAMICHAND** consistently improves EHTP performance over existing baselines, demonstrating that deblurring discrepancy can serve as an effective bridge between visual restoration and world-dynamics learning in dynamic interaction scenes.

1 INTRODUCTION

With the development of mixed reality and embodied AI, egocentric hand trajectory prediction (EHTP) has become important for anticipating user intent, supporting timely interaction, and learning from human behavior. Despite recent progress (Bao et al., 2025; Ma et al., 2026a; Chen et al., 2026; Ma et al., 2026b), EHTP remains challenging in dynamic interaction scenes. Moving objects, flexible hand adjustments, and head-induced viewpoint changes within the rapid hand-object interaction process jointly produce complex visual dynamics, making future hand waypoints difficult to infer reliably from past images and observed hand positions alone.

These challenges suggest that reliable EHTP requires reasoning beyond the observed hand motion itself to capture how the underlying hand-object interaction state evolves over time. World modeling provides a natural framework for such predictive representation learning (Bardes et al., 2024; Li et al., 2025; Yu et al., 2026). Recent world modeling approaches have shown that learning the temporal evolution of world states can capture spatiotemporal regularities beyond task-specific labels, benefiting downstream applications (Hansen et al., 2024; Hafner et al., 2025; Ye et al., 2026). However, learning such predictive dynamics through appearance-level state prediction is unreliable in high-speed egocentric interactions. Fast hand-object motion and rapid viewpoint changes introduce pronounced motion blur, which disturbs the observed interaction patterns and makes future appearances highly uncertain. Thus, conventional prediction objectives may force the model to reconstruct ambiguous appearances, rather than focus on the motion cues that are most relevant to EHTP.

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.

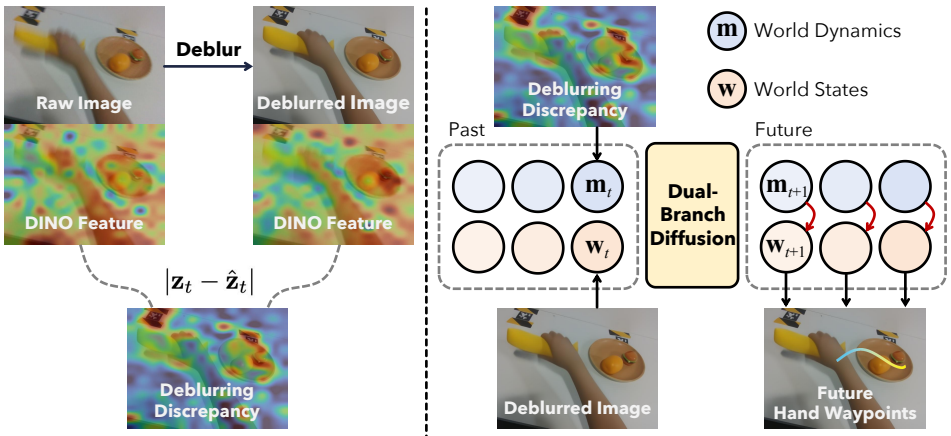


Figure 1: (a) Deblurring discrepancy provides a useful proxy for blur-induced world dynamics. (b) **DynamicHand** learns world dynamics from deblurring discrepancy to better predict future world states for EHTP with a dual-branch diffusion model.

Image deblurring can alleviate the visual degradation caused by motion blur (Zamir et al., 2021; Chen et al., 2022; Kong et al., 2025). Nonetheless, we argue that motion blur is not merely a nuisance factor to be removed. While it manifests as visual ambiguity in image space, it originates from structured temporal integration during image formation and therefore carries motion cues about hands, objects, and the head-mounted camera. **This motivates us to exploit motion blur to better understand how the world evolves rather than simply suppress it**.

To this end, as shown in Fig. 1(a), we first introduce ***deblurring discrepancy***, defined as the feature-level residual between blurred and deblurred egocentric visual representations. Rather than treating deblurred features as the only visual state, we regard deblurring discrepancy as a complementary dynamics-aware representation that captures the visual components modified by restoration. These components highlight motion-corrupted regions and thus encode interaction dynamics that are difficult to infer from deblurred appearance alone. Then, we propose **DynamicHand**, a deblurring-discrepancy-guided EHTP framework with a dual-branch predictive structure. As shown in Fig. 1(b), one branch models future world states from deblurred visual representations, while the other branch predicts future evolution of deblurring-discrepancy features. The predicted discrepancy features are used as dynamics-aware conditions for the world-state branch, enabling hand trajectory prediction to exploit both restored appearance and blur-induced motion cues. We also design a discrepancy-aware consistency model (DACM) to transfer the prediction performance to a lightweight student model. This enables efficient inference while preserving world dynamic cues learned by the full model.

Our contributions are fourfold: (1) We identify motion blur as a structured source of motion cues associated with world dynamics rather than merely visual degradation, and introduce deblurring discrepancy for a complementary dynamic state in EHTP. (2) We propose **DynamicHand**, a dual-branch framework that jointly forecasts deblurred world states and deblurring-discrepancy features, using the latter as world-dynamics-aware conditioning for hand trajectory prediction. (3) We design a discrepancy-aware consistency model that distills discrepancy-guided dynamics into a lightweight student for efficient inference. (4) Extensive experiments on public datasets and our new dynamic interaction benchmark demonstrate consistent gains over existing EHTP baselines, validating deblurring discrepancy as an effective bridge between visual restoration and world-dynamics learning.

2 Related Work

Egocentric Hand Trajectory Prediction. Understanding human hand motion patterns from egocentric views has become increasingly important for inferring human intent and guiding robotic manipulation in mixed reality and embodied AI (Gu et al., 2024; Li et al., 2026; Ma et al., 2026a). Early works (Liu et al., 2022; Bahl et al., 2023; Bao et al., 2023) attend to using Transformer blocks to encode egocentric images and past hand positions, autoregressively predicting future trajectories. To mitigate the interference of head-mounted camera egomotion, Diff-IP2D (Ma et al., 2025b), EMAG (Hatano et al., 2024), and MADiff (Ma et al., 2025a) propose to integrate past head-motion

representations into hand motion forecasting. Inspired by these works, Uni-Hand (Ma et al., 2026a) further explicitly predicts future camera egomotion to better capture head-hand coordination with a latent world-action-model paradigm. To improve prediction performance by scaling up the model capacity, HandsOnVLM (Bao et al., 2025) and EgoMAN (Chen et al., 2026) both fine-tune powerful VLMs (Huang et al., 2024; Bai et al., 2025) by formulating hand trajectory prediction as a vision-language reasoning problem. Despite this progress, the existing methods are substantially degraded by motion blur and scene-level uncertainty in dynamic interaction scenes with rapid hand-object movements. In contrast, we regard motion blur as a source of motion cues and introduce latent world-dynamics modeling based on deblurring discrepancy for EHTP.

Motion Forecasting in Dynamic Interaction. Dynamic interaction scenes (e.g., moving-object grasping and handshaking) are characterized by non-stationary targets, fast relative motion, and tightly coupled subject-object dynamics, requiring models to capture evolving hand-object states and interaction dynamics. Early robotics studies address dynamic manipulation by estimating object motion and optimizing end-effector trajectories (Sleiman et al., 2019; Yan et al., 2024). Recently, DynamicVLA (Xie et al., 2026) introduces temporal reasoning and closed-loop action streaming for manipulating moving objects. However, these methods largely assume clean visual observations and robot-centric action supervision, leaving motion blur in dynamic egocentric forecasting unexplored. To our knowledge, we are the first to exploit motion blur as a predictive cue for hand forecasting in dynamic interactions, capturing coupled dynamics of hands, objects, and the egocentric camera.

3 METHODOLOGY

Our goal is to learn world dynamics from deblurring discrepancy between raw egocentric observations and their deblurred counterparts, to facilitate EHTP. We begin by formulating deblurring discrepancy as a feature residual that exposes blur-induced dynamic cues hidden in observations (Sec. 3.1). Building on this, we introduce a dual-branch predictive framework that conditions world state modeling and hand trajectory forecasting on predicted discrepancy states (Sec. 3.2). Ultimately, we describe a discrepancy-aware consistency model that distills discrepancy-guided dynamics from the full predictor into a lightweight student model for efficient inference (Sec. 3.3).

3.1 DEBLURRING DISCREPANCY FORMULATION

Given an egocentric video clip $\mathbf{X} = \{I_t\}_{t=1}^{T}$, where I_t denotes the raw RGB observation at time t, we first obtain its deblurred counterpart $\hat{\mathbf{X}} = \{\hat{I}_t\}_{t=1}^{T}$ using an off-the-shelf deblurring model. We then extract visual representations from both observations with a frozen visual encoder $\Phi(\cdot)$:

$$\mathbf{z}_t = \Phi(I_t), \qquad \hat{\mathbf{z}}_t = \Phi(\hat{I}_t), \tag{1}$$

where $\mathbf{z}_t, \hat{\mathbf{z}}_t \in \mathbb{R}^{N\times C}$ denote patch-level features with N spatial tokens and C channels. We define ***deblurring discrepancy*** as the feature-level residual induced by the deblurring transformation:

$$\Delta_t = \psi(\mathbf{z}_t, \hat{\mathbf{z}}_t) = |\mathbf{z}_t - \hat{\mathbf{z}}_t|, \tag{2}$$

where $\Delta_t \in \mathbb{R}^{N\times C}$ preserves spatially localized differences between the blurred and deblurred representations with the absolute residual. The discrepancy sequence is denoted as $\mathbf{\Delta}_{1:T} = \{\Delta_t\}_{t=1}^{T}$. This formulation is motivated by the image formation process. Motion blur can be viewed as temporal integration of latent sharp observations during exposure:

$$I_t(\mathbf{x}) \approx \int_{\tau\in\Omega_t} K_t(\tau, \mathbf{x})\, S_\tau\big(\mathcal{W}_{t\leftarrow\tau}(\mathbf{x})\big) d\tau, \tag{3}$$

where $\mathbf{x}$ denotes a pixel location in the image plane, S_τ denotes the latent sharp scene state at time τ, $\mathcal{W}_{t\leftarrow\tau}$ is the motion-induced warp from time τ to the reference time t, $K_t(\tau, \mathbf{x})$ is the exposure weighting kernel, and Ω_t is the exposure interval. The warp $\mathcal{W}_{t\leftarrow\tau}$ is determined by the relative motion of hands, objects, and the head-mounted camera. Thus, although temporal integration appears as visual ambiguity, it is governed by structured interaction dynamics during image formation. Here, we use *world dynamics* in a task-oriented latent predictive sense, referring to temporal evolution of interaction states relevant to hand forecasting, rather than a complete physical state representation. Following the principle of latent world modeling (Hansen et al., 2024; Sun et al., 2026), we do not explicitly recover the exposure kernel or motion-induced warp. Instead, we encode their task-relevant effects into a predictive latent state. Accordingly, we encode each patch-level deblurring discrepancy with a learnable discrepancy encoder E_Δ:

$$\mathbf{m}_t = E_\Delta(\Delta_t), \qquad \mathbf{M}_{1:T} = \{\mathbf{m}_t\}_{t=1}^{T}, \tag{4}$$

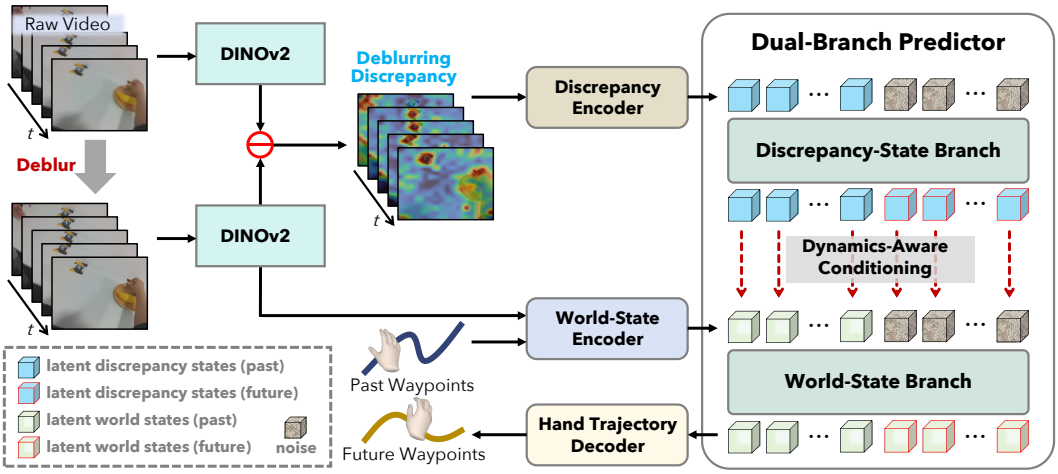


Figure 2: System overview of **DYNAMICHAND**. The Discrepancy-State Branch forecasts future deblurring-discrepancy states, which condition the World-State Branch for EHTP. For clearer visualization, the patch-level discrepancy maps are upsampled to the original image resolution.

where $\mathbf{m}_t \in \mathbb{R}^d$ denotes the discrepancy state at time t. Unlike $\hat{\mathbf{z}}_t$ which describes the restored visual state, $\mathbf{m}_t$ encodes the feature response induced by removing motion blur. It provides a compact proxy for blur-induced dynamic cues associated with the within-exposure motion that produces the blurred observation. We therefore use $\mathbf{M}_{1:T}$ as a dynamics-aware signal for predicting future world states and hand trajectories. With the discrepancy states defined above, our work can be formulated as conditional future trajectory modeling. Let $\mathbf{H}_{1:T_o}$ denote the observed hand trajectories, $\hat{\mathbf{Z}}_{1:T_o}$ denote the deblurred visual states, and $\mathbf{M}_{1:T_o}$ denote the deblurring-discrepancy states extracted from the observed frames. Our goal is to predict future hand trajectories $\mathbf{H}_{T_o+1:T}$ by modeling

$$p_\theta\left(\mathbf{H}_{T_o+1:T} \mid \mathbf{H}_{1:T_o}, \hat{\mathbf{Z}}_{1:T_o}, \mathbf{M}_{1:T_o}\right). \tag{5}$$

This conditional formulation makes deblurring discrepancy an explicit part of the predictive state, rather than treating deblurring as a purely visual preprocessing step.

3.2 DYNAMICHAND: DUAL-BRANCH PREDICTIVE FRAMEWORK

Based on the above formulation, we design **DYNAMICHAND** as a dual-branch predictive framework that separates restored world-state modeling from blur-induced world-dynamics modeling. Deblurred visual features provide a cleaner description of the current world state, whereas deblurring discrepancy provides blur-induced motion cues associated with world dynamics. As shown in Fig. 2, instead of collapsing them into a single representation, **DYNAMICHAND** models their evolution through two coupled branches: a World-State (WS) Branch for deblurred latent world-state prediction, and a Discrepancy-State (DS) Branch for deblurring-discrepancy state prediction.

Given observed frames $\{I_t\}_{t=1}^{T_o}$ and their deblurred counterparts $\{\hat{I}_t\}_{t=1}^{T_o}$, we extract deblurred visual features $\hat{\mathbf{z}}_t$ and discrepancy features Δ_t. The discrepancy features are encoded into latent discrepancy states $\mathbf{m}_t$ as introduced in Sec. 3.1. Meanwhile, the world-state encoder fuses deblurred visual features $\hat{\mathbf{z}}_t$ with observed hand waypoints $\mathbf{h}_t$:

$$\mathbf{w}_t = E_W\left(\hat{\mathbf{z}}_t, \mathbf{h}_t, \mathbf{q}\right), \quad t = 1, \ldots, T_o, \tag{6}$$

where $\mathbf{q}$ denotes the queried hand keypoint identity (Ma et al., 2026a), and $\mathbf{w}_t$ represents the deblurred visual world state incorporating the observed hand motion context.

To predict the future, **DYNAMICHAND** constructs two latent sequences, $\mathbf{W}_{1:T}$ and $\mathbf{M}_{1:T}$, where the observed states are anchored and the future states are generated by latent diffusion models of their respective branches. The Discrepancy-State Branch first forecasts the future discrepancy states:

$$\hat{\mathbf{M}}_{T_o+1:T} = F_\Delta\left(\mathbf{M}_{1:T_o}, \boldsymbol{\epsilon}_\Delta\right), \tag{7}$$

where $\boldsymbol{\epsilon}_\Delta$ denotes the initial noise for future discrepancy states. To improve prediction efficiency, as shown in Fig. 3(a), we use stacked Mamba layers (Gu & Dao, 2023) as the denoising model in

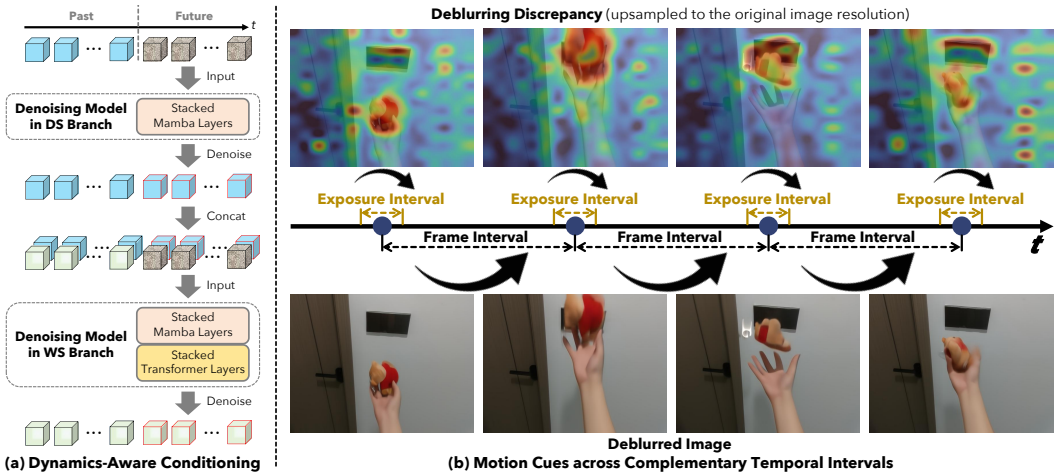


Figure 3: Dynamics-aware conditioning (a) is achieved by denoising future discrepancy states and world states sequentially with stacked Mamba and Transformer layers. **DYNAMICHAND** jointly captures motion cues across complementary temporal intervals (b) with the dual-branch framework.

the Discrepancy-State Branch. The full dynamics-aware condition $\tilde{\mathbf{m}}_t \in \tilde{\mathbf{M}}_{1:T}$ is then formed by concatenating observed discrepancy states $\mathbf{m}_t$ and predicted future counterparts $\hat{\mathbf{m}}_t$, which captures potential information about fast hand-object motion and viewpoint changes of dynamic interaction.

Conditioned on $\tilde{\mathbf{M}}_{1:T}$, the World-State Branch further predicts future deblurred latent world states:

$$\hat{\mathbf{W}}_{T_o+1:T} = F_W\left(\mathbf{W}_{1:T_o}, \tilde{\mathbf{M}}_{1:T}, \boldsymbol{\epsilon}_W\right), \tag{8}$$

where $\boldsymbol{\epsilon}_W$ is the initial noise for future world states. Concretely, the discrepancy states are first concatenated with the past world states and future noises along the channel dimension, serving as dynamics-aware conditioning. Then, as shown in Fig. 3(a), stacked Mamba (Gu & Dao, 2023) and Transformer (Vaswani et al., 2017) layers are utilized as the denoising model to convert the future noises to the predicted world states. Since world dynamics mainly reflect short-term state changes induced by rapid motion around the current moment, they can be effectively modeled with the uni-directional selective scan of Mamba without requiring long-range memory. In contrast, world states encode scene-level semantic evolution and hand-object context over a longer horizon, where Transformer layers are additionally introduced to enhance long-range temporal dependency modeling. The coupling of world dynamics and world states in our paradigm allows the model to use restored appearance for semantic stability while using deblurring discrepancy for motion sensitivity.

Finally, the predicted future world states are decoded into hand trajectories:

$$\hat{\mathbf{h}}_t = D_H(\hat{\mathbf{w}}_t), \quad t = T_o + 1, \ldots, T, \tag{9}$$

where the decoder D_H functions as an inverse-dynamics model (Ye et al., 2026). As can be noted, by explicitly predicting latents in both branches, **DYNAMICHAND** avoids direct future image reconstruction while still learning the temporal structure needed for hand trajectory forecasting. Fig. 3(b) also illustrates that the dual-branch predictive framework captures motion cues across complementary temporal intervals. The Discrepancy-State Branch captures local world changes within each exposure interval, while the World-State Branch models the evolution across frame intervals. By coupling intra-frame blur-induced dynamics with inter-frame temporal prediction, **DYNAMICHAND** obtains a richer dynamics-aware representation for hand trajectory forecasting.

3.3 DISCREPANCY-AWARE CONSISTENCY DISTILLATION

Although the dual-branch predictor learns discrepancy-guided world dynamics effectively, it requires deblurring, visual feature extraction, and iterative diffusion sampling, which limits inference efficiency. Thus, we introduce a *Discrepancy-Aware Consistency Model* (DACM) to distill the full predictor into a lightweight raw-RGB-only student.

Given a training sequence, the frozen teacher takes deblurred visual features and deblurring-discrepancy features as privileged inputs, and produces clean teacher world-state and discrepancy-state latent targets, $\mathbf{u}^{\mathrm{T}}_{1:T}$ and $\mathbf{v}^{\mathrm{T}}_{1:T}$, with its two branches. The student encodes the raw RGB sequence

directly with a lightweight temporal encoder:

$$\mathbf{s}_{1:T}, \mathbf{r}_{1:T} = E_{\mathrm{RGB}}(\mathbf{X}_{1:T}), \tag{10}$$

where $\mathbf{s}$ and $\mathbf{r}$ denote scene and motion representations, respectively. We concatenate them with a one-hot hand target identity to form the student's world-state inputs. The observed motion representations also contribute to the dynamics condition. Considering there is no future input observation during real deployment, we replace all future RGB frames with the last observed frame during training and test before encoding. The student follows the teacher's discrepancy-to-world-state reasoning order, but performs consistency denoising (Song et al., 2023) in a single step. Given a sampled diffusion step k, the student first maps the noisy discrepancy latent directly to its clean estimate:

$$\hat{\mathbf{v}}^{\mathrm{S}}_{0|k} = G^{\mathrm{S}}_{\Delta}\left(\mathbf{y}^{\mathrm{S}}_{k}, k\right), \tag{11}$$

where $\mathbf{y}^{\mathrm{S}}_{k}$ denotes the noisy input to the Discrepancy-State Branch, and $\hat{\mathbf{v}}^{\mathrm{S}}_{0|k}$ denotes its one-step estimate at the clean endpoint. The predicted future discrepancy states are combined with the observed RGB-based motion representations to construct the dynamics condition $\tilde{\mathbf{r}}^{\mathrm{S}}_{1:T}$. The World-State Branch then directly estimates the clean world-state latent under this condition:

$$\hat{\mathbf{u}}^{\mathrm{S}}_{0|k} = G^{\mathrm{S}}_{W}\left(\mathbf{x}^{\mathrm{S}}_{k}, \tilde{\mathbf{r}}^{\mathrm{S}}_{1:T}, k\right), \tag{12}$$

where $\mathbf{x}^{\mathrm{S}}_{k}$ denotes the noisy input to the World-State Branch. DACM thus preserves the teacher's discrepancy-to-world-state reasoning structure while replacing deblurring, visual feature extraction, and iterative diffusion sampling with raw-RGB encoding and one-step consistency denoising.

To emphasize samples containing stronger dynamic cues, DACM weights distillation according to discrepancy strength and temporal variation. Let $\mathbf{m}_t$ denote the teacher discrepancy representation at video frame t, and let $\mathcal{U}$ denote the set of future-frame indices. We compute

$$e = \frac{1}{|\mathcal{U}|}\sum_{t\in\mathcal{U}} \|\mathbf{m}_t\|_2, \qquad d = \frac{1}{|\mathcal{U}|}\sum_{t\in\mathcal{U}} \|\mathbf{m}_t - \mathbf{m}_{t-1}\|_2, \tag{13}$$

and obtain the discrepancy-aware weight

$$\omega = 1 + \alpha \cdot \sigma\left(\mathrm{Norm}(e + \lambda d)\right), \tag{14}$$

where $\mathrm{Norm}(\cdot)$ denotes batch-wise standardization. e measures discrepancy strength and d captures its temporal variation. Samples with stronger or more rapidly varying blur-induced dynamics thus receive larger distillation weights, prioritizing dynamically informative sequences.

DACM is trained with latent distillation, consistency learning, and trajectory supervision, where the discrepancy-aware weight ω is applied to the latent distillation and consistency objectives. Please refer to Sec. A.1 for detailed training objectives of the full model and DACM. At inference time, the student model uses only observed raw RGB frames and performs one-step prediction for future hand waypoints. This facilitates efficient reasoning in applications with stringent real-time requirements.

4 Experiments

DynamicHand Configuration. We employ NAFNet (Chen et al., 2022) to deblur raw videos and adopt DINOv2 (Oquab et al., 2023) for visual representations by default. Deblurred DINOv2 features are used as world-state inputs, while the deblurring-discrepancy features are computed from the feature-level residual between raw and deblurred observations. The dual-branch predictor of the full model uses latent diffusion with 1000 diffusion steps for training and 10 steps for DDIM inference (Song et al., 2020). All latent dimensions are set to 1024. The raw-RGB temporal encoder of the DACM student consists of four convolution layers, a depth-wise temporal convolution module, and two projection heads that produce scene and motion representations. Additional training protocol and architecture details are provided in Sec. A.1 and Sec. A.2.

Datasets and Baselines. We evaluate **DynamicHand** on the Ego4D hand forecasting benchmark (Grauman et al., 2022), EgoPAT3D-DT (Li et al., 2022; Bao et al., 2023), and our new dynamic interaction benchmark, **EgoDyn**. Ego4D and EgoPAT3D-DT already contain diverse scenarios affected by motion blur, and their evaluation protocols are modified from Ma et al. (2026a) (see Sec. A.3). In contrast, as shown in Tab. 1, **EgoDyn** specifically collects more interaction clips involving dynamic objects for comprehensive validation. We construct it by recording egocentric RGB videos at 30 FPS using a head-mounted Intel RealSense D435i. Following Ma et al., we predict hand

Table 1: Statistics of our **EgoDyn** benchmark. Different numbers of past and future frames are used across scenarios to enable extensive evaluation under varying observation and prediction horizons.

Scenario	Past	Future	Videos	Training Clips	Validation Clips	Test Clips	Total Clips
Moving-Object Grasping	5	10	637	18,536	200	1,634	20,370
One-Hand Tossing	3	6	639	6,131	200	523	6,854
Table-Tennis Bouncing	3	6	567	6,398	200	657	7,255
Handshaking	3	6	639	9,502	200	814	10,516
High-Five	1	3	600	2,751	200	246	3,197
Human-Robot Handover	3	10	651	7,392	200	616	8,208
Total	–	–	**3,733**	**50,710**	**1,200**	**4,490**	**56,400**

Table 2: Quantitative comparison on our **EgoDyn** benchmark. The best result is shown in **bold**, and the second-best result is underlined.

Approach	Moving-Object Grasping		One-Hand Tossing		Table-Tennis Bouncing	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
CVM (Schöller et al.)	0.174	0.300	0.177	0.312	0.056	0.095
HandsOnVLM (Bao et al.)	0.331	0.388	0.195	0.216	0.281	0.300
EMAG (Hatano et al.)	0.137	0.221	0.103	0.176	0.050	0.066
USST (Bao et al.)	0.106	0.154	0.071	0.104	0.043	0.062
S-Mamba (Wang et al.)	0.083	0.114	0.054	0.077	0.050	0.069
EgoMAN (Chen et al.)	0.082	0.110	0.063	0.101	0.054	0.077
Uni-Hand (Ma et al.)	0.079	0.105	0.055	0.072	0.048	0.063
DYNAMICHAND (DACM)	0.076	**0.094**	0.053	0.067	0.041	0.060
DYNAMICHAND (Full)	**0.074**	0.103	**0.049**	**0.065**	**0.039**	**0.056**

Approach	Handshaking		High-Five		Human-Robot Handover	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
CVM (Schöller et al.)	0.085	0.140	0.174	0.225	0.111	0.213
HandsOnVLM (Bao et al.)	0.203	0.208	0.337	0.340	0.192	0.224
EMAG (Hatano et al.)	0.081	0.123	0.111	0.139	0.064	0.091
USST (Bao et al.)	0.066	0.097	0.161	0.204	0.046	0.066
S-Mamba (Wang et al.)	0.058	0.074	0.084	0.065	0.044	0.058
EgoMAN (Chen et al.)	0.063	0.088	0.080	0.069	0.046	0.065
Uni-Hand (Ma et al.)	0.054	0.069	0.079	0.071	0.038	0.047
DYNAMICHAND (DACM)	0.058	0.069	0.072	0.057	0.039	0.045
DYNAMICHAND (Full)	**0.046**	**0.060**	**0.066**	**0.050**	**0.031**	**0.039**

centers on Ego4D and EgoPAT3D-DT, while attending to three representative hand joints (i.e., the wrist, thumb tip, and middle-finger tip) for **EgoDyn**. We compare **DYNAMICHAND** with existing EHTP methods including CVM (Schöller et al., 2020), HandsOnVLM (Bao et al., 2025), USST (Bao et al., 2023), S-Mamba (Wang et al., 2025), Diff-IP2D (Ma et al., 2025b), EMAG (Hatano et al., 2024), EgoMAN (Chen et al., 2026), and a latent world action model, Uni-Hand (Ma et al., 2026a). For all image-based baselines, we use the same deblurred images as visual inputs (see Sec. A.11).

4.1 MAIN RESULTS

Following prior works (Bao et al., 2023; Ma et al., 2025a; 2026a), we evaluate EHTP performance using Average Displacement Error (ADE) and Final Displacement Error (FDE) between predicted and GT future hand/joint waypoints in normalized image coordinates. As shown in Tab. 2, **DYNAMICHAND** (Full) consistently outperforms the baselines across all **EgoDyn** scenarios, despite their diverse motion characteristics and prediction horizons. This demonstrates that **DYNAMICHAND**'s performance advantage is not limited to a specific type of dynamic interaction pattern. More qualitative comparisons can be found in Fig. 4 and Sec. A.4. Despite using only raw RGB observations and one-step inference, **DYNAMICHAND** (DACM) remains competitive with the strongest baselines. This suggests that our distillation strategy effectively transfers the teacher's discrepancy-to-world-state reasoning into the lightweight student. As reported in Tab. 3, **DYNAMICHAND** (Full) also achieves the best performance across all metrics on Ego4D and EgoPAT3D-DT. The consistent improvements across seen and unseen environments of EgoPAT3D-DT indicate that the learned discrepancy-guided dynamics generalize beyond the interaction scenarios presented during training.

4.2 ABLATION STUDY

Discrepancy-Guided Dual-Branch Modeling. Tab. 4 shows the contributions of deblurring discrepancy and world states to hand forecasting. The Discrepancy-State Branch alone performs poorly, confirming that discrepancy cannot entirely replace the appearance and semantic context encoded by world states. The World-State Branch provides a substantially stronger basis, while adding observed discrepancy yields only modest improvements. This indicates that historical discrepancy is informative but insufficient to describe how interaction dynamics will evolve beyond the observation

Table 3: Quantitative comparison on the Ego4D and EgoPAT3D-DT datasets.

Approach	Ego4D		EgoPAT3D-DT (Seen)		EgoPAT3D-DT (Unseen)	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
CVM (Schöller et al.)	0.223	0.256	0.117	0.219	0.086	0.164
HandsOnVLM (Bao et al.)	0.200	0.216	0.174	0.202	0.145	0.163
USST (Bao et al.)	0.172	0.190	0.077	0.127	0.061	0.105
S-Mamba (Wang et al.)	0.179	0.192	0.082	0.128	0.069	0.105
Diff-IP2D (Ma et al.)	0.186	0.204	0.087	0.139	0.074	0.114
Uni-Hand (Ma et al.)	0.179	0.190	0.078	0.127	0.062	0.107
DYNAMICHAND (DACM)	0.172	0.191	0.076	0.122	0.064	0.103
DYNAMICHAND (Full)	**0.164**	**0.181**	**0.071**	**0.117**	**0.055**	**0.097**

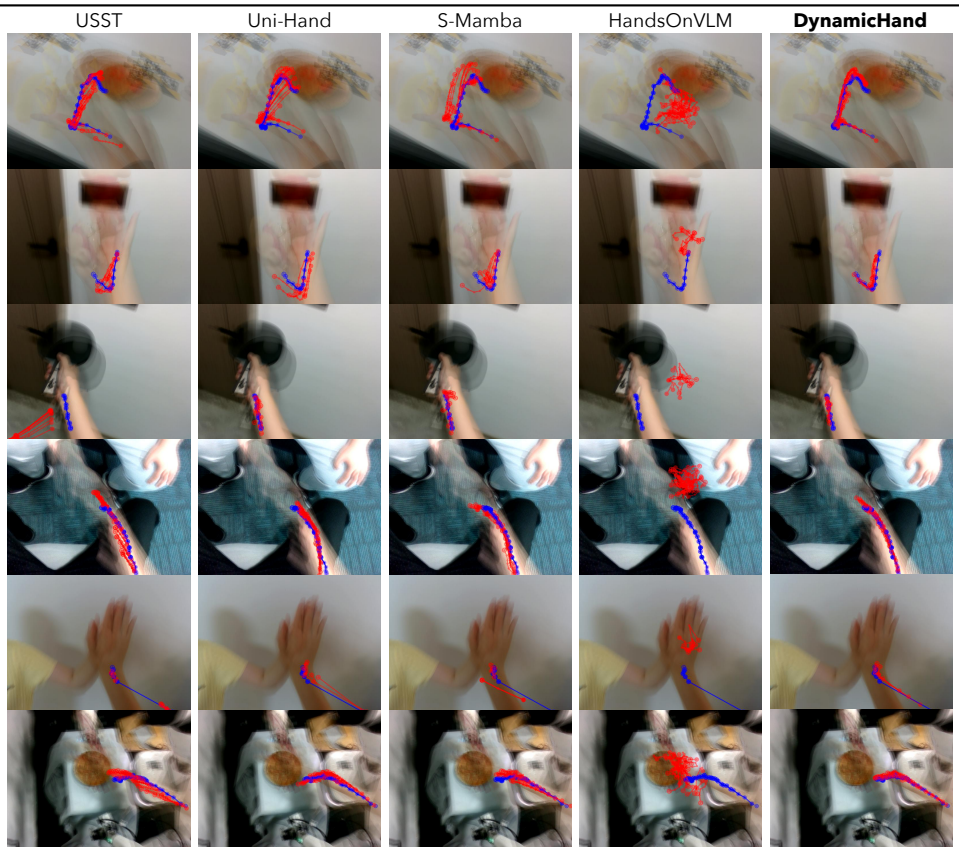


Figure 4: Qualitative comparison of hand trajectory prediction on **EgoDyn**. Red and blue curves denote predicted and GT *wrist* trajectories of all the clips within each video, respectively. Temporal ghost backgrounds visualize scene dynamics by overlaying frames sampled from each video.

horizon. The full **DYNAMICHAND** instead forecasts future discrepancy states and uses them to condition world-state prediction. Compared with the strongest ablated variant, it reduces ADE/FDE by 9.8%/11.8% on Handshaking and 11.4%/13.3% on Human-Robot Handover. These results highlight the benefit of propagating discrepancy-guided dynamics into future world-state prediction.

Discrepancy-Aware Consistency Model. Tab. 2 and Tab. 3 have presented the performance of **DYNAMICHAND** (DACM). Tab. 4 also ablates discrepancy-aware weighting in DACM. Removing the weight ω yields the largest overall degradation, demonstrating the benefit of allocating stronger supervision to dynamically informative samples. Among its components, discrepancy strength e contributes more substantially by identifying the magnitude of blur-induced dynamics, while temporal variation d complements it by describing how these dynamics evolve, particularly improving FDE. Their combination achieves the best overall performance, showing that effective distillation depends on both the magnitude and temporal evolution of deblurring discrepancy.

Visual Encoders. We further ablate the visual encoder in Tab. 5. DINOv2 achieves the best overall balance. DINOv3 and LingBot-Vision also perform competitively, whereas V-JEPA 2 produces consistently larger errors. These results suggest that the patch-level representations of DINOv2 provide a stable feature space in which the deblurring discrepancy preserves motion cues without sacrificing

Table 4: Ablation study of discrepancy-guided dual-branch modeling in the full model and discrepancy-aware weighting in DACM on **EgoDyn**. The *World-State Branch only* variant entirely removes dynamics-aware conditioning.

Variant	Handshaking		Human-Robot Handover	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓
Discrepancy-State Branch only	0.073	0.102	0.076	0.091
World-State Branch only	0.053	0.074	0.035	0.047
Observed discrepancy conditioning	0.051	0.068	0.035	0.045
DYNAMICHAND (Full)	**0.046**	**0.060**	**0.031**	**0.039**
w/o discrepancy-aware weight ω	0.063	0.076	0.042	0.051
w/o discrepancy strength e	0.061	0.073	0.042	0.049
w/o temporal variation d	**0.058**	0.072	0.040	0.046
DYNAMICHAND (DACM)	**0.058**	**0.069**	**0.039**	**0.045**

Table 5: Ablation study of different visual feature encoders on **EgoDyn**.

Visual Encoder	Handshaking		Human-Robot Handover	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓
V-JEPA 2 (Assran et al.)	0.051	0.068	0.037	0.044
LingBot-Vision (Fu et al.)	0.046	0.064	**0.031**	**0.039**
DINOv2 (Oquab et al.)	0.046	**0.060**	**0.031**	**0.039**
DINOv3 (Siméoni et al.)	**0.045**	0.062	0.032	**0.039**

semantic context. In contrast, V-JEPA 2 is optimized to learn temporally predictive and abstract representations that suppress low-level appearance variations, hindering the deblurring discrepancy estimation (see Fig. 5). More ablation studies are provided in the Appendix.

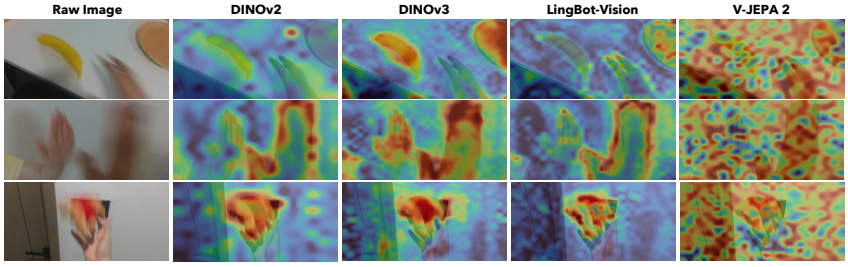


Figure 5: Visualization of deblurring discrepancy from different visual encoders. The feature-level discrepancies are overlaid on the deblurred images. Warmer colors indicate larger feature changes.

4.3 Inference Efficiency

Table 6: Per-sample inference runtime measured on one NVIDIA A100 GPU.

Approach	Runtime (ms) ↓
HandsOnVLM (Bao et al.)	5,500.1
USST (Bao et al.)	52.2
S-Mamba (Wang et al.)	**6.8**
Uni-Hand (Ma et al.)	92.3
DYNAMICHAND (DACM)	17.0
DYNAMICHAND (Full)	98.9

Tab. 6 compares the runtime of different methods under the same forecasting setting of Moving-Object Grasping in **EgoDyn**. **DYNAMICHAND** (Full) results in 98.9 ms per prediction. Our DACM effectively reduces the runtime to 17.0 ms by raw-RGB encoding and one-step consistency prediction. **DYNAMICHAND** (DACM) is also substantially faster than USST, Uni-Hand, and HandsOn-VLM while retaining competitive prediction accuracy, as shown in Tab. 2. These results demonstrate that DACM strikes a favorable balance between trajectory accuracy and inference efficiency.

5 Conclusion

In this paper, we propose **DYNAMICHAND**, an egocentric hand trajectory prediction framework that reinterprets motion blur as structured evidence of world dynamics, rather than mere visual degradation. To expose this evidence, we introduce *deblurring discrepancy*, the feature residual between raw and deblurred observations. **DYNAMICHAND** factorizes predictive modeling into two coupled branches, which connect motion occurring within each exposure interval to world-state evolution across frames. We further introduce DACM to transfer the discrepancy-to-world-state reasoning into a raw-RGB student with one-step inference. Experiments on public datasets and our new benchmark demonstrate consistent improvements in prediction accuracy together with a favorable accuracy-efficiency trade-off. We hope our exploration of deblurring discrepancy inspires future work to seek predictive structure in visual degradations traditionally treated only as nuisances. In future work, we plan to transfer this predictive capability to robotic tasks of dynamic object manipulation, where anticipating interaction dynamics is critical for timely and adaptive control.

AI USE STATEMENT

We used generative AI tools to assist with language editing, polishing, and drafting portions of the manuscript based on technical content, experimental results, and ideas provided by the authors. These tools were not used to conduct experiments, generate datasets, or establish mathematical claims. All AI-assisted text was critically reviewed, revised, and verified by the authors, who take full responsibility for the technical accuracy, citations, claims, and final content of the paper.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, pp. 13778–13790, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chen Bao, Jiarui Xu, Xiaolong Wang, Abhinav Gupta, and Homanga Bharadhwaj. Handsonvlm: Vision-language models for hand-object interaction prediction. *Transactions on Machine Learning Research*, 2025.

Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13702–13711, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Mingfei Chen, Yifan Wang, Zhengqin Li, Homanga Bharadhwaj, Yujin Chen, Chuan Qin, Ziyi Kou, Yuan Tian, Eric Whitmire, Rajinder Sodhi, et al. Flowing from reasoning to motion: Learning 3d hand trajectory prediction from egocentric human interaction videos. In *ECCV*, 2026.

Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception. *arXiv preprint arXiv:2607.05247*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, Priya Sundaresan, Peng Xu, Hao Su, Karol Hausman, Chelsea Finn, Quan Vuong, and Ted Xiao. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. In *International Conference on Learning Representations*, 2024.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Masashi Hatano, Ryo Hachiuma, and Hideo Saito. Emag: Ego-motion aware and generalizable 2d hand forecasting from egocentric videos. In *ECCVW*, 2024.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lingshun Kong, Jiangxin Dong, Jinhui Tang, Ming-Hsuan Yang, and Jinshan Pan. Efficient visual state space model for image deblurring. In *Proceedings of the computer vision and pattern recognition conference*, pp. 12710–12719, 2025.

Hao Li, Ganlong Zhao, Yufei Liu, Haotian Hou, Guoquan Ye, Tongyan Fang, Chunxiao Liu, Siyuan Huang, Jianbo Liu, Xiaogang Wang, et al. Ace-ego-0: Unifying egocentric human and robotic data for vla pretraining. *arXiv preprint arXiv:2606.17200*, 2026.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025.

Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *CVPR*, pp. 20971–20980, 2022.

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3282–3292, 2022.

Junyi Ma, Xieyuanli Chen, Wentao Bao, Jingyi Xu, and Hesheng Wang. Madiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Junyi Ma, Xieyuanli Chen, Jingyi Xu, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4291–4298. IEEE, 2025b.

Junyi Ma, Wentao Bao, Jingyi Xu, Guanzhong Sun, Yu Zheng, Erhang Zhang, Xieyuanli Chen, and Hesheng Wang. Uni-hand: Universal hand motion forecasting in egocentric views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Junyi Ma, Erhang Zhang, Haoran Yang, Ditao Li, Chenyang Xu, Guangming Wang, and Hesheng Wang. Robot learning from human videos: A survey. *arXiv preprint arXiv:2604.27621*, 2026b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Karl Pearson. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58:240–242, 1895.

Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jean-Pierre Sleiman, Jan Carius, Ruben Grandia, Martin Wermelinger, and Marco Hutter. Contact-implicit trajectory optimization for dynamic object manipulation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 6814–6821. IEEE, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Charles Spearman. The proof and measurement of association between two things. *International journal of epidemiology*, 39(5):1137–1150, 2010.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. Is mamba effective for time series forecasting? *Neurocomputing*, 619:129178, 2025.

Haozhe Xie, Beichen Wen, Jiarui Zheng, Zhaoxi Chen, Fangzhou Hong, Haiwen Diao, and Ziwei Liu. Dynamicvla: A vision-language-action model for dynamic object manipulation. *arXiv preprint arXiv:2601.22153*, 2026.

Lei Yan, Theodoros Stouraitis, Joao Moura, Wenfu Xu, Michael Gienger, and Sethu Vijayakumar. Impact-aware bimanual catching of large-momentum objects. *IEEE Transactions on Robotics*, 40:2543–2563, 2024.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Xiao Yu, Yichen Zhang, Mingzhang Wang, Shifang Zhao, Weizhe Liu, Yuyang Yin, Zhongwei Ren, Ning An, Xinglong Wu, Hao Liu, Houwen Peng, Yao Zhao, Jianchao Yang, Shuicheng Yan, Yunchao Wei, and Xiaojie Jin. From seeing to knowing the world: A survey of vision world models. *Preprints*, April 2026. doi: 10.20944/preprints202604.2072.v1. URL `https://doi.org/10.20944/preprints202604.2072.v1`.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

A Appendix

A.1 Training Recipe of **DynamicHand**

We train **DynamicHand** in two stages. The first stage optimizes the full dual-branch predictor, while the second freezes it as the teacher and distills its predictive structure into the DACM student.

Full Dual-Branch Predictor. During diffusion training, noise is added only to future latent states, while observed states remain anchored. The Discrepancy-State Branch first reconstructs the clean discrepancy states, which are combined with the observed discrepancy sequence to condition the World-State Branch. We denote the reconstruction losses of the two branches collectively by $\mathcal{L}_{\text{diff}}$. The predicted world states are decoded into future hand waypoints and supervised by the trajectory loss $\mathcal{L}_{\text{traj}}$. For the discrepancy states, an auxiliary residual loss $\mathcal{L}_{\text{disc}}$ aligns the predicted future states with their encoded targets. Together with the diffusion-prior regularization $\mathcal{L}_{\text{prior}}$ (Ma et al., 2025b), the full objective is summarized as

$$\mathcal{L}_{\text{Full}} = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \lambda_{\text{disc}}\mathcal{L}_{\text{disc}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} \tag{15}$$

DACM Student. The frozen teacher provides clean world-state and discrepancy-state targets. The noisy inputs $\mathbf{x}_k^{\text{S}}$ and $\mathbf{y}_k^{\text{S}}$ are constructed by retaining the observed raw-RGB student states and perturbing only the future teacher world-state and discrepancy-state targets according to the diffusion schedule. For each training sample, the inputs at k_h and k_l are generated from the same clean teacher targets using the same Gaussian noise. Thus, they form a paired sample along the same noising trajectory, rather than two independently corrupted inputs. The student performs one-step denoising at both noise levels. The latent distillation loss $\mathcal{L}_{\text{latent}}$ matches the student predictions at k_h to the clean teacher targets, while $\mathcal{L}_{\text{cons}}$ aligns the predictions at k_h and k_l, treating the latter as a stop-gradient consistency target. Both losses supervise the World-State and Discrepancy-State Branches. At inference, the future latent states are initialized with Gaussian noise at the terminal timestep, while the observed states remain anchored by the raw-RGB representations. The decoded world states are further constrained by waypoint, velocity, and smoothness losses, denoted by $\mathcal{L}_{\text{traj}}$, $\mathcal{L}_{\text{vel}}$, and $\mathcal{L}_{\text{smooth}}$, respectively. An auxiliary latent residual loss $\mathcal{L}_{\text{mot}}$ preserves the teacher's discrepancy-state dynamics.

The detached sample-wise weight ω_i defined in Eq. (14) of Sec. 3.3 is applied to latent distillation and consistency learning, directing the student to prioritize dynamics transfer on samples with stronger or more rapidly varying discrepancy. We leave the trajectory-level objectives uniformly weighted to preserve unbiased supervision in the hand-coordinate space. Therefore, the DACM objective is

$$\begin{aligned}\mathcal{L}_{\text{DACM}} = \mathbb{E}_i\Big[\omega_i \Big(\lambda_{\text{lat}}\mathcal{L}_{\text{latent}}^{(i)} + \lambda_{\text{con}}\mathcal{L}_{\text{cons}}^{(i)}\Big) \\ + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}}^{(i)} + \lambda_{\text{mot}}\mathcal{L}_{\text{mot}}^{(i)} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}^{(i)} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}^{(i)}\Big].\end{aligned} \tag{16}$$

All losses are computed independently for multiple prediction targets (e.g., the wrist, thumb tip, and middle-finger tip), and then averaged across them.

In the first stage, the full **DYNAMICHAND** is optimized with the AdamW optimizer (Kingma & Ba, 2014) using a learning rate of 5×10^{-5} and weight decay 10^{-4} for 1,000 epochs. We use a loss-aware timestep sampler for diffusion training. In the second stage, we train DACM by freezing the full **DYNAMICHAND** as the teacher. We sample a high diffusion timestep $k_h \in [20, 1000)$ and set a lower timestep $k_l = \lfloor 0.5k_h \rfloor$ for consistency learning. We train DACM for 300 epochs with the AdamW optimizer, learning rate 5×10^{-5}, weight decay 10^{-4}, and gradient clipping with maximum norm 1.0.

A.2 ADDITIONAL ARCHITECTURE DETAILS OF DYNAMICHAND

We employ a frozen DINOv2 encoder to extract a 16×16 grid of patch tokens from each raw and deblurred frame, yielding 256 tokens with 768 channels. The deblurred tokens are spatially averaged into a 768-dimensional frame descriptor, while their patch-level residual is retained for discrepancy encoding. All subsequent latent representations have 1024 dimensions. The discrepancy encoder reshapes each residual into a $768 \times 16 \times 16$ feature map and processes it with three convolutional blocks whose channel dimensions are $768 \to 256 \to 256 \to 1024$ and strides are 1, 2, and 2, respectively. Each convolution is followed by batch normalization and GELU activation, and adaptive average pooling produces a 1024-dimensional discrepancy state. For world-state encoding, the deblurred frame descriptor is linearly projected to 1024 dimensions, while each observed 2D hand waypoint is encoded by an MLP with GELU activation. These features are concatenated with one-hot vectors identifying the queried hand keypoint (Ma et al., 2026a) and projected to 1024 dimensions. The Discrepancy-State Branch comprises two Mamba layers. In the World-State Branch,

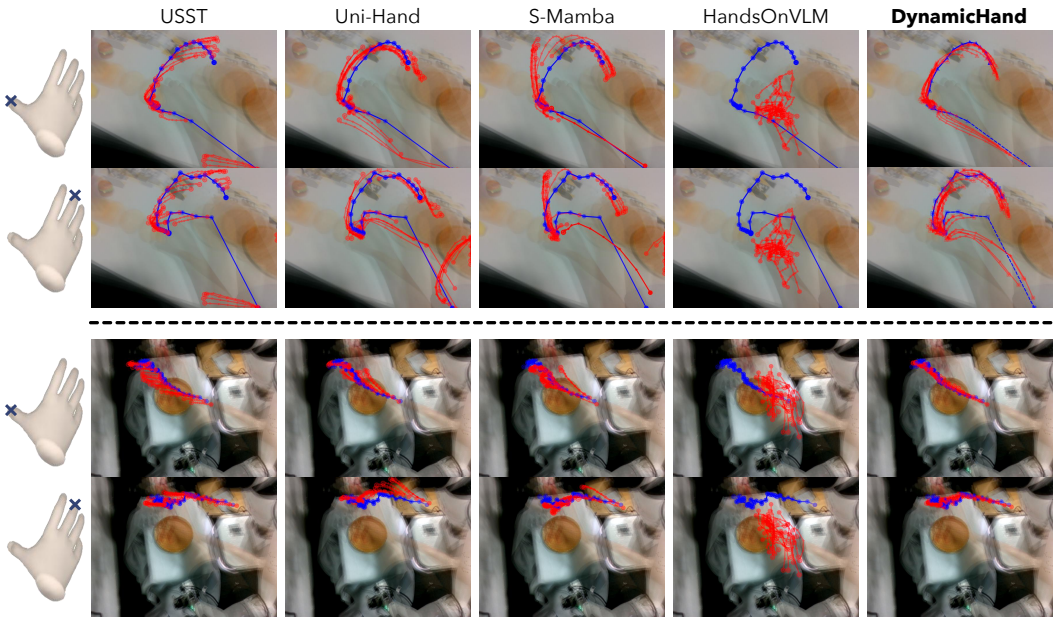


Figure 6: Qualitative comparison of *thumb-tip* and *middle-finger-tip* trajectory prediction on **EgoDyn**. Red and blue curves denote predicted and ground-truth trajectories, respectively.

each world state is concatenated with its discrepancy condition and processed by two Mamba layers and two Transformer decoder blocks. For clearer visualization, all deblurring-discrepancy feature maps shown in the figures of the main text are upsampled to the original image resolution for better display only.

A.3 Additional Dataset and Evaluation Details

Our evaluation protocols for Ego4D and EgoPAT3D-DT are adapted from Uni-Hand (Ma et al., 2026a). We retain its video-level assignments for the training, validation, and test sets, including the seen and unseen test partitions of EgoPAT3D-DT. We only modify the observation and prediction horizons and regenerate clips accordingly. For EgoPAT3D-DT, each clip contains 5 observed frames and 10 future frames. The training, seen-test, and unseen-test partitions contain 6,234, 1,554, and 2,246 videos, respectively, yielding 140,005, 32,895, and 44,165 clips. For Ego4D, each clip comprises 3 observed frames and 2 future frames. The retained training and test lists contain 10,018 and 500 video clips, of which 9,736 and 492 provide sufficient valid hand observations. For all methods, each test sample is evaluated over 10 independent inference runs, and the reported ADE and FDE are averaged over these runs. We apply the same inference protocol to all baselines for a consistent comparison.

Our proposed **EgoDyn** is collaboratively collected by eight participants using head-mounted Intel RealSense D435i at 30 FPS. Recordings from multiple participants are distributed across the dataset splits to maintain diverse motion patterns within each split. Automatic exposure is enabled throughout data collection to maintain visible hand-object interactions under varying illumination while preserving motion blur that naturally arises from rapid interaction and camera motion. To obtain trajectory annotations, we first deblur each video and then apply HaMeR (Pavlakos et al., 2024) to detect hand joints in every frame. We manually inspect the resulting trajectories and correct joint annotations with noticeable localization errors or identity inconsistencies. Clips with severe ambiguity that cannot be confidently corrected are discarded. Although these trajectories are generated with a hand detector, they directly localize hands in the actual future frames and therefore provide a meaningful reference for evaluating whether a model can anticipate future image-plane hand positions. Ma et al. (2026a) have further empirically validated that this detector-based labeling protocol does not introduce systematic evaluation bias. The training, validation, and test splits are mutually video-disjoint, with no video shared across any two splits. Subject identities are not explicitly enforced to be disjoint across splits, while each split contains recordings from multiple participants to maintain cross-subject diversity.

Table 7: Relative error reductions achieved by dynamics-aware conditioning over the World-State-Branch-only variant on **EgoDyn**.

	Motion Blur Intensity		Deblurring Discrepancy Strength	
Level	**ADE Reduction ↑**	**FDE Reduction ↑**	**ADE Reduction ↑**	**FDE Reduction ↑**
Low	4.7%	5.2%	4.3%	5.8%
Medium	12.1%	14.4%	10.9%	13.5%
High	20.0%	23.3%	22.4%	25.1%

A.4 Visualizations of More Prediction Targets

Fig. 6 extends the wrist-only visualization in Fig. 4 of the main text to fingertip targets. Predicting the fingertip is particularly challenging because its trajectory reflects both overall hand movement and local articulation. Across both thumb-tip and middle-finger-tip targets, **DynamicHand** tracks the ground-truth motion more closely, especially around changes in direction. Competing methods exhibit greater drift or scattered predictions. This validates that discrepancy-guided dynamics remain effective beyond a single hand keypoint. Additional qualitative results are provided in the supplementary video.

A.5 Impact of Motion Blur Intensity and Deblurring Discrepancy Strength on Prediction Performance

Motion Blur Intensity. To present whether our proposed dynamics-aware conditioning becomes more beneficial under stronger motion blur, we train **DynamicHand** (Full) and the World-State-Branch-only variant once on the complete **EgoDyn** training set and evaluate them on manually categorized low-, medium-, and high-blur test subsets. This avoids confounding blur intensity with subset-specific training. When trained jointly on all scenarios, **DynamicHand** (Full) achieves an ADE of 0.054 and an FDE of 0.074. As shown in Tab. 7, dynamics-aware conditioning reduces ADE/FDE over the World-State-Branch-only variant by 4.7%/5.2% under low blur, 12.1%/14.4% under medium blur, and 20.0%/23.3% under high blur. The increasing gains indicate that dynamics-aware conditioning becomes more effective as motion blur intensifies.

Deblurring Discrepancy Strength. We further perform an automatic, representation-level analysis using discrepancy strength. For each test clip, we compute the average discrepancy strength over observed frames only, $e_{\text{obs}} = \frac{1}{T_o}\sum_{t=1}^{T_o}\|\mathbf{m}_t\|_2$, and divide the clips into three equal-sized tertiles. No future observation is used to determine these subsets. Tab. 7 shows that the relative ADE/FDE reductions increase from 4.3%/5.8% in the low-strength subset to 10.9%/13.5% in the medium-strength subset and 22.4%/25.1% in the high-strength subset. This suggests that observed discrepancy strength can identify interactions in which blur-induced dynamics contribute more substantially to future trajectory prediction.

Both perceptual blur intensity and observed discrepancy strength exhibit the same monotonic trend. Stronger blur-related signals lead to larger gains from our devised dynamics-aware conditioning. Their agreement supports our central design choice of exploiting motion blur rather than merely suppressing it, and shows that deblurring discrepancy provides a meaningful indicator of when blur-induced motion cues are most valuable.

A.6 Ablation Study on Image Restoration Methods

We further validate whether discrepancy-guided prediction depends on a particular deblurring model by comparing Restormer (Zamir et al., 2022) and NAFNet (Chen et al., 2022) in Tab. 8. NAFNet achieves lower errors on Ego4D and the EgoPAT3D-DT unseen split, while Restormer performs slightly better on the seen split. The differences are modest across all three evaluations. This suggests that the benefit of deblurring discrepancy comes from modeling the change between raw and restored representations, rather than relying on a specific restoration architecture.

Table 8: Ablation study of deblurring models on the Ego4D and EgoPAT3D-DT datasets.

Approach	Ego4D		EgoPAT3D-DT (Seen)		EgoPAT3D-DT (Unseen)	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ADE ↓	FDE ↓
DYNAMICHAND with Restormer (Zamir et al., 2022)	0.167	0.185	**0.069**	**0.114**	0.058	0.101
DYNAMICHAND with NAFNet (Chen et al., 2022)	**0.164**	**0.181**	0.071	0.117	**0.055**	**0.097**

Table 9: Comparison with alternative motion representations on **EgoDyn**.

Motion Representation	Handshaking		Human-Robot Handover	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓
None (World-State Branch only)	0.053	0.074	0.035	0.047
Raw visual feature	0.060	0.083	0.046	0.059
Adjacent-frame feature difference	0.058	0.084	0.041	0.054
Optical flow (SEA-RAFT)	0.050	0.076	0.034	0.052
Deblurring discrepancy (Ours)	**0.046**	**0.060**	**0.031**	**0.039**

A.7 Comparison with Alternative Motion Representations

We further conduct experiments to compare deblurring discrepancy with alternative representations under the same dual-branch predictive framework. In addition to adjacent-frame feature difference and optical flow, we introduce a raw-visual-feature control to explore whether the performance gain simply comes from providing an additional visual representation or predictive branch. Specifically, for the raw visual feature variant, we directly feed DINOv2 features extracted from the raw images into the auxiliary predictive branch, using the same encoder and branch architecture as the deblurring-discrepancy variant. For adjacent-frame feature difference, we compute the absolute difference between DINOv2 features extracted from consecutive deblurred frames. For optical flow, we use SEA-RAFT (Wang et al., 2024) to estimate dense flow between consecutive deblurred frames, and project the resulting flow representation to the same feature dimension before feeding it into the discrepancy encoder. All other network components and training settings are kept unchanged.

As shown in Tab. 9, directly using raw visual features in the auxiliary branch performs even worse than the World-State-Branch-only baseline. This result indicates that the benefit of our dual-branch design does not simply arise from introducing an additional visual representation or increasing the predictive capacity of the model. Raw visual features do not explicitly isolate the motion-related components that are complementary to the deblurred world states, and therefore introduce redundant or interfering information into world-state prediction. Adjacent-frame feature difference also performs noticeably worse than the World-State-Branch-only baseline, suggesting that inter-frame differences may entangle hand-object motion with camera-induced appearance changes, semantic variations, and temporal misalignment. Optical flow provides a more explicit geometric motion cue and is therefore more competitive. Nonetheless, it still remains inferior to deblurring discrepancy, as optical flow primarily captures inter-frame motion and does not explicitly isolate the within-exposure dynamics encoded by motion blur, while also being susceptible to rapid motion, occlusion, and large viewpoint changes.

In contrast, deblurring discrepancy explicitly measures the feature changes induced by removing motion blur within the same frame. Such a formulation suppresses appearance information already represented by the deblurred world state while emphasizing restoration-sensitive components associated with the within-exposure motion. It also avoids the cross-frame misalignment inherent in inter-frame motion representations. Consequently, deblurring discrepancy achieves the best performance across both scenarios, supporting that its benefit comes from providing a more complementary and dynamics-sensitive representation rather than merely from an additional feature branch.

A.8 Controlled Synthetic-Blur Validation of Deblurring Discrepancy

To directly demonstrate whether deblurring discrepancy responds to motion-induced blur, we conduct a controlled experiment on the Moving-Object Grasping task of our **EgoDyn**. Since ground-truth sharp frames are unavailable, we first apply the NAFNet deblurring model to each captured frame and use the resulting deblurred image as a fixed reference. Starting from this deblurred reference, we simulate a centered exposure by averaging 16 translated copies of the same image. We vary the prescribed total displacement over $d \in \{0, 2, 4, 6, 8, 10, 12\}$ pixels and the translation direction

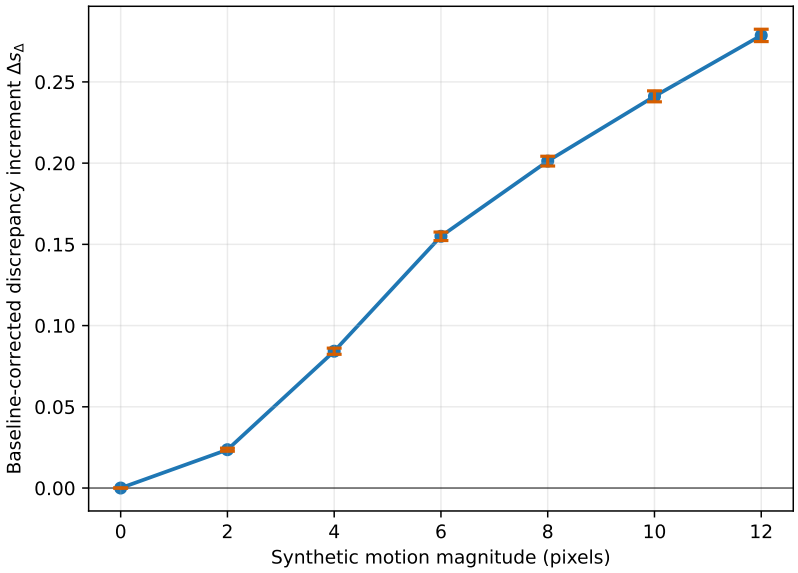


Figure 7: Controlled synthetic-blur validation after paired zero-motion baseline correction. The mean deblurring-discrepancy increment increases with the prescribed within-exposure displacement. For each source image, its direction-averaged discrepancy at $d = 0$ is subtracted from the discrepancy measured at each displacement magnitude. Error bars indicate 95% confidence intervals over the paired per-image increments.

over eight angles at 45° intervals. Thus, the reference image content and scene semantics remain fixed while only the additionally introduced synthetic displacement is varied.

We process each synthetic observation with the same NAFNet deblurring model and DINOv2 encoder used in our **DYNAMICHAND** pipeline. For a blurred image $B_{d,\theta}$ and its deblurred result $\hat{I}_{d,\theta}$, we measure the discrepancy strength as

$$s_\Delta(I, d, \theta) = \frac{1}{NC} \left\| \Phi(B_{d,\theta}) - \Phi(\hat{I}_{d,\theta}) \right\|_1, \tag{17}$$

where N and C are the numbers of patch tokens and feature channels. For each source image and displacement magnitude, we first average $s_\Delta(I, d, \theta)$ over the eight directions, yielding $\bar{s}_\Delta(I, d)$. Because reapplying the deblurring operator can alter the reference image even when $d = 0$, we remove this image-specific restoration response using a paired zero-motion baseline:

$$\Delta s_\Delta(I, d) = \bar{s}_\Delta(I, d) - \bar{s}_\Delta(I, 0). \tag{18}$$

The resulting quantity measures the discrepancy increment associated with the additionally introduced controlled blur. Pearson and Spearman correlations (Pearson, 1895; Spearman, 2010) are computed over all image-magnitude pairs $\{(d, \Delta s_\Delta(I, d))\}$.

As shown in Fig. 7, the baseline-corrected discrepancy increment increases consistently with the prescribed displacement. Before baseline correction, the raw ℓ_1 discrepancy exhibits a Pearson correlation of 0.790 and a Spearman rank correlation of 0.804 with d. After subtracting the paired zero-motion response, these correlations increase substantially to 0.918 and 0.927, respectively, over all the image-magnitude pairs. Importantly, the image content and scene semantics are identical across different displacement levels, with the prescribed within-exposure translation being the only systematically varied factor. The stronger correlations after paired baseline correction indicate that removing the image-specific zero-motion restoration response isolates a discrepancy component that is more closely associated with the controlled motion magnitude. Therefore, the observed monotonic relationship provides direct representation-level evidence supporting our interpretation of deblurring discrepancy as a proxy for blur-induced within-exposure dynamics, rather than merely an arbitrary feature residual introduced by image restoration.

Table 10: Subject-disjoint evaluation on the One-Hand Tossing scenario of **EgoDyn**. "Original" denotes the test set in Tab. 2, while "Subject-Disjoint" contains 2,038 clips from two entirely unseen participants. Δ denotes the relative error increase.

Approach	Original Test		Subject-Disjoint Test		Relative Increase	
	ADE ↓	FDE ↓	ADE ↓	FDE ↓	ΔADE ↓	ΔFDE ↓
CVM (Schöller et al.)	0.177	0.312	0.198	0.348	11.9%	11.5%
HandsOnVLM (Bao et al.)	0.195	0.216	0.220	0.243	12.8%	12.5%
EMAG (Hatano et al.)	0.103	0.176	0.119	0.201	15.5%	14.2%
USST (Bao et al.)	0.071	0.104	0.126	0.241	77.5%	131.7%
S-Mamba (Wang et al.)	0.054	0.077	0.068	0.099	25.9%	28.6%
EgoMAN (Chen et al.)	0.063	0.101	0.072	0.114	14.3%	12.9%
Uni-Hand (Ma et al.)	0.055	0.072	0.062	0.082	12.7%	13.9%
DYNAMICHAND (DACM)	0.053	0.067	0.057	0.073	7.5%	9.0%
DYNAMICHAND (Full)	**0.049**	**0.065**	**0.052**	**0.069**	**6.1%**	**6.2%**

Table 11: Comparison between direct trajectory supervision and discrepancy-aware consistency distillation.

Approach	ADE ↓	FDE ↓
Direct Student	0.090	0.088
DYNAMICHAND (DACM)	**0.072**	**0.057**

A.9 SUBJECT-DISJOINT EVALUATION ON EGODYN

Although the original **EgoDyn** test set already contains clips from participants unseen during training, it is not fully subject-disjoint because some participant identities remain shared across the training and test sets. We therefore conduct a more stringent evaluation with two new participants who do not appear in any original **EgoDyn** split. These participants provide 200 additional One-Hand Tossing videos, from which 2,038 clips are extracted following the same three-frame observation and six-frame prediction protocol. All methods retain the models trained on the original **EgoDyn** training set and are evaluated directly on the new subjects without fine-tuning or test-time adaptation.

As shown in Tab. 10, **DYNAMICHAND** (Full) achieves the best subject-disjoint performance, with an ADE of 0.052 and an FDE of 0.069. Moreover, **DYNAMICHAND** (Full) exhibits the smallest relative error increases of 6.1%/6.2%, followed by **DYNAMICHAND** (DACM) with 7.5%/9.0%. In contrast, existing baselines undergo increases of 11.9-77.5% in ADE and 11.5-131.7% in FDE. These results indicate that discrepancy-guided world dynamics transfer more reliably to unseen subjects, reducing reliance on subject-specific appearance and motion patterns.

A.10 COMPARISON WITH DIRECT STUDENT

To determine whether the performance of DACM originates from teacher-based dynamics transfer rather than the lightweight student architecture itself, we construct a *Direct Student* baseline. We conduct this controlled ablation on the High-Five task of **EgoDyn**, which contains rapid hand motion and abrupt trajectory changes. The Direct Student uses exactly the same raw-RGB encoder, student architecture, one-step inference procedure, optimizer, and training schedule as DACM, but is trained without the frozen teacher. Specifically, we remove the teacher-dependent latent distillation loss $\mathcal{L}_{\text{latent}}$, consistency loss $\mathcal{L}_{\text{cons}}$, auxiliary motion-latent loss $\mathcal{L}_{\text{mot}}$, and discrepancy-aware weight ω. Its training objective therefore contains only trajectory-level supervision.

As shown in Tab. 11, DACM reduces ADE from 0.090 to 0.072 and FDE from 0.088 to 0.057, corresponding to relative reductions of 20.0% and 35.2%, respectively. The larger improvement in FDE indicates that teacher-guided dynamics transfer is particularly beneficial for accurately predicting trajectory endpoints under the rapid and abrupt hand-motion changes of High-Five interactions. Since the two variants share the same student architecture, raw-RGB observations, one-step inference procedure, and training configuration, these results demonstrate that the performance gain cannot be attributed solely to model capacity or inference cost. Instead, they provide evidence that the combined latent distillation, cross-timestep consistency learning, and discrepancy-aware supervision in DACM effectively transfer the teacher's discrepancy-guided dynamics into the lightweight student.

Table 12: Comparison of visual input configurations on **EgoDyn**. **DYNAMICHAND** (Full) accesses both views through its discrepancy-guided framework.

Approach	Visual Input	Handshaking		Human-Robot Handover	
		ADE ↓	FDE ↓	ADE ↓	FDE ↓
EMAG (Hatano et al.)	Raw	0.084	0.126	0.068	0.089
	Deblurred	0.081	0.123	0.064	0.091
	Raw + Deblurred	0.081	0.124	0.062	0.092
USST (Bao et al.)	Raw	0.070	0.101	0.051	0.074
	Deblurred	0.066	0.097	0.046	0.066
	Raw + Deblurred	0.066	0.097	0.048	0.067
S-Mamba (Wang et al.)	Raw	0.057	0.075	0.051	0.062
	Deblurred	0.058	0.074	0.044	0.058
	Raw + Deblurred	0.054	0.076	0.047	0.062
EgoMAN (Chen et al.)	Raw	0.067	0.100	0.046	0.073
	Deblurred	0.063	0.088	0.046	0.065
	Raw + Deblurred	0.063	0.088	0.046	0.065
Uni-Hand (Ma et al.)	Raw	0.061	0.070	0.043	0.045
	Deblurred	0.054	0.069	0.038	0.047
	Raw + Deblurred	0.053	0.067	0.039	0.046
DYNAMICHAND (Full)	Raw + Deblurred	**0.046**	**0.060**	**0.031**	**0.039**

A.11 EFFECT OF RAW AND DEBLURRED INPUTS ON BASELINE PERFORMANCE

To examine whether our comparisons depend on supplying deblurred observations to competing methods, we evaluate three input configurations on the original Handshaking and Human-Robot Handover splits of **EgoDyn**, i.e., raw observations, deblurred observations, and raw-deblurred feature fusion. For the fusion variant, both views are processed by the same visual encoder, and their aligned features are concatenated along the channel dimension and linearly projected back to the baseline's original feature dimension. The projection is trained jointly with the trainable baseline components. Each variant is trained and evaluated using its corresponding input configuration, with identical data splits, observation horizons, hand-coordinate inputs, supervision, and training budgets. Checkpoints are selected using the same validation protocol. CVM is excluded because it does not consume visual inputs.

Tab. 12 shows that deblurring generally improves baseline performance, while concatenating raw and deblurred features yields limited and inconsistent additional gains. **DYNAMICHAND** (Full) achieves the lowest errors across both scenarios, even when competing methods retain raw observations or access both visual views. Its advantage therefore does not depend on restricting baselines to deblurred inputs. Together with the raw-feature control in Sec. A.7, these results support the value of explicitly isolating restoration-induced feature changes and forecasting their evolution to provide complementary dynamics for world-state prediction.

A.12 SCOPE AND LIMITATIONS OF WORLD-DYNAMICS MODELING

Our use of *world dynamics* follows the task-oriented latent predictive perspective of world modeling, rather than explicit physical system identification. In particular, **DYNAMICHAND** does not aim to recover a complete physical state, dense motion field, or physically identifiable governing equations from egocentric observations. Instead, it learns task-relevant latent dynamics by combining predictive representation learning with deblurring discrepancy, which provides an indirect but informative cue about within-exposure relative motion among the hands, objects, and camera. Therefore, the learned dynamics should be interpreted as predictive latent structure that is useful for future hand trajectory forecasting, rather than as a physically identifiable model of the environment.

This formulation deliberately prioritizes predictive utility over explicit physical reconstruction. Our controlled synthetic-blur analysis and comparisons with alternative motion representations provide evidence that deblurring discrepancy captures motion-sensitive information beyond conventional appearance features, while the consistent gains across datasets and dynamic interaction scenarios demonstrate that such latent dynamics are useful for EHTP. A natural direction for future work is to

further connect these learned latent dynamics with explicit geometric or physical variables, and to extend the framework toward action-conditioned prediction and dynamic robotic manipulation.