EGOWAM: GEOMETRY-GROUNDED WORLD ACTION MODELS FOR EGOCENTRIC INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models are typically learned in image-space state representations, such as pixels or 2D features. In egocentric hand-object interaction, however, such view-dependent states are heavily perturbed by arbitrary head motion, causing learned dynamics to entangle physical interactions with incidental viewpoint shifts. We present EgoWAM, a world action model that learns geometry-grounded dynamics in a 3D semantic field. This field represents the world by anchoring dense semantic features to scene geometry to reduce the influence of viewpoint changes. EgoWAM autoregressively predicts future 3D semantic fields and hand motions, with a diffusion-based action head that models uncertainty in future hand motions and provides a context-aware prior for sampling candidate actions during planning. We then score candidate actions using a spatial-semantic optimal transport objective, which establishes soft correspondences between predicted and goal states by aligning semantic features under a 3D transport cost. Experiments on both rigid and articulated egocentric hand-object interaction datasets show improved performance in world-state prediction, motion forecasting, and goal-directed planning.

1 INTRODUCTION

World models provide agents with internal simulators for reasoning about consequences before acting. They have been explored in robot manipulation (Yang et al., 2024b; Hafner et al., 2025), autonomous driving (Gao et al., 2024; Wang et al., 2024b), and video generation (Bruce et al., 2024; Agarwal et al., 2025). For embodied interaction, a world model should not only predict accurate future states, but also support sampling context-aware actions toward goal-directed trajectories. In this work, we focus on egocentric hand-object interaction, where a world model is expected to sample plausible hand motions from first-person observations and predicts their effects on nearby objects.

A key design choice is the state representation. Existing world models mostly predict image-space states, such as RGB pixels (Bar et al., 2025; Bai et al., 2025) or 2D visual features (Zhou et al., 2025; Baldassarre et al., 2025). This is reasonable for static-camera setups where pixel changes largely reflect the underlying scene motion. In egocentric cameras, however, visual changes are caused by both human-object interactions and incidental head movements. As a result, the same physical interaction state may appear very different across views, while visually similar images may correspond to different 3D scene configurations. Consequently, dynamics represented in image-space states inherently entangle true interaction dynamics with arbitrary viewpoint shifts, limiting the world model's ability to capture and reason about the true physical evolution of the scene.

We argue that egocentric world-action modeling should be grounded in the geometry of interaction rather than the image plane of a moving camera. To this end, we represent the world as a 3D semantic field, lifting visual semantics from individual views onto a shared 3D coordinate frame. The resulting representation factors out camera motion while capturing the spatial structure relevant to action-induced changes. Recent visual foundation models (Siméoni et al., 2025; Wang et al., 2025a) make such fields readily obtainable from visual observations at scale. World-action modeling can then be formulated directly in this space: learning how actions transform a continuous semantic-geometric field, and reasoning over those transformations to select actions toward desired future states.

Building on this formulation, we introduce EgoWAM, a geometry-grounded world action model for egocentric hand-object interaction. To ground actions in the same 3D space as the world state, EgoWAM represents hand motion with MANO (Romero et al., 2022) and introduces a hand-scene

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.

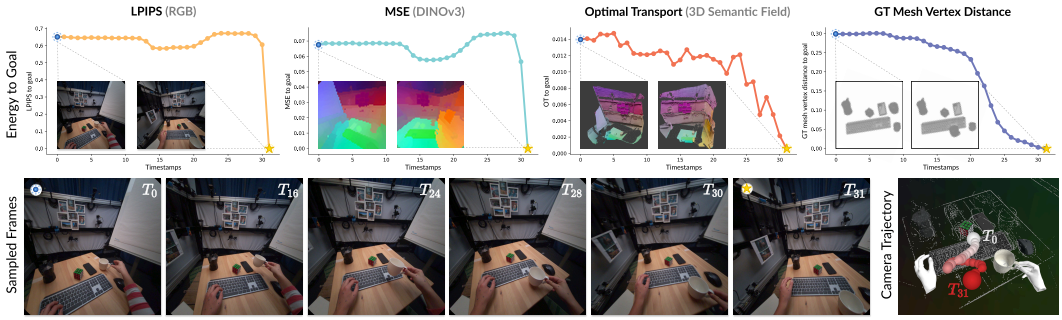


Figure 1: **Progress-to-goal energy curves** computed in four representation spaces on an egocentric HOI sequence. The camera trajectory reveals substantial arbitrary ego-motion, making the sampled frames wobble despite gradual hand-object state changes. Image LPIPS and DINO MSE remain largely uninformative and change sharply only near the final frames when the view becomes visually similar to the goal. In contrast, our spatial-semantic optimal transport energy in 3D semantic field better matches the overall decreasing trend of the GT mesh transport distance.

interaction field that explicitly encodes the local geometric relationship between the hand surface and the surrounding scene. EgoWAM learns autoregressive dynamics jointly over hand motion, interaction field, and 3D semantic field. At each timestep, the autoregressive context conditions a lightweight diffusion head that models the next hand motion delta as a distribution. A sampled delta updates the action, which in turn conditions the prediction of subsequent world-state tokens.

For model-based planning, candidate hand motions are sampled from the learned diffusion prior rather than the full MANO parameter space, and rolled out through EgoWAM to predict consequent 3D semantic fields. Selecting actions toward a goal requires measuring the discrepancy between predicted and target states, but viewpoint changes produce different samplings of the scene surface, making explicit point correspondence unavailable. We therefore introduce a spatial-semantic optimal transport energy that establishes soft, spatially coherent correspondences based on semantic similarity and 3D displacement, providing a viewpoint-robust measure of interaction progress (Fig. 1).

Experiments on HOT3D (Banerjee et al., 2025) and ARCTIC (Fan et al., 2023) demonstrate the effectiveness of EgoWAM across rigid and articulated hand-object interactions. Compared with action-conditioned world models and joint world-action models, EgoWAM achieves the highest retrieval precision for world-state prediction, and the lowest hand joint position and root translation errors on both datasets, showing that geometry-grounded modeling benefits both scene prediction and hand-motion generation. Ablation studies further support the complementary roles of geometry, semantics, and local interaction modeling, and show that the diffusion prior and spatial-semantic transport energy improve planning accuracy and efficiency.

Our contributions are threefold. First, we formulate egocentric world-action modeling in 3D semantic fields to reduce the influence of arbitrary camera motion on interaction modeling. Second, we introduce EgoWAM, an autoregressive prediction model over continuous hand actions and 3D semantic world states, using a diffusion action head to model motion uncertainty and provide a context-aware action prior for planning. Third, we propose a spatial-semantic optimal transport energy that establishes soft correspondences between differently sampled scene surfaces, measuring goal discrepancy through semantic similarity and 3D displacement for goal-directed control.

2 Related Works

World Models. World models learn predictive dynamics for future simulation and planning. Generative world models predict future observations in RGB space, with supervision on visual generation fidelity (Bar et al., 2025; Agarwal et al., 2025). In contrast, JEPA-style models predict in latent feature spaces, either using pretrained representations such as DINO (Zhou et al., 2025; Baldassarre et al., 2025) or jointly learning video representations and predictions from scratch (Assran et al., 2025; Balestriero & LeCun, 2025). More recently, world action models further incorporate actions into the prediction target, making the learned dynamics more directly applicable to control and planning. This

formulation has also been explored in both image-space models (Li et al., 2025; Yuan et al., 2026) and feature-space models (Chen et al., 2025; Wang et al., 2025b).

Visual Representations. Visual representation is central to predictive world modeling, as it determines what state information is preserved for forecasting and reasoning. In 2D, image-level models such as DINO (Caron et al., 2021; Oquab et al., 2024) and MAE (He et al., 2022) learn semantic features from large-scale pretraining, while video-level models (Assran et al., 2025; Tong et al., 2022) further capture temporal structure. For 3D, NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023) provide powerful scene representations, but their reliance on multi-view observations and per-scene optimization limits their scalability for massive predictive training. Recent pointmap-based feed-forward models offer a more scalable alternative. Recent methods (Wang et al., 2025a; Zhang et al., 2025; Xie et al., 2026) demonstrate strong reconstruction quality and generalizability, making pointmap an effective geometry representation for large-scale world modeling.

Egocentric Interaction. Egocentric vision provides a natural agentic setting, where models must understand dynamic visual input and plan actions from the embodied first-person perspective. Large-scale egocentric datasets have enabled progress in this direction, including skilled or daily full-body motion datasets such as EgoExo-4D (Grauman et al., 2024) and Nymeria (Ma et al., 2024), as well as hand-object interaction (HOI) datasets such as HOT3D (Banerjee et al., 2025) and ARCTIC (Fan et al., 2023). These datasets support a wide range of tasks, including egocentric pose estimation (Hong et al., 2025; Li et al., 2023), human action forecasting (Liu et al., 2022; Ashutosh et al., 2025), and egocentric world models (Chen et al., 2025; Bai et al., 2025).

Optimal Transport. Optimal transport (OT) computes a soft correspondence between two weighted point sets by finding how mass should be transported from one set to the other at minimum cost. Although exact optimal transport is computationally expensive, entropic regularization enables efficient approximation with Sinkhorn iterations (Cuturi, 2013). OT has been widely used as a structured loss in vision and robotic learning, including Wasserstein losses for multi-label classification (Frogner et al., 2015), task assignment for multi-robot dense reconstruction (Dong et al., 2019), and imitation learning by matching expert and agent state-action distributions (Dadashi et al., 2021).

3 EgoWAM: World-Action Dynamics in 3D Semantic Field

We introduce EgoWAM for learning and planning egocentric HOI. Our formulation represents the world as a 3D semantic field, decoupling the world state from the motion of the egocentric camera (Sec. 3.1). This representation supports a spatial-semantic OT energy for measuring state discrepancy through soft correspondence, providing the objective for model-based planning (Sec. 3.3). EgoWAM learns autoregressive world-action prediction in this representation space, together with a denoising diffusion prior for context-aware sampling of candidate hand motions (Sec. 3.2).

3.1 Geometry-Grounded State Space for Egocentric Interaction

Egocentric observations entangle interaction dynamics with camera motion. As illustrated in Fig. 1, this makes distances in image or 2D feature space poorly aligned with the underlying interaction: gradual hand-object changes may induce only small differences, while incidental viewpoint changes can dominate them. We therefore ground the world state in 3D geometry, providing a stable coordinate frame in which interaction dynamics can be modeled independently of camera motion.

World state representation. As shown in Fig. 2(left), we represent the world state at timestep t as a dense 3D semantic field that anchors semantic descriptors to metric 3D locations.

$$s_t = (f_t, p_t), \quad f_t \in \mathbb{R}^{H\times W\times C}, \quad p_t \in \mathbb{R}^{H\times W\times 3}. \tag{1}$$

Here, f_t denotes the DINO (Siméoni et al., 2025) feature map, and p_t denotes the 3D pointmap (Wang et al., 2024a). Each feature $f_t(h, w)$ is assigned to a 3D point $p_t(h, w)$ in a common coordinate frame. DINO features provide semantic correspondence across time, while pointmaps provide the metric geometry needed to reason about hand-object proximity and transportation between states.

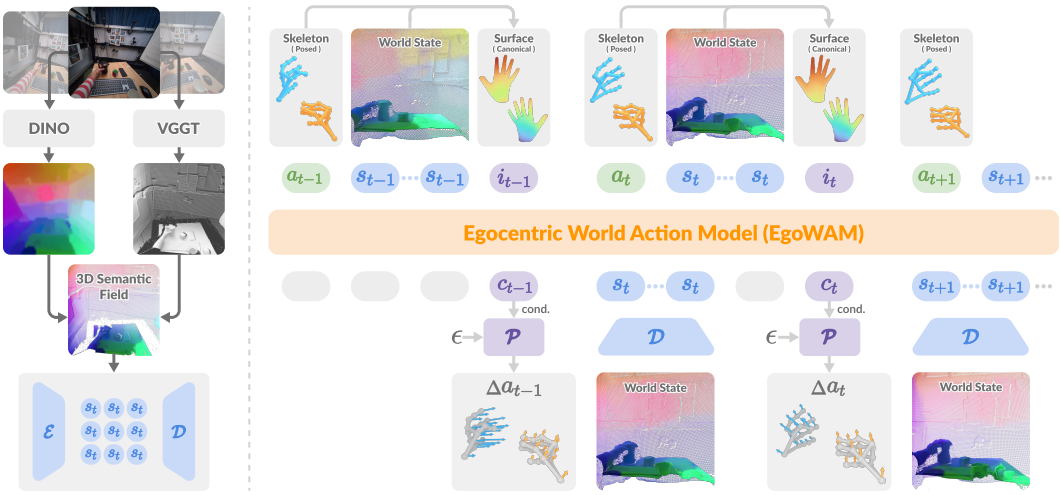


Figure 2: **EgoWAM architecture. (Left)** Given egocentric observations, we construct 3D semantic fields by lifting dense semantic features and scene geometry into a view-stable world-coordinate representation. A VAE further compresses this field into tokens s. **(Right)** EgoWAM models the joint evolution of world states and hand actions using a causal Transformer. Given hand pose a_{t-1} and world state s_{t-1}, we first convert the skeletal pose into a surface mesh and compute an interaction field i_{t-1} that captures hand-scene local geometric relationship. This field, alongside the history of actions and states, is processed into a context token that conditions a diffusion network $\mathcal{P}$, which samples a pose increment Δa_{t-1} to update the hand pose to a_t. The model then predicts the consequent world state s_t conditioned on the updated action and the preceding sequence.

Action representation. We represent hand motion with MANO (Romero et al., 2022), a compact and physically consistent parameterization of articulated 3D hand geometry. At each time step t, the hand state is $a_t = (\boldsymbol{\theta}_t, \mathbf{R}_t, \mathbf{t}_t)$, where $\boldsymbol{\theta}_t$ denotes the articulated pose, and $(\mathbf{R}_t, \mathbf{t}_t)$ is the global rotation and translation. Together with a fixed hand shape parameter $\boldsymbol{\beta}$, these parameters determine the 3D hand surface used to model hand-scene interaction.

Hand-scene interaction field. While MANO provides the hand surface, its spatial relationship with the scene is not explicitly represented. We therefore introduce a hand-scene interaction field that captures local scene geometry around the hand surface. Given hand vertices $\{\mathbf{v}_k\}_{k=1}^{V}$ and scene points $\{\mathbf{p}_i\}_{i=1}^{H\times W}$, we compute their proximity using a Gaussian kernel $\alpha_{ik} = w_i \exp(-\|\mathbf{p}_i - \mathbf{v}_k\|^2/2\sigma^2)$, where w_i downweights hand and background regions. Normalizing the weights as $\tilde{\alpha}_{ik} = \alpha_{ik}/\sum_j \alpha_{jk}$, we define an interaction feature $\mathbf{h}_k = (m_k, d_k, \mathbf{u}_k)$ at each hand surface vertex:

$$m_k = \sum_i \alpha_{ik}, \quad d_k = \sum_i \tilde{\alpha}_{ik}\|\mathbf{p}_i - \mathbf{v}_k\|, \quad \mathbf{u}_k = \sum_i \tilde{\alpha}_{ik} \frac{\mathbf{p}_i - \mathbf{v}_k}{\|\mathbf{p}_i - \mathbf{v}_k\|}. \tag{2}$$

Here, m_k captures how much scene geometry is nearby, d_k measures the average distance, and $\mathbf{u}_k$ encodes the dominant direction. Field $i_t = \{\mathbf{h}_k^t\}_{k=1}^{V}$ explicitly couples the hand action with the local 3D scene geometry, providing a geometry-grounded representation for learning interaction dynamics.

3.2 Autoregressive World–Action Prediction with Uncertainty

Building upon our geometry-grounded world-action representations, we learn a causal dynamics model that jointly predicts how the hands move and how the world state changes. Egocentric hand-object interaction has an inherently interleaved causal structure: the current world state constrains plausible hand motion, while the resulting hand motion changes the future world state. We therefore model future interaction as an interleaved autoregressive process:

$$p(a_{1:T}, s_{1:T} \mid a_0, s_0) = \prod_{t=0}^{T-1} p(\Delta a_t \mid a_{\le t}, s_{\le t}, i_{\le t})\, p(s_{t+1} \mid a_{\le t+1}, s_{\le t}, i_{\le t}), \tag{3}$$

where Δa_t is the relative MANO motion that updates the hand state from a_t to a_{t+1}.

Multimodal autoregressive architecture. We instantiate EgoWAM as a decoder-only Transformer. The backbone adopts a language-model backbone (InternLM, 2023) with designs like causal attention and RoPE. Different from language models over discrete word indices, we represent all tokens as continuous embeddings. This design preserves the inherently continuous nature of hand kinematics and scene evolution, thereby avoiding the information loss typically introduced by token discretization.

EgoWAM models hand-object dynamics over short temporal chunks, where each step corresponds to L consecutive frames. We train a spatio-temporal VAE to compress each world-state chunk into compact tokens. We then use separate lightweight MLP adapters to project the three modalities (hand motion, world state, and interaction field) into a shared Transformer hidden space. For notational simplicity, we do not introduce separate symbols for the encoded or adapted representations, and use a_t, s_t, and i_t to denote the corresponding tokens used by the Transformer. As illustrated in Fig. 2, we arrange these tokens in an interleaved causal sequence $[a_0,\ s_0,\ i_0,\ a_1,\ s_1,\ i_1,\ \ldots]$.

Modeling action uncertainty. Given the same observed hand-object state, multiple future motion sequences can be plausible; likewise, the same goal may be reached through different motions. A deterministic regressor tends to average these alternatives, often producing a motion that matches none of them. We therefore model relative MANO updates as a conditional distribution, from which diverse and plausible action trajectories can be sampled.

To parameterize this conditional action distribution, the causal Transformer summarizes the rollout history of hand states, world states, and interaction fields into a context token c_t at each step. Conditioned on c_t, a lightweight denoising diffusion network models the next MANO delta as $\Delta a_t \sim \mathcal{P}_\psi(\Delta a_t \mid c_t)$, $\Delta a_t = a_{t+1} - a_t$. Following Diffusion Loss (Li et al., 2024), we train this conditional action model by corrupting the GT with Gaussian noise and predicting the injected noise:

$$x_k = \sqrt{\bar{\alpha}_k}\Delta a_t + \sqrt{1-\bar{\alpha}_k}\epsilon, \quad \mathcal{L}_{\text{act}} = \mathbb{E}_{t,k,\epsilon}\left[\|\epsilon - \epsilon_\psi(x_k, k, c_t)\|_2^2\right], \tag{4}$$

where k denotes the diffusion timestep and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. At inference time, the reverse diffusion process samples plausible action deltas from $\mathcal{P}_\psi(\Delta a_t \mid c_t)$, which are subsequently composed with the current MANO state to advance the rollout.

Training objectives. We train EgoWAM to predict the next hand motion and the consequent world state simultaneously. For action prediction, we apply the Diffusion Loss in Eq. 4. For world prediction, we supervise the predicted world-state tokens with MSE. The autoregressive backbone and the diffusion network are jointly trained by backpropagation. The total loss is:

$$\mathcal{L} = \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \lambda_{\text{state}} \sum_t \|\hat{s}_{t+1} - s_{t+1}\|_2^2 . \tag{5}$$

We optimize this objective using both teacher-forcing and rollout training, following standard autoregressive practice. Teacher forcing learns one-step transition under ground-truth histories, whereas rollout training exposes the model to its own predictions to correct accumulated errors.

3.3 Goal-Directed Planning with Spatial-Semantic Transport Energy

The preceding sections describe how EgoWAM predicts the consequences of hand motion, here we use these predictions for goal-directed control. At test time, candidate action trajectories are rolled out with the frozen world model and scored against a goal (Zhou et al., 2025; Bar et al., 2025). This comparison is challenging under viewpoint changes and without pointwise correspondence: image-space or 2D feature distances poorly capture 3D displacement (Fig. 1). We therefore introduce a spatial-semantic transport energy over 3D semantic fields to guide diffusion-based action search.

Action planning at test time. Given the current 3D semantic field s_0, hand state a_0, and a goal field s^*, we optimize a trajectory of MANO deltas $A = \Delta a_{0:H-1}$ over a planning horizon H. Rolling A through EgoWAM produces fields $\hat{s}_{1:H}(A)$, and planning minimizes the terminal discrepancy

$$A^* = \arg\min_A \mathcal{E}_{\text{OT}}\big(\hat{s}_H(A), s^*\big). \tag{6}$$

where $\mathcal{E}_{\text{OT}}$ is the spatial-semantic transport energy defined below.

We solve Eq. 6 within a model predictive control (MPC) framework, using a population-based search inspired by Yang et al. (2024a). At each planning step, candidate trajectories are sampled from

EgoWAM's action prior and scored using this energy. We keep the best candidates, add a moderate amount of diffusion noise to them, and denoise again to obtain nearby but plausible variations. Repeating this process gradually focuses the search on motions that better reach the goal. The diffusion model remains frozen, and the goal guides planning only through candidate scoring and selection. After the search, we execute the first action of the best trajectory and then replan from the new world state. Appendix A.2 provides additional details and pseudocode.

Spatial-semantic transport energy. Although the predicted and goal fields share a common world coordinate frame, their surfaces are independently sampled and therefore lack pointwise correspondence. A hard correspondence induced by global rigid alignment is also inappropriate: an interaction may move multiple rigid bodies or articulated parts independently, and thus cannot be explained by a single $\mathrm{Sim}(3)$ transformation. We therefore compare the two fields through a soft correspondence that jointly accounts for semantic identity and 3D displacement.

Let the predicted field be $\hat{s} = \{(\hat{\mathbf{p}}_i, \hat{\mathbf{f}}_i)\}_{i=1}^{N}$ and the goal field be $s^* = \{(\mathbf{p}_j^*, \mathbf{f}_j^*)\}_{j=1}^{M}$, where $\mathbf{p} \in \mathbb{R}^3$ is a 3D point and $\mathbf{f}$ is its visual semantic feature. To avoid the viewpoint- and depth-dependent bias of uniform masses on image-grid pointmaps, we construct each mass distribution by weighting points according to their local 3D surface area. Specifically, each 2×2 cell in the pointmap is triangulated into 3D triangles, and each triangle's area is uniformly distributed to its vertices. Triangles spanning large depth discontinuities are discarded to avoid assigning mass across invalid surface boundaries. The resulting masses are normalized into probability distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. We define the set of feasible balanced transport plans as $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \left\{\mathbf{P} \in \mathbb{R}_+^{N \times M} \;\middle|\; \mathbf{P}\mathbf{1} = \boldsymbol{\mu},\ \mathbf{P}^\top \mathbf{1} = \boldsymbol{\nu}\right\}$, where each entry P_{ij} represents how much mass is transported from predicted point i to goal point j. The spatial-semantic transport energy is the entropically regularized optimal transport cost

$$\mathcal{E}_{\mathrm{OT}}(\hat{s}, s^*) = \min_{\mathbf{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \sum_{i,j} P_{ij}\, C_{ij} + \varepsilon \sum_{i,j} P_{ij}(\log P_{ij} - 1), \tag{7}$$

with the pairwise cost C_{ij} defined as

$$C_{ij} = \lambda_p \|\hat{\mathbf{p}}_i - \mathbf{p}_j^*\|_2^2 + \lambda_f (1 - \langle \hat{\mathbf{f}}_i, \mathbf{f}_j^* \rangle). \tag{8}$$

The first term measures spatial displacement, the second term penalizes semantic mismatch, with λ_p and λ_f controlling their trade-off. The entropic regularization coefficient ε encourages smooth correspondences and enables efficient computation with Sinkhorn iterations (Cuturi, 2013). As shown in Fig. 1, this energy tracks goal progress more consistently than LPIPS and DINO MSE and follows the GT mesh transport distance more closely.

4 Experiments and Results

4.1 Experiment Setting

Datasets. We evaluate EgoWAM on HOT3D (Banerjee et al., 2025) and ARCTIC (Fan et al., 2023), two egocentric HOI datasets in which subjects perform structured action sequences on rigid and articulated objects. Both datasets provide object CAD models along with per-timestamp 6DoF poses, MANO parameters, and camera poses. We use these annotations to render foreground pointmaps in the world-coordinate frame. Since GT annotations for the official test sets are not publicly available, we construct our own evaluation splits. For HOT3D, we randomly hold out three subjects for testing. For ARCTIC, we use the official validation split as our test set.

Metrics. For *hand-motion prediction*, we report MPJPE to measure the average 3D joint position error, and MPJVE to measure the temporal consistency of joint velocities. We further use F@15 that measures the fraction of joints predicted within 15mm of the ground truth. Finally, we report root translation error (RTE) to evaluate the accuracy of the predicted global hand trajectory. For *world-state prediction*, we use different evaluation protocols and metrics for baseline comparison (Sec. 4.2) and ablation studies (Sec. 4.3), which we detail in the corresponding subsections.

Implementation details. We train a 100M-parameter EgoWAM as our default configuration to ensure efficient inference, with architectural details provided in Appendix A.3. We investigate the effect of model scale in Sec. 4.3. The chunk length L is set to four and each sample consists of six

Table 1: Quantitative comparison with baselines on HOT3D and ARCTIC for world-state and hand-motion prediction. We evaluate world-state prediction using local temporal retrieval and pixel-space metrics, and hand-motion prediction. Best results are in bold and second-best results are underlined.

Dataset	Model	World State Prediction						Hand Motion Prediction			
		Top-1↑	Top-5↑	mAP↑	ℓ_1-RGB↓	LPIPS↓	PSNR↑	MPJPE↓	MPJVE↓	F@15↑	RTE↓
HOT3D	DINO-WM	0.038	0.290	0.178	0.090	0.299	16.861	547.426	641.861	0.005	530.854
	V-JEPA2-AC	0.073	0.423	0.251	0.074	0.244	18.427	205.780	168.990	0.037	142.443
	NWM	0.083	0.315	0.227	0.175	0.582	13.340	68.653	89.427	0.277	66.789
	Fast-WAM	0.026	0.313	0.206	0.085	0.376	18.175	52.279	10.165	0.509	44.477
	UVA	<u>0.294</u>	**0.881**	<u>0.531</u>	<u>0.056</u>	<u>0.236</u>	<u>20.757</u>	<u>21.267</u>	<u>10.060</u>	<u>0.814</u>	<u>17.426</u>
	EgoWAM	**0.352**	<u>0.870</u>	**0.557**	**0.052**	**0.212**	**21.216**	**18.632**	**7.379**	**0.836**	**14.514**
ARCTIC	DINO-WM	0.055	0.480	0.263	0.152	0.503	12.059	507.304	575.451	0.000	486.780
	V-JEPA2-AC	0.098	0.473	0.284	0.142	0.483	12.337	218.042	155.330	0.031	168.644
	NWM	**0.212**	<u>0.649</u>	<u>0.407</u>	0.081	<u>0.333</u>	16.459	74.679	87.549	0.234	72.337
	Fast-WAM	0.013	0.035	0.108	0.122	0.503	15.128	49.982	<u>9.759</u>	0.559	41.075
	UVA	0.014	0.537	0.215	<u>0.062</u>	0.343	<u>18.970</u>	<u>28.425</u>	13.942	**0.733**	<u>23.499</u>
	EgoWAM	<u>0.193</u>	**0.798**	**0.426**	**0.050**	**0.227**	**20.168**	**28.360**	**8.926**	<u>0.712</u>	**22.785**

chunks. For the hand–scene interaction field, we identify hand regions from semantic features and downweight them from the scene geometry. We downsample the MANO mesh following Fan et al. (2023) to reduce the computational cost while preserving the hand structure. We set $\lambda_{\text{act}} = \lambda_{\text{state}} = 1$ in Eq. 5 and $\lambda_p = 0.5, \lambda_f = 0.5$ in Eq. 8. We train EgoWAM with batch size of 64/48 for teacher-forcing/rollout fine-tuning, 200k/50k iterations on HOT3D and 100k/20k iterations on ARCTIC, respectively. Rollout fine-tuning is performed over a randomly sampled chunk at each iteration.

4.2 Comparison with Baselines

Baselines. We compare EgoWAM with two groups of baselines. First, we consider action-conditioned world models, including DINO-WM (Zhou et al., 2025), V-JEPA2-AC (Assran et al., 2025), and NWM (Bar et al., 2025), which predict future states conditioned on given actions. During planning, their candidate actions are sampled from a context-independent Gaussian proposal. Second, we compare with world-action models, including UVA (Li et al., 2025) and Fast-WAM (Yuan et al., 2026), which jointly predict future world states and actions. For a fair comparison, all baselines are trained from scratch on HOT3D and ARCTIC using the same training schedule as EgoWAM, with the exception of Fast-WAM, which employs a pretrained 5B-parameter video DiT.

World-state evaluation protocol. Our baselines predict future states in heterogeneous spaces, including DINO features, V-JEPA features, and RGB pixels, whereas EgoWAM operates on 3D semantic fields, making prediction errors across representations not directly comparable. We therefore introduce two complementary evaluation protocols.

Local temporal retrieval. We cast next-state prediction as a local temporal retrieval problem. Given an initial state and ground-truth actions, each model predicts the next state in its native representation space. We construct a 20-frame candidate window containing the target frame, 9 preceding frames, and 10 subsequent frames, and rank the candidates by their distance to the prediction in the corresponding representation space. We report *Top-1* and *Top-5* accuracy, indicating whether the target is ranked first or within the top five candidates, respectively, together with mean average precision (*mAP*) to characterize ranking quality. This protocol provides a representation-agnostic measure of temporal prediction accuracy based on relative ranking rather than raw distances.

Pixel-space probing. As a complementary evaluation, we decode predicted representations into a common RGB space, enabling direct comparison with GT videos. Specifically, we train V-RAE (Guo et al., 2026), which constructs generative latent spaces from frozen visual representations, for DINO-WM, V-JEPA, and EgoWAM to decode their predictions into videos. NWM, Fast-WAM, and UVA already generate videos with their native video-generation models and therefore require no additional decoder. We evaluate all predicted videos against GT using ℓ_1 error, LPIPS, and PSNR. Note that this experiment only serves as a common pixel-space probe for cross-model comparison.

Table 2: Ablation study of EgoWAM. Removing each component generally degrades state or action prediction. Dashes denote metrics that are not applicable when the modality is removed.

Ablation	World State Prediction				Hand Motion Prediction			
	MSE ↓	Cosine Sim. ↑	3D Err. ↓	Interp. Depth ↓	MPJPE ↓	MPJVE ↓	F@15 ↑	RTE ↓
Full model	**0.010**	**0.926**	**26.087**	**17.015**	**18.632**	7.379	0.836	**14.514**
w/o Diffusion Loss	0.012	0.906	36.398	20.359	19.427	**6.123**	0.836	15.127
w/o Interaction field	0.013	0.896	37.660	18.284	19.450	9.694	**0.847**	15.673
w/o 3D pointmap	0.012	0.905	—	—	23.173	15.146	0.839	19.626
w/o DINO feature	—	—	52.874	20.896	20.421	7.986	0.824	16.033

Action evaluation protocol. We evaluate action prediction by comparing predicted actions against recorded human motions. For action-conditioned world models, we provide L frames of observation history and use future L frames as the goal state. We then use CEM to plan an action sequence toward this goal. For world-action models, we directly use their predicted action sequences. In both cases, the recorded human motion serves as the GT action sequence, against which we measure prediction errors introduced in 4.1. Since planning is performed in the high-dimensional MANO parameter space, CEM-based optimization is computationally expensive, requiring approximately 15 minutes per sequence to converge. We therefore evaluate all methods on the same set of 100 randomly sampled sequences, with their metadata provided in the supplementary material.

Results analysis. Tab. 1 shows that EgoWAM achieves the strongest overall world-state prediction. On HOT3D, it obtains the best Top-1 accuracy and mAP (0.352 and 0.557), with Top-5 accuracy only slightly below UVA (0.870 vs. 0.881). On ARCTIC, it achieves the best Top-5 accuracy and mAP (0.798 and 0.426), while its Top-1 accuracy is second to NWM (0.193 vs. 0.212). The video-space probe yields a consistent advantage: EgoWAM is best on all three video metrics for both datasets, reaching ℓ_1-RGB/LPIPS/PSNR of 0.052/0.212/21.216 on HOT3D and 0.050/0.227/20.168 on ARCTIC. These measurements demonstrate the consistent advantage of our world-state representation across complementary evaluation spaces. EgoWAM also yields strong hand-motion prediction. It ranks first on all four metrics on HOT3D, and on ARCTIC it achieves the best MPJPE, MPJVE, and RTE (28.360, 8.926, and 22.785), with F@15 slightly below UVA (0.712 vs. 0.733).

In contrast, action-conditioned baselines obtain non-trivial state predictions with GT actions but struggle when planning actions, because Gaussian CEM proposals must search the high-dimensional articulated hand-motion space without a learned prior, which makes it hard to discover the feasible hand motions within a limited number of samples. Fast-WAM also shows comparatively weak retrieval performance. Empirically, it captures coarse local hand–object motion but struggles with ego-camera-induced global motion. We hypothesize that this limitation stems from its generation strategy, which initializes future frames by replicating the initial frame, thereby introducing a static-scene bias that hinders the generation of large camera-induced transformations. Overall, geometry-grounded state modeling and a learned action prior jointly support accurate state and action prediction.

4.3 Ablation Study and Scalability Analysis

World-state evaluation metrics. For the following studies, we evaluate EgoWAM's predicted 3D semantic fields using feature-level, spatial, and contact-related metrics. Specifically, we measure MSE and cosine similarity between predicted and GT semantic features, as well as the 3D position error of the predicted pointmap. To assess physical plausibility near hand-object contacts, we additionally report the mean interpenetration depth between the predicted scene geometry and the GT hand mesh, quantifying the extent to which the prediction intrudes into the hand volume.

Ablation study on model designs. Table 2 ablates the key components of EgoWAM on HOT3D, revealing complementary benefits for hand-motion and world-state prediction. Geometry and semantics play distinct roles: removing point geometry increases MPJPE from 18.632 to 23.173 and RTE from 14.514 to 19.626, while removing DINO semantics more than doubles the 3D position error from 26.087 to 52.874. These results suggest that spatial structure constrains motion accuracy, whereas semantic cues are essential for predicting future scene states. The interaction field further links hand motion to scene changes: removing it increases semantic MSE from 0.010 to 0.013 and reduces cosine similarity from 0.926 to 0.896. Interpenetration depth also rises from 17.015 to 18.284,

Table 4: Scaling behavior of EgoWAM. Increasing model size from 100M to 600M consistently improves world-state prediction and most hand-motion prediction metrics.

Dataset	Params	World State Prediction				Hand Motion Prediction			
		MSE ↓	Cosine Sim. ↑	3D Err. ↓	Interp. Depth ↓	MPJPE ↓	MPJVE ↓	F@15 ↑	RTE ↓
HOT3D	100M	0.00965	0.92570	26.087	17.015	18.632	7.379	0.836	14.514
	300M	0.00962	0.92592	25.275	16.936	17.374	6.546	0.847	13.732
	600M	**0.00959**	**0.92618**	**24.262**	**16.669**	**16.646**	**6.509**	**0.858**	**13.247**
ARCTIC	100M	0.00949	0.92945	32.185	6.970	28.360	8.926	0.712	22.785
	300M	0.00896	0.93405	31.951	6.914	19.089	**7.190**	0.807	15.319
	600M	**0.00891**	**0.93411**	**31.893**	**6.884**	**18.613**	7.227	**0.813**	**14.877**

supporting the value of explicit proximity cues for contact-aware prediction. Diffusion-based action modeling benefits future-state prediction as well: without the Diffusion Loss, semantic MSE rises to 0.012 and 3D position error to 36.398. The full model achieves the best results across all world-state metrics, together with the lowest MPJPE and RTE, supporting the joint use of all components.

Table 3: Ablation of planning samplers and energies.

Sampler	Energy	Steps	DINO MSE	3D Err.	MPJPE	OT Cost
Ours	OT	**1926**	0.09527	**0.032**	**23.81**	**0.790**
Ours	OT-sem.	**1926**	**0.09524**	0.034	23.84	0.791
Ours	OT-pnt.	**1926**	0.09529	0.033	23.85	0.792
Gaussian	OT	2450	0.09528	**0.032**	23.86	0.791

Ablation study on planning choices. Tab. 3 compares our diffusion-based local search strategy (Sec. 3.3) with an independent Gaussian sampler and assesses the contribution of each term in the spatial-semantic transport energy (Eq. 8). All variants use identical initialization and compute budgets, with a population size of 4, 2 elites per round, and 2 optimization rounds. We evaluate all variants on the same randomly sampled sequences used in Sec. 4.2. Our sampler with the full spatial-semantic OT energy achieves the lowest MPJPE (23.81) and OT cost (0.790), while matching the Gaussian sampler's 3D error (0.032). Compared with Gaussian sampling under the same energy, it uses 21.4% fewer steps (1,926 vs. 2,450) and yields slightly lower DINO MSE and MPJPE. Among the energy variants, the semantic-only term achieves the lowest DINO MSE (0.09524), but incurs higher 3D error and MPJPE than the full energy. Using only the spatial term also slightly increases both errors. Overall, the full energy provides the best balance across semantic, spatial, and hand-motion accuracy, while our sampling strategy maintains competitive accuracy with fewer steps.

Scaling behavior. Scaling EgoWAM improves multiple aspects of world-action prediction (Tab. 4). Increasing model size from 100M to 600M consistently improves all four world-state metrics on both datasets, indicating gains in semantic fidelity, geometric accuracy, and contact plausibility. These improvements are accompanied by more accurate hand poses and trajectories: MPJPE decreases by 10.7% on HOT3D and 34.4% on ARCTIC, while RTE decreases by 8.7% and 34.7%, respectively. However, the gains depend on the dataset and scaling interval. On HOT3D, every metric improves at each model size. On ARCTIC, most of the improvement occurs between 100M and 300M, scaling further yields smaller gains, with MPJVE slightly increasing from 7.190 to 7.227.

5 Conclusion

We presented EgoWAM, a world action model that grounds egocentric hand-object dynamics and planning in 3D semantic fields. By anchoring visual semantics in a shared 3D coordinate frame, this representation reduces the influence of constant camera motion on both state prediction and goal comparison. EgoWAM interleaves hand-motion generation and world-state prediction autoregressively, using a diffusion action head to model motion uncertainty and provide context-aware action proposals. Our spatial-semantic optimal transport energy measures goal discrepancy through soft correspondences that account for semantic similarity and 3D displacement. Experiments demonstrate improved world-state and hand-motion prediction across rigid and articulated interactions, while planning ablations show that combining the learned action prior with the transport energy reduces terminal state error with fewer search iterations. These results support grounding both predictive dynamics and planning objectives in the geometry of interaction.

AI USE STATEMENT

We used ChatGPT and Codex for language polishing, assistance with selected coding tasks, and parts of the data processing workflow. The AI-assisted coding tasks did not include the implementation of the core method proposed in this paper. We did not use these tools to develop or discuss the research methodology, design experiments, or interpret results. This work does not involve synthetic data generation or mathematical proofs. The authors manually reviewed the AI-assisted code to verify its correctness and visually inspected the processed data to check the outcomes of AI-assisted data processing. We take full responsibility for the final content of this work, including text, code, and other artifacts produced with the assistance of these tools.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. FIction: 4D future interaction prediction from video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17613–17625, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: DINO as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. 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*, pp. 7061–7071, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Lu Chen, Yizhou Wang, Shixiang Tang, Qianhong Ma, Tong He, Wanli Ouyang, Xiaowei Zhou, Hujun Bao, and Sida Peng. EgoAgent: a joint predictive agent model in egocentric worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6970–6980, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. *International Conference on Learning Representations*, 9, 2021.

Siyan Dong, Kai Xu, Qiang Zhou, Andrea Tagliasacchi, Shiqing Xin, Matthias Nießner, and Baoquan Chen. Multi-robot collaborative dense scene reconstruction. *ACM Transactions on Graphics*, 38 (4):1–16, 2019.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in Neural Information Processing Systems*, 28, 2015.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Minghui Guo, Shengqiong Wu, and Hao Fei. V-RAE: Rethinking video latent spaces for generation. *arXiv preprint arXiv:2608.13556*, 2026.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. EgoLM: Multi-modal language model of egocentric motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5344–5354, 2025.

Team InternLM. InternLM: A multilingual language model with progressively enhanced capabilities, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17142–17151, 2023.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

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.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

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. *Transactions on Machine Learning Research Journal*, 2024.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

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.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. Learning real-world action-video dynamics with heterogeneous masked autoregression. *arXiv preprint arXiv:2502.04296*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024b.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, et al. Scal3R: Scalable test-time training for large-scale 3D reconstruction. *arXiv preprint arXiv:2604.08542*, 2026.

Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, and Katerina Fragkiadaki. Diffusion-ES: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following. *arXiv preprint arXiv:2402.06559*, 2024a.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. 2024b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one d4rt at a time. *arXiv preprint arXiv:2512.08924*, 2025.

Gaoyue Zhou, Hengkai Pan, Yann Lecun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, pp. 79115–79135. PMLR, 2025.

A Appendix

A.1 Visualization of 3D Semantic Field and Optimal Transport

In Fig. 3, we visualize each 3D semantic field through complementary semantic, geometric, and mass-based renderings in a sequence with four frames, focusing only on the foreground scene. To focus the analysis on scene dynamics rather than the human hand, we identify hand pixels using a DINO-feature-based classifier and remove them from the foreground visualization and transport computation.

For the semantic component, DINO descriptors are projected onto a shared three-dimensional PCA basis fitted over valid foreground pixels from all displayed frames, and the resulting coordinates are robustly normalized using global percentiles before being rendered as RGB. For the geometric component, we render a coordinate-colorized pointmap by mapping the aligned XYZ coordinates of foreground points to RGB under the same global normalization scheme, so colors remain comparable across time. We also visualize the surface-area mass used in our OT formulation by triangulating each 2×2 pointmap cell, rejecting invalid or depth-discontinuous triangles, distributing triangle area to incident vertices, and normalizing the resulting mass over valid foreground pixels.

Finally, to inspect the transport plan, we overlay sampled OT correspondences between two frames: source pixels are connected either to their mass-weighted barycentric target locations or to their strongest transported target points, with line width, endpoint size, opacity, and color scaled by transported mass.

A.2 Planning with Diffusion-Based Local Search

We describe the planner used in Sec. 3.3. Let c_0 denote the current interaction context and let $A = (\Delta a_0, \ldots, \Delta a_{H-1})$ be a candidate MANO-delta trajectory. Rolling out A with EgoWAM gives a terminal world state $\hat{s}_H(A)$, which is scored by

$$J(A) = \mathcal{E}_{\mathrm{OT}}(\hat{s}_H(A), s^*)\,. \tag{9}$$

The planner seeks a low-cost trajectory; the goal affects the search through this score rather than through the diffusion model itself.

Search procedure. We begin by sampling N trajectories from the learned action prior. After rolling them out with the world model, we retain the M trajectories with the lowest costs. Instead of fitting a Gaussian to these trajectories as in standard CEM, we use them directly as parents for the next round. Each parent is slightly corrupted and then denoised again by the frozen action model, producing a new trajectory near a previously successful one. The new trajectories are scored and selected in the same way, and the process is repeated for a small number of rounds.

Concretely, for an elite trajectory A and a diffusion noise level ℓ, we compute

$$x_\ell = \sqrt{\bar{\alpha}_\ell} A + \sqrt{1 - \bar{\alpha}_\ell}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \qquad A' = \mathrm{Denoise}_\psi(x_\ell, \ell, c_0). \tag{10}$$

The child A' is then evaluated from scratch; denoising does not by itself guarantee a better trajectory. A larger ℓ produces a broader change, while a smaller ℓ stays closer to the parent. We therefore decrease the noise level over search rounds, moving from exploration to local refinement.

Algorithm 1 Planning with diffusion-based local search

Require: Context c_0, goal s^*, population size N, number of elites M, search rounds R
1: Sample N trajectories from the action prior $p_\psi(A \mid c_0)$
2: **for** $r = 1, \ldots, R$ **do**
3: Roll out and score every trajectory using Eq. 9
4: Save the best trajectory found so far
5: Keep the M lowest-cost trajectories as elites
6: **if** $r < R$ **then**
7: Resample parents from the elites
8: Partially noise and denoise each parent using Eq. 10
9: **end if**
10: **end for**
11: **return** the best evaluated trajectory

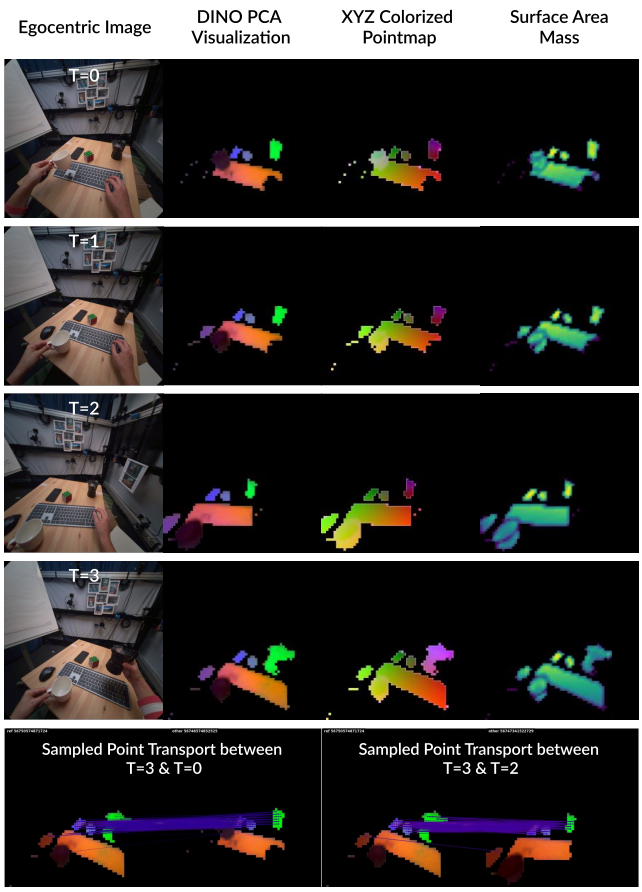


Figure 3: **Visualization of 3D semantic fields and optimal transport.** For each frame, we show the egocentric RGB image, DINO PCA visualization, coordinate-colorized XYZ pointmap, and surface-area mass used by our OT energy. Only foreground regions are visualized; hand pixels are identified by a DINO-feature-based classifier and removed. Bottom: sampled correspondences from the entropic OT plan, with line strength proportional to transported mass.

All noising and denoising are performed in the same normalized MANO-delta representation used during training. Because EgoWAM predicts one action delta at a time, trajectory mutation is also carried out autoregressively: after generating a child delta, we update the predicted history before generating the next one. We return the best evaluated trajectory rather than averaging the elites, and execute only its first action before replanning.

A.3 EgoWAM Architecture and Training Configurations

Architecture configurations. EgoWAM uses a causal transformer as the dynamics model over continuous hand-state and world-state tokens. Hand-state vectors and compact world-state latents are first projected into a shared transformer embedding space. Future hand motion is decoded by a conditional diffusion head, while future world-state latents are predicted by a lightweight latent head.

Following the InternLM-style scaling recipe, the transformer keeps the standard pre-norm block with RMSNorm, rotary position embedding, FlashAttention, and a SwiGLU feed-forward network. Tab. 5 summarizes the neural architecture of the EgoWAM-100M dynamics model and its MANO diffusion decoder.

Tab. 6 reports the lightweight input and prediction modules around the transformer. MANO shape and pose vectors are projected into the transformer dimension by small LayerNorm–MLP adapters.

Table 5: Neural architecture configuration of the EgoWAM autoregressive model.

Component	Configuration	EgoWAM-100M
Causal Transformer for modeling interaction dynamics		
Backbone	transformer blocks	22
Embedding	hidden dimension	640
Self-attention	heads / head dimension	10 / 64
Self-attention	attention implementation	causal FlashAttention
Position encoding	rotary embedding	1D RoPE
Normalization	pre-norm type / epsilon	RMSNorm / 10^{-5}
Feed-forward	activation and gating	SwiGLU
Feed-forward	expansion ratio / hidden width	$8/3$ / 1706
Regularization	attention dropout / residual dropout	0.0 / 0.0
Output head	world latent projection	LayerNorm + Linear, $640 \to 640$
Denoising diffusion network for action generation		
Target projection	condition dimension	640
Input projection	delta token to denoiser width	Linear, $54 \to 512$
Timestep embedding	sinusoidal embedding MLP	$256 \to 512 \to 512$
Condition embedding	transformer hidden state projection	Linear, $640 \to 512$
Denoiser trunk	AdaLN residual MLP blocks	4 blocks, width 512
Residual MLP	per-block MLP	Linear + SiLU + Linear
AdaLN modulation	per-block modulation	SiLU + Linear, $512 \to 1536$
Final layer	adaptive LayerNorm + Linear	$512 \to 108$
Sampling schedule	noise schedule / steps	cosine / 350

Table 6: Auxiliary neural modules used by EgoWAM.

Module	Norm & activation	Output shape / configuration
Hand-state tokenization		
MANO shape adapter	LN, GELU	$10 \to d_{\text{model}} \to d_{\text{model}}$
MANO pose adapter	LN, GELU	$54 \to d_{\text{model}} \to d_{\text{model}}$
Hand embedding	–	learned left/right embedding in d_{model}
World-state tokenizer		
Encoder stem	GN, GELU	Conv3D, $1027 \to 512$, kernel $(1,3,3)$, stride 1
Encoder block 0	GN, GELU	two Conv3D layers, channels 512, kernel $(1,3,3)$
Encoder downsample 1	GN, GELU	Conv3D, $512 \to 512$, kernel $(1,3,3)$, stride $(1,2,2)$
Encoder temporal mix	GN, GELU	two Conv3D layers, channels 512, kernel $(3,3,3)$
Encoder downsample 2	GN, GELU	Conv3D, $512 \to 640$, kernel $(4,3,3)$, stride $(4,2,2)$
Posterior head	–	$1 \times 1 \times 1$ Conv3D, $640 \to 1280$ for $\mu, \log \sigma^2$
Latent input adapter	LN	$640 \to d_{\text{model}}$
Latent prediction adapter	LN	$d_{\text{model}} \to 640$
Decoder bottleneck	GN, GELU	$1 \times 1 \times 1$ Conv3D and 2.5D residual block, channels 640
Decoder upsample 1	GN, GELU	ConvTranspose3D, $640 \to 512$, kernel $(4,4,4)$, stride $(4,2,2)$
Decoder temporal mix	GN, GELU	learned time embedding + two Conv3D layers, kernel $(3,3,3)$
Decoder upsample 2	GN, GELU	ConvTranspose3D, $512 \to 512$, kernel $(1,4,4)$, stride $(1,2,2)$
Decoder output	–	Conv3D, $512 \to 1027$, kernel $(1,3,3)$
Interaction field tokenizer		
Hand segmentation head	–	1×1 Conv2D, $1024 \to 1$ logit
Interaction descriptor	–	195 MANO vertices $\times$ 5 features per hand
Interaction adapter	LN	$975 \to d_{\text{model}}$

The frozen ST-VAE uses 3D convolutional encoder/decoder stages with 2.5D residual blocks and temporal mixing to tokenize world states.

Training configurations. Tab. 7 lists the hyper-parameters used for the ARCTIC 100M rollout training run. The dynamics model is optimized with AdamW using $\beta_1 = 0.9$, $\beta_2 = 0.999$, a base learning rate of 1×10^{-5}, and a cosine schedule with a minimum learning rate of 1×10^{-8}. Weight decay is scheduled from 0.04 to 0.4, and gradients are clipped at 1.0.

Table 7: Hyper-parameters for training EgoWAM.

Training configuration	EgoWAM-100M rollout
Training recipe	
Optimizer	AdamW
Optimizer momentum	$\beta_1 = 0.9,\ \beta_2 = 0.999$
Optimizer epsilon	1×10^{-8}
Model dtype	Float-32
Learning hyper-parameters	
Base learning rate	1.0×10^{-5}
Minimum learning rate	1.0×10^{-8}
Learning-rate schedule	cosine
Warm-up ratio	0.1
Base weight decay	0.04
End weight decay	0.4
Gradient clip	1.0

Compute resources. All experiments are conducted on NVIDIA H20 GPUs, each with 140GB memory. For the 100M-parameter EgoWAM, we use 8 GPUs for teacher-forcing training and 16 GPUs for rollout fine-tuning. For the 300M-parameter model, we use 8 GPUs for teacher-forcing training and 32 GPUs for rollout fine-tuning. For the 600M-parameter model, we use 16 GPUs for teacher-forcing training and 32 GPUs for rollout fine-tuning. The full training pipeline, including teacher forcing and rollout fine-tuning, takes approximately one week for each model scale.

A.4 Baseline Training and Evaluation Details

We evaluate EgoWAM against two categories of baselines. The first category includes action-conditioned world models, namely DINO-WM (Zhou et al., 2025), V-JEPA 2-AC (Assran et al., 2025), and NWM (Bar et al., 2025). These methods predict future visual observations or latent representations conditioned on candidate action sequences, and perform test-time planning by selecting actions based on the predicted outcomes. The second category consists of world-action models, including UVA (Li et al., 2025) and the recent Fast-WAM (Yuan et al., 2026), which jointly model future world states and action trajectories. For a fair comparison, we train or fine-tune all baselines on HOT3D and ARCTIC using the same training schedule as EgoWAM. We provide the implementation and training details below. For HOT3D, the sampled test subject IDs are `P0002`, `P0003`, and `P0014`. We use all recordings performed by these subjects as the test set, and use the remaining subjects for training.

For DINO-WM, we use a frozen DINOv2 ViT-L/14 visual encoder and a ViT-L transition predictor, resulting in 626.9M parameters in total. The model takes a 4-frame visual history with frame skip 1 as input, and uses future absolute MANO poses and camera extrinsics as action conditions. We do not include proprioceptive states.

For NWM, we train the CDiT-L/2 Navigation World Model, which has approximately 687M parameters. We adapt it as an egocentric image-latent diffusion world model. The model uses 4 RGB history frames as context and conditions on future absolute MANO hand poses and camera extrinsics. The action dimensionality is dataset-specific, with 74 dimensions for HOT3D and 134 dimensions for ARCTIC.

For V-JEPA 2-AC, we freeze a pretrained V-JEPA2 ViT-L video encoder with 303.9M parameters and train only a 24-layer action-conditioned predictor. The predictor contains 304.5M parameters for HOT3D and 304.7M parameters for ARCTIC. We train the model on 5-frame clips, using the first 4 frames as history context. The predictor is conditioned on MANO action/state variables and ground-truth camera extrinsics, and is trained to predict future visual representations and roll out MANO states.

For UVA, we follow the official two-stage training recipe and adapt it to egocentric hand–object videos. The model uses a frozen KL-16 image VAE with 66.5M parameters and a trainable MAR-Base autoregressive transformer backbone with two diffusion heads attached: one for next-frame

visual tokens and one for actions, totaling 261.3M parameters for HOT3D and 261.5M parameters for ARCTIC. We use a video horizon of $H{=}8$ frames (first 4 as history, remaining 4 as future targets) and the same delta-MANO action parameterization as EgoWAM, giving an action dimensionality of 48 on HOT3D and 108 on ARCTIC.

For Fast-WAM, we fine-tune from the released checkpoint, a Mixture-of-Transformers world–action model built on the Wan2.2 TI2V-5B text-to-video diffusion transformer. The model couples a video expert (Wan2.2 video DiT, 5.00B parameters) with a parallel ActionDiT (1.02B parameters) through MoT mixed attention, totaling 6.02B trainable parameters; the Wan2.2 VAE and T5 text encoder are kept frozen and external. Since HOT3D and ARCTIC do not provide per-clip language annotations, we use a single fixed instruction per dataset (e.g., "manipulating objects with both hands in HOT3D / ARCTIC egocentric video") as text condition, encoded once offline by the T5 and reused as a constant cross-attention context for every clip.