ObjectForesight: Predicting Future 3D Object Trajectories from Human Videos

Rustin Soraki[1], **Homanga Bharadhwaj**[2,*], **Ali Farhadi**[1,*], **Roozbeh Mottaghi**[1,*]

[1]University of Washington, [2]Carnegie Mellon University

[*]Equal advising

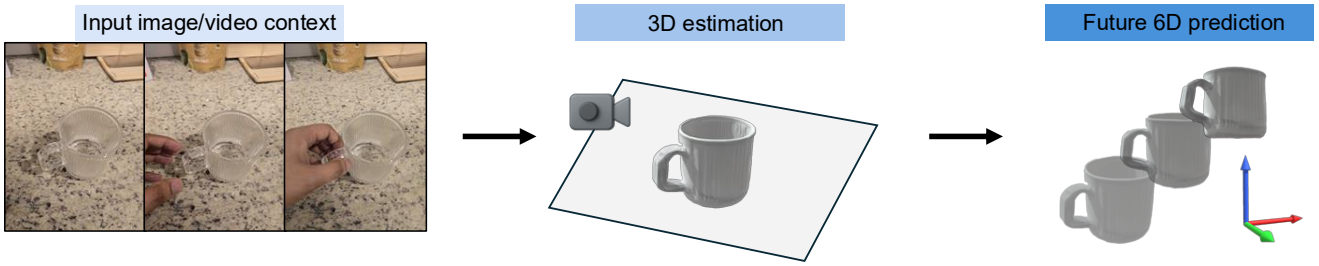

Figure 1. We introduce **ObjectForesight**, a framework for predicting future 3D object trajectories from a video context of past motion. We first estimate the object's 3D shape and initial pose, and then predicts its future 6D poses over time. There are three key contributions: (1) introducing and formalizing the task of 3D object dynamics prediction from human videos, (2) a 3D object-centric dynamics model for future prediction of 6-DoF trajectories, (3) a large-scale dataset of 2 million+ object-centric 3D trajectories with pseudo-groundtruth. Video results are in the website objectforesight.github.io

Abstract

*Humans can effortlessly anticipate how objects might move or change through interaction—imagining a cup being lifted, a knife slicing, or a lid being closed. We aim to endow computational systems with a similar ability to predict plausible future object motions directly from passive visual observation. We introduce **ObjectForesight**, a 3D object-centric dynamics model that predicts future 6-DoF poses and trajectories of rigid objects from short egocentric video sequences. Unlike conventional world/dynamics models that operate in pixel or latent space, ObjectForesight represents the world explicitly in 3D at the object level, enabling geometrically grounded and temporally coherent predictions that capture object affordances and trajectories. To train such a model at scale, we leverage recent advances in segmentation, mesh reconstruction, and 3D pose estimation to curate a dataset of 2 million+ short clips with pseudo-ground-truth 3D object trajectories. Through extensive experiments, we show that ObjectForesight achieves significant gains in accuracy, geometric consistency, and generalization to unseen objects and scenes—establishing a scalable framework for learning physically grounded, object-centric dynamics models directly from observation.* objectforesight.github.io

1. Introduction

Humans possess an intuitive understanding of how the world around them can change through interaction. When we see a cup on a table, we can effortlessly imagine it being picked up, tilted, or placed elsewhere. Watching a hand reach toward a knife, we can anticipate the knife's motion and the transformation of the objects it touches. Such inferences go beyond recognizing what *is* — they reflect our ability to imagine what *can be*. This capacity to mentally simulate object interactions is central to intelligent behavior, allowing us to plan, predict, and act effectively in the physical world.

Our goal in this work is to endow computational systems with a similar capability: to infer and predict plausible *future* configurations of objects from passive visual observation. **We focus on the problem of predicting 3D object dynamics** — *learning how objects can move and interact in 3D space as a result of human actions, without directly modeling the human motion itself.* Rather than learning explicit manipulation trajectories or low-level control policies, we seek to model their *effects*: the diverse, physically coherent object motions that arise from everyday interactions.

To this end, we present **ObjectForesight**, a 3D object-centric forward dynamics model that learns to predict future 6-DoF trajectories of *rigid* objects from egocentric human

videos. Given a sequence of RGB frames and an object mesh, **ObjectForesight** predicts a temporally coherent sequence of future object poses — effectively imagining how the object may move in the near future (Fig. 1). Operating in object-centric coordinates allows the model to generalize across varied objects, scenes, and manipulation styles, capturing the underlying semantics of object affordances.

For training **ObjectForesight**, a key challenge is data: There are no large-scale, clean, and physically grounded 3D interaction datasets. Existing robot datasets capture limited, scripted manipulations with explicit action supervision [32], while internet-scale human video corpora on their own, though rich and diverse, lack aligned 3D information such as object poses, camera geometry, or depth [15, 38]. To address this, we develop a scalable data curation pipeline that transforms passive human videos into structured 3D motion supervision. Specifically, we extract 2 million short clips (2–3 seconds each) from the EPIC-Kitchens dataset [9], automatically detecting hands [47] and identifying objects in contact using SAM [35]. We then recover 3D object meshes and poses with TRELLIS [43], and estimate camera motion and monocular depth using SpaTrackerv2 [45]. By expressing object poses relative to the first-frame camera coordinates, we effectively disentangle ego-motion from object motion. This process converts ordinary egocentric videos into a large-scale dataset of 3D object trajectories — the first at this level of scale, fidelity, and semantic diversity.

ObjectForesight integrates a Diffusion Transformer (DiT) [33] with a geometry-aware 3D point encoder, PointTransformerV3 [42], to jointly reason about object motion and surrounding scene context. Given a short history of RGB frames with corresponding monocular depth maps and a mask of the object in the anchor frame, the model encodes the local 3D geometry of the scene and the object's recent motion into a unified representation. Conditioned on this visual and spatial context, **ObjectForesight** predicts a distribution over future 6-DoF object poses through a denoising diffusion process. This formulation enables robust, multi-modal prediction of dynamically feasible and physically consistent object motions, maintaining geometric fidelity and temporal coherence across predicted trajectories.

In summary, *we introduce the task of predicting future 3D object dynamics from videos — a core capability for embodied visual reasoning, and build models and datasets towards this task.* Our key contributions are as follows:

- We introduce and formalize the task of **3D object dynamics prediction from human videos**, establishing a standardized setting for learning how objects move in the real world. This formulation enables models to leverage the vast amount of in-the-wild egocentric video data to learn physical interaction priors without requiring explicit action supervision.
- We propose **ObjectForesight**, a 3D object-centric dynamics model that predicts future 6-DoF trajectories of objects from short egocentric video snippets and monocular geometry.
- We construct a large-scale dataset of object-centric 3D trajectories from 2 million EPIC-Kitchens clips, using automatic object segmentation and pose estimation to recover high-quality 3D motion supervision from generic interaction videos.

Across extensive experiments in daily human activities, **ObjectForesight** produces accurate, stable, and physically coherent 6-DoF trajectories in diverse real-world scenes. The diffusion-based formulation outperforms autoregressive models and video-generation approaches, offering sharper long-horizon consistency and better multimodal prediction. These results show that large-scale observational data, combined with explicit 3D reasoning, provides a strong foundation for reliable and scalable object-centric motion forecasting.

2. Related Works

Extracting Representations from Human Videos. Large-scale egocentric datasets such as Something-Something [15], YouCook [10], EPIC-Kitchens [9], EGTEA [27], and Ego4D [16] have enabled learning rich representations of human–object interactions directly from video. Early work focused on recovering 3D hand and object poses [13, 19, 22, 37, 50] and reconstructing object geometry [20, 21, 24, 44], providing geometric supervision for understanding interaction. Advances in tracking and scene flow [12, 18, 23, 45] further enable dense motion estimation across time, while recent segmentation and reconstruction systems such as SAM [35], TRELLIS [43], and very recently SAM3D [8] make it possible to automatically extract 3D trajectories of objects from in-the-wild images and videos. Our work is closely related in that it leverages these advances to *curate a dataset of object-centric 3D trajectories at scale*, transforming ordinary human videos into a resource for training predictive models of object dynamics. By building upon existing 3D pose estimation and reconstruction pipelines, we focus not on estimating geometry itself, but on learning how objects move and interact over time.

Predicting Manipulation Cues from Human Videos. Another line of research focuses on predicting or reasoning about manipulation cues and affordances from human videos. Classical works in affordance learning [5, 14, 28–31] study how objects are grasped, where contact occurs, how hands move in the future [3, 7, 28] or which parts of an object afford specific actions. More recent approaches [4, 11, 38] learn to anticipate manipulation outcomes or future contact regions, connecting perception to physical reasoning. Such methods primarily operate in 2D or intermediate feature space, forecasting human or

object-centric cues that signal future interactions. Our work shares the goal of extracting predictive signals from human videos but differs in focus. Rather than predicting contact maps or categorical actions, we aim to learn a continuous model of *3D object dynamics*—how objects themselves move in space as a result of human interactions. By grounding prediction in SE(3) pose space and explicit geometry, we extend affordance learning toward physically coherent, object-centric reasoning about future motion.

World Models and Trajectory Representations. Building models of how the world evolves in response to interaction has long been a core challenge in both computer vision and robotics. Recent efforts in visual world modeling have primarily focused on learning predictive representations either at the pixel level through video generation [39, 40] or in latent spaces through representation learning [1, 17, 25]. While such approaches capture temporal dependencies, they often lack explicit 3D grounding and object-level motion prediction. In contrast, our work develops an *explicit 3D object trajectory model* that operates in SE(3) space. Instead of predicting future pixels or abstract latent codes, our method explicitly models object evolution in 6-DoF pose space, and unlike implicit language conditioning [46], conditions explicitly on predicted object geometry and past motion context, offering a physically grounded representation well-suited for integration into robotic manipulation frameworks [4, 26].

3. Method

We aim to learn a forward dynamics model that predicts future 3D poses of rigid objects from passive human videos. The task involves inferring plausible 6-DoF trajectories conditioned on observed object geometry, local scene context, and a short history of object motion. Since no dataset exists for this setting, we construct a large-scale dataset of 3D object trajectories from egocentric human activity videos using off-the-shelf vision models (Sec. 3.2). We then train a diffusion-based transformer model (Sec. 3.3) that learns to sample diverse, physically consistent future trajectories conditioned on visual and geometric context.

3.1. Overview

ObjectForesight tackles the problem of predicting future 3D object motion from short windows of egocentric video. Given C observed frames and a prediction horizon of H, the goal is to model a distribution over the next H future 6-DoF poses of a manipulated object. All frames in the window are expressed in the anchor-frame (first frame of the prediction horizon) camera coordinates, allowing us to isolate true object motion from ego-motion. In our default setting, we use $C{=}3$ and $H{=}8$.

Formally, we observe images $\mathcal{I}_{1:C}$ and their corresponding object poses

$$\mathbf{P}_{1:C} = [\mathbf{p}_1, \ldots, \mathbf{p}_C], \qquad \mathbf{p}_t \in \mathrm{SE}(3),$$

where each pose token $\mathbf{p}_t = [x_t, y_t, z_t, \mathbf{r}_{t,6\mathrm{D}}]$ contains translation and a continuous 6D rotation representation [48]. Depth from the anchor frame is backprojected to form a point cloud $\mathbf{X}$, and normalized object bounding boxes $\mathbf{B}_{1:P}$ provide coarse spatial cues. The forecasting target is the future sequence

$$\mathbf{P}_{\text{future}} = [\mathbf{p}_{t_a}, \ldots, \mathbf{p}_{t_a+H-1}], \quad t_a = C{+}1.$$

ObjectForesight contributes both *data* and *modeling*: (i) a large-scale pipeline that converts raw egocentric videos into metrically grounded, anchor-frame–canonicalized 6-DoF trajectories; and (ii) a geometry-aware diffusion model that predicts future object motion from these trajectories.

Our predictive architecture combines a context-conditioned geometry encoder over the anchor-frame point cloud with a Diffusion Transformer (DiT) temporal backbone. The encoder conditions point features on the recent motion context (FiLM) and pools them into an object-centric scene embedding $\mathbf{z}_{\text{geom}}$, while the DiT models a distribution over future pose sequences conditioned on $\mathbf{z}_{\text{geom}}$ and an explicit pose-token prefix.

The model operates in a depth-normalized pose space for stability, and uses a cosine noise schedule with v-parameterized denoising. At inference, DDIM sampling produces smooth, diverse, and physically coherent 3D trajectories.

3.2. Data Curation: From Egocentric Video to 3D 6-DoF Object Trajectories

Our curation pipeline converts in-the-wild egocentric videos into clean, metrically grounded trajectories of hand-manipulated objects (Fig. 2). Starting from EPIC-Kitchens action segments, we apply a sequence of automatic extraction and quality gates to recover temporally coherent 6-DoF poses. We summarize the key stages below.

Action segment prefiltering. We begin from annotated single-activity segments and discard clips longer than 10 seconds to limit drift and ensure short, interaction-centric windows.

Hand–object discovery with EgoHOS. For each remaining clip, we run EgoHOS [47] to segment hands and candidate manipulated objects frame-wise. Frames without hands or without any object hypotheses are removed. This yields per-frame masks for (i) active hand(s) and (ii) plausible manipulated objects.

Robust object masks with temporal consensus. We initialize SAM2 [35] using point prompts derived from

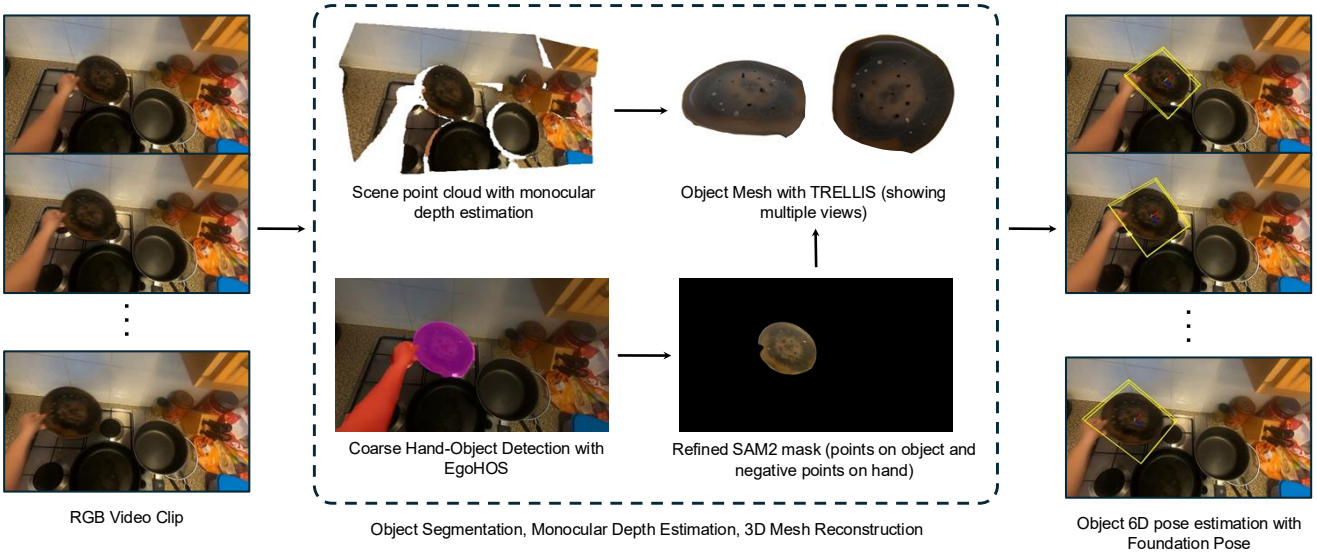

Figure 2. **Data curation pipeline from egocentric video to 3D object trajectories.** Starting from EPIC-Kitchens action segments, we detect hands and objects, refine masks, and filter for clear manipulations. We then reconstruct an object mesh, recover metric depth and camera geometry, and do 6-DoF pose estimation and tracking. Sliding windows over these tracks yield short, clean, anchor-frame–canonicalized 6-DoF trajectories used to train **ObjectForesight**.

EgoHOS masks and propagate a single object instance through the clip. Positive prompts come from the interior of the EgoHOS object mask; negative prompts are drawn from the hand mask, the other-hand object (if present), and a thin ring around the object boundary. To mitigate occasional EgoHOS failures, we form *temporal consensus prompts*, intersections of masks over a small temporal window, which bias SAM2 toward temporally stable shapes. Newly proposed SAM2 masks must have low IoU with the active tracks to prevent duplication. The result is a temporally smooth, occlusion-resilient object mask sequence.

VLM gating for manipulation and view quality. We apply a two-stage VLM-based filter using InternVL3 [49]. First, at the video level, we check whether the highlighted object is actually moved by hand; static objects are discarded. Second, at the frame level, we crop around the object and evaluate visibility (no blur, limited occlusion). Frames passing this test form the set of *clean views*.

Object 3D reconstruction from clean views. TRELLIS [43] reconstructs a 3D object mesh from clean views. The mesh is *not* used during **ObjectForesight** training; it only serves as a geometric template for model-based pose estimation under occlusion.

Model-based 6-DoF pose with metric depth and amodal masks. SpaTrackerV2 [45] provides metric depth and camera geometry. We use Diffusion-VAS [6] to complete amodal object masks. Pose initialization and tracking use FoundationPose [41] with three modifications for egocentric video:

(i) *Metric scale estimation.* TRELLIS meshes lack scale; we estimate scale by comparing masked depth points to mesh radii across neighboring frames (robust weighted median), then refine via depth–silhouette alignment.

(ii) *Multi-view initialization.* We pick up to five clean views, run FoundationPose initialization, and refine each using depth alignment and silhouette consistency. We choose the best by FoundationPose score, with an IoU-based override. Low-IoU cases are discarded.

(iii) *Bidirectional tracking with re-registration.* From the best initialization we track forward and backward. If projection IoU drops below 0.1, we trigger local re-registration using the current mask. This produces temporally coherent pose tracks with explicit re-registration events.

Trajectory slicing and final quality control. We slide a window of length $C+H$ along each track. A window is kept if it lies within a single registration segment and maintains stable projection IoU (no drop > 0.1). All poses are re-expressed in the anchor-frame camera coordinates to remove ego-motion.

Outcome. This automatic pipeline enforces (i) manipulation validity (VLM gating), (ii) mask fidelity (SAM2 with temporal consensus and amodal completion), and (iii) metric, temporally coherent poses (FoundationPose with depth, geometry, and re-registration). The result is a large collection of short, clean, object-centric trajectories suitable for training multi-modal 3D dynamics models.

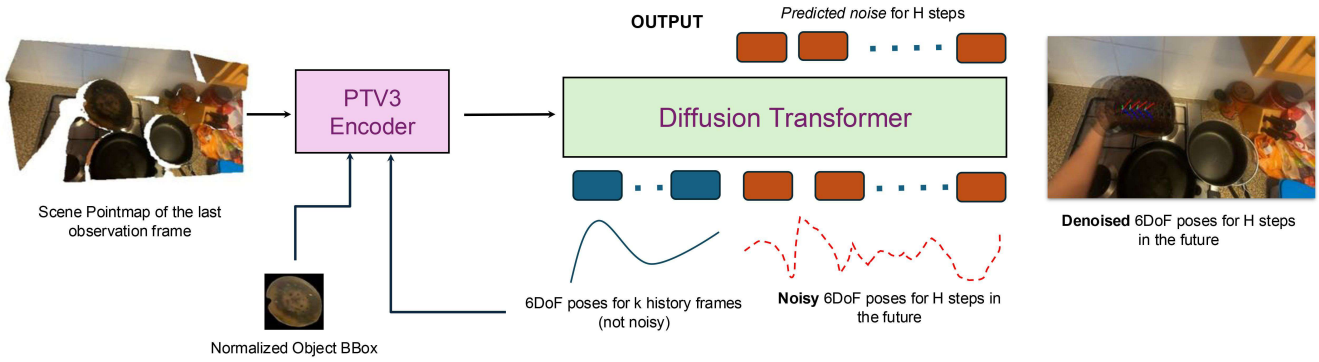

Figure 3. **Model architecture.** Given past pose tokens and their normalized bounding boxes, we summarize motion context with anchor-query attention and use it to guide object-centric pooling in a PointTransformerV3 encoder, producing a geometry-aware scene embedding. A diffusion transformer (DiT, AdaLN-Zero) then denoises future depth-normalized pose tokens, conditioned on the scene embedding and an explicit prefix of past pose tokens. This design allows **ObjectForesight** to generate diverse, physically coherent, and temporally smooth 3D motion predictions.

3.3. Predicting Future Trajectories in 3D

Our forecasting model learns to generate diverse and physically coherent future pose sequences conditioned on the current geometric and motion context. It combines a geometry-aware encoder with a diffusion-based transformer operating on object-centric, depth-normalized 9D pose tokens. An outline of the model is depicted in Fig. 3.

Scene and Context Encoding. Given an anchor-frame point cloud $\mathbf{X} \in \mathbb{R}^{N\times 3}$, conditioning pose tokens $\mathbf{P}_{1:t_a} \in \mathbb{R}^{t_a\times 9}$, and corresponding normalized bounding boxes $\mathbf{B}_{1:t_a} \in \mathbb{R}^{t_a\times 4}$, our goal is to construct a compact representation that summarizes both the recent motion and the 3D scene structure. Here N is the number of points sampled from the anchor-frame depth map. We use C pre-anchor context frames, so the anchor index is $t_a = C{+}1$.

For each frame k in the conditioning sequence, we concatenate the 9D pose token and 4D box into a 13D vector and project it into a D-dimensional context space. We then pool the conditioning sequence with attention: the anchor token queries all conditioning tokens, and we add a sinusoidal embedding of the relative time to the anchor with a learnable scale. This yields a single context vector $\mathbf{ctx} \in \mathbb{R}^D$.

We feed the point cloud $\mathbf{X}$ into a PointTransformerV3 encoder [42]. Each point is represented by its anchor-camera coordinates and its coordinates in the estimated anchor object frame, enabling object-centric reasoning. We also provide the encoder with $\mathbf{ctx}$ which conditions the point cloud features on it using feature-wise linear modulation (FiLM) [34]. We then pool point features into a global scene embedding $\mathbf{z}_{\text{geom}} \in \mathbb{R}^{512}$ using an object-centric attention head that matches point features to a query derived from $\mathbf{ctx}$ and biases weights toward points near the object. $\mathbf{z}_{\text{geom}}$ is then used as a conditioning signal, injected using AdaLN-Zero [33] inside the DiT blocks.

Tokenization of Pose Sequences. We operate on object-centric pose tokens expressed in the anchor-frame camera coordinates. Each pose $\mathbf{p}_t = [x_t, y_t, z_t, \mathbf{r}_{t,6\text{D}}]$ is reparameterized into a depth-normalized token:

$$\mathbf{y}_t = [u_t, v_t, s_t, \mathbf{r}_{t,6\text{D}}], \quad u_t = \tfrac{x_t}{z_t},\ v_t = \tfrac{y_t}{z_t},\ s_t = \log z_t,$$

which reduces the dynamic range of translation and improves numerical stability in egocentric perspectives. For the future horizon of length H, we form $\mathbf{Y}_{\text{future}} = [\mathbf{y}_{t_a}, \ldots, \mathbf{y}_{t_a+H-1}]$. We then apply channel-wise standardization using statistics $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathbb{R}^9$ estimated over the first training batches (and fixed thereafter).

We apply the same depth reparameterization and standardization to the conditioning poses, yielding normalized context tokens $\tilde{\mathbf{P}}_{1:t_a}$. These tokens are embedded and prepended as a prefix to the future sequence inside the transformer, giving the DiT access to the full conditioning history while $\mathbf{B}_{1:t_a}$ contributes to the pooled context vector $\mathbf{ctx}$.

Forward Diffusion Process and Cosine Schedule. Let $\tilde{\mathbf{Y}}_0 \in \mathbb{R}^{H\times 9}$ be the clean normalized future sequence for the batch. Following the standard diffusion framework, we define a forward noising process with a cosine β-schedule and sample timesteps uniformly from $\{0, \ldots, T{-}1\}$ (we use $T{=}1000$). The DiT processes the noised sequence $\tilde{\mathbf{Y}}_t$ as a length-H sequence of 9D tokens, conditioned on the timestep embedding, the geometric embedding $\mathbf{z}_{\text{geom}}$, and the normalized context tokens $\tilde{\mathbf{P}}_{1:t_a}$.

Tokens are embedded into a latent sequence and augmented with learned absolute positions, a token-type embedding (context vs. future), and a signed anchor-relative time embedding. Conditioning is injected via AdaLN-Zero

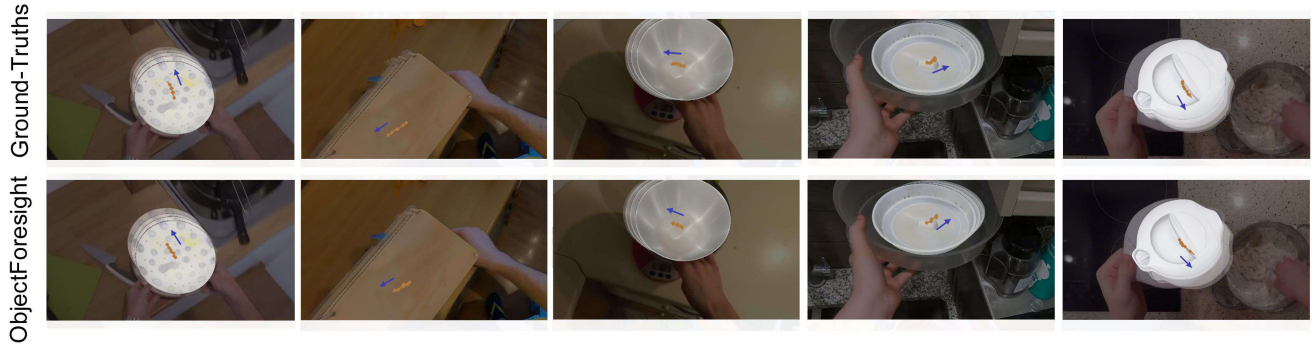

Figure 4. **Qualitative results from ObjectForesight.** Given only the past context and the anchor-frame geometry, **ObjectForesight** generates physically plausible and semantically meaningful 6-DoF trajectories of manipulated objects. For each sequence, we overlay 8 predicted poses on the last observed frame, illustrating both (i) the projected coordinate axes and (ii) the transformed object mesh, with increasing blur indicating further steps into the future. The images are zoomed in for clarity, and the arrows indicate the direction of motion. These results and additional results on HOT3D-clips are best viewed as videos in the website objectforesight.github.io

where a lightweight MLP combines timestep and scene embeddings into per-layer normalization modulations and gated residuals within each transformer block.

v-Parameterization with p2 Weighting. Instead of predicting the noise $\boldsymbol{\epsilon}$ directly, we adopt v-parameterization, which stabilizes training across timesteps. We train with an SNR-weighted regression loss (p2 reweighting [36]) and additionally apply horizon-aware weighting that linearly increases toward later forecast steps (from 1 to 3 across the horizon). During training and DDIM sampling, we reconstruct $\hat{\tilde{\mathbf{Y}}}_0$ from the predicted $\mathbf{v}_\theta$ using the standard closed-form relation.

Denormalization and Pose Decoding. To obtain physical 9D pose tokens from the network outputs, we invert both the standardization and the depth reparameterization. First,

$$\hat{\mathbf{Y}}_0 = \hat{\tilde{\mathbf{Y}}}_0 \odot \boldsymbol{\sigma} + \boldsymbol{\mu}, \tag{1}$$

where $\odot$ denotes elementwise multiplication. Each decoded token $\hat{\mathbf{y}}_t = [\hat{u}_t, \hat{v}_t, \hat{s}_t, \hat{\mathbf{r}}_{t,6\mathrm{D}}]$ is then mapped back to $(\hat{x}_t, \hat{y}_t, \hat{z}_t)$ via

$$\hat{z}_t = \exp(\hat{s}_t), \qquad \hat{x}_t = \hat{u}_t \hat{z}_t, \qquad \hat{y}_t = \hat{v}_t \hat{z}_t.$$

The full 9D pose token is then $[\hat{x}_t, \hat{y}_t, \hat{z}_t, \hat{\mathbf{r}}_{t,6\mathrm{D}}]$.

SE(3) Losses. Since we predict poses in anchor-frame camera coordinates, we can supervise the decoded SE(3) trajectory directly. For each future step k, we convert the predicted 6D rotation to $\hat{\mathbf{R}}_k \in \mathrm{SO}(3)$ and measure translation error $\|\mathbf{t}_k - \hat{\mathbf{t}}_k\|_2$ and rotation error via the SO(3) geodesic angle $\mathrm{d}_{\mathrm{geo}}(\mathbf{R}_k, \hat{\mathbf{R}}_k)$. We average both errors over the horizon (converting degrees to radians) and use

$$\mathcal{L}_{\mathrm{aux}} = \mathbb{E}\big[\bar{\alpha}_t \left(\lambda_R \, \overline{d}_{\mathrm{geo}} + \lambda_{\mathrm{trans}} \, \overline{e}_{\mathrm{trans}}\right)\big],$$

where $\overline{d}_{\mathrm{geo}}$ is the horizon-averaged geodesic rotation error (in radians) and $\overline{e}_{\mathrm{trans}}$ is the horizon-averaged translation error. The expectation is over training samples and sampled diffusion steps, λ_R and λ_{trans} balance rotation and translation, and $\bar{\alpha}_t$ downweights very noisy steps where $\hat{\tilde{\mathbf{Y}}}_0$ reconstruction is less reliable. We regularize dynamics with SE(3) velocity and acceleration losses on increments (also weighted by $\bar{\alpha}_t$): let $\Delta\mathbf{t}_k = \mathbf{t}_{k+1} - \mathbf{t}_k$, $\Delta\mathbf{R}_k = \mathbf{R}_k^\top \mathbf{R}_{k+1}$, and Δ^2 denotes second differences, then

$$\mathcal{L}_{\mathrm{vel}} = \overline{\|\Delta\mathbf{t}_k - \Delta\hat{\mathbf{t}}_k\|_2^2 + \mathrm{d}_{\mathrm{geo}}(\Delta\mathbf{R}_k, \Delta\hat{\mathbf{R}}_k)^2},$$
$$\mathcal{L}_{\mathrm{acc}} = \overline{\|\Delta^2\mathbf{t}_k - \Delta^2\hat{\mathbf{t}}_k\|_2^2 + \mathrm{d}_{\mathrm{geo}}(\Delta^2\mathbf{R}_k, \Delta^2\hat{\mathbf{R}}_k)^2}$$

where $\overline{\cdot}$ averages over valid timesteps k. A small depth-floor penalty is also applied to discourage degenerate solutions with extremely small depth:

$$\mathcal{L}_{z_{\min}} = 0.01 \, \mathrm{ReLU}(z_{\min} - \hat{z}_t).$$

Sampling and Total Objective. At inference time we start from Gaussian noise and perform deterministic DDIM sampling with S denoising steps (we use S=50). We use S evenly spaced timesteps from the training schedule and iteratively denoise under the same conditioning $(\mathbf{z}_{\mathrm{geom}}, \tilde{\mathbf{P}}_{1:t_a})$. At each step we predict $\mathbf{v}_\theta$, reconstruct $\hat{\tilde{\mathbf{Y}}}_0$, and apply the DDIM update to produce $\hat{\mathbf{P}}_{\mathrm{future}}$ in the anchor frame.

The complete training objective combines the main diffusion loss $\mathcal{L}_{\mathrm{v}}$ with pose-space supervision and smoothness terms:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{v}} + \mathcal{L}_{\mathrm{aux}} + \mathcal{L}_{z_{\min}} + \lambda_{\mathrm{vel}} \, \mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{acc}} \, \mathcal{L}_{\mathrm{acc}}.$$

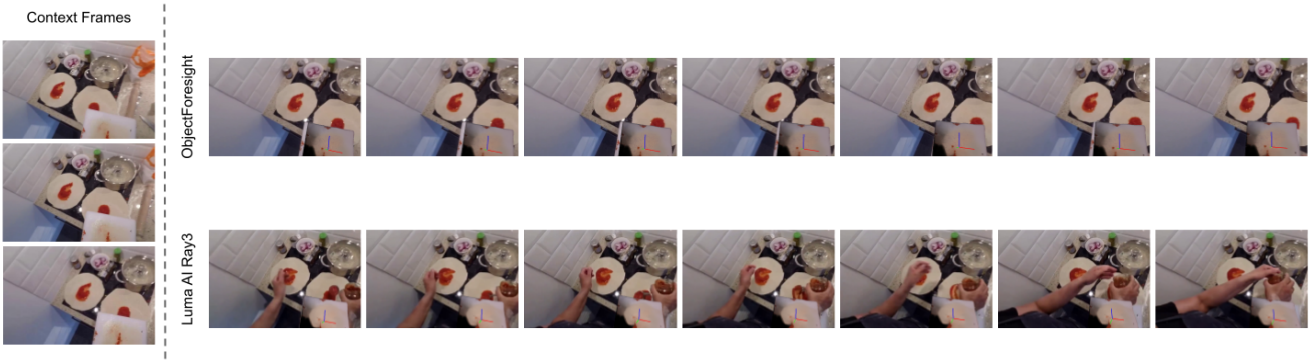

Figure 5. **Visual comparison between generations from ObjectForesight and Luma AI Ray3.** Both methods are conditioned on the same three-frame context. **ObjectForesight** generates future 3D object poses, while Luma AI Ray3 generates a short video. We then apply our pose extraction pipeline to the generated video. Under this procedure, **ObjectForesight** yields temporally consistent pose trajectories, whereas poses extracted from Ray3 generations are less consistent. Results are best viewed as videos in the website objectforesight.github.io

We use $\lambda_R = 2.0$, $\lambda_{\text{trans}} = 20.0$, $\lambda_{\text{vel}} = 0.5$, and $\lambda_{\text{acc}} = 0.1$. The SE(3) auxiliary loss and the smoothness losses are computed on the decoded pose sequence (after denormalization and depth decoding), while $\mathcal{L}_{\text{v}}$ is applied in the normalized token space.

Why Diffusion? Diffusion-based modeling is well suited to 3D interaction dynamics. Given identical 3D conditioning, multiple future motions can be plausible (e.g., a mug can be picked up, slid, or rotated). Our DiT captures this inherently one-to-many nature while encouraging temporally smooth, physically plausible trajectories. In our experiments, it yields more plausiblegeometrically consistent predictions than an autoregressive transformer baseline trained on the same object-centric representation.

Summary. By combining an object-centric 3D scene encoder with an AdaLN-Zero conditioned diffusion transformer over depth-normalized pose tokens, **ObjectForesight** learns a rich conditional distribution over future object motion. The architecture explicitly leverages metric geometry, camera coordinates, and pose history to generate accurate, diverse, and physically plausible 6-DoF trajectories in real-world egocentric scenes.

4. Experiments

Our experiments aim to answer three questions:

1. Is the curated dataset of object-centric 3D trajectories reliable and diverse?
2. Are the predicted future object motions plausible and physically consistent?
3. Does the model generalize beyond the distribution of curated scenes?

4.1. Curated EpicK Dataset Details

We curate a large-scale collection of object-centric 3D motion trajectories from egocentric videos using an automated eight-stage pipeline (Table 2). From **76K** EPIC-Kitchens action segments, we retain **72K** short clips ($\leq$10s) with visible hands to ensure the presence of interactions. Object masks and 2D tracks are obtained using SAM2, yielding **229K** raw tracks before quality filtering reduces them to **112K**. Using TRELLIS, we reconstruct **71K** object meshes and obtain **59K** pose-aligned tracks. Sliding-window extraction produces **3.06M** raw (3+8)-step 3D trajectories, which are further filtered to **2.07M** high-quality trajectories used for training and evaluation.

4.2. HOT3D-Clips

We also train and evaluate **ObjectForesight** on HOT3D-Clips [2] to validate that the model can learn future 3D motion from cleaner trajectories. For the HOT3D experiments, we skip frames to convert the clips to 6 fps and then extract the same $(C+H)$-frame windows used in our main setting.

4.3. Baselines, Ablations, and Metrics

We evaluate three categories of models:

1. **ObjectForesight-DiT**: our diffusion transformer for multimodal trajectory prediction
2. **ObjectForesight-AR**: an autoregressive transformer without diffusion
3. **Video-generation baseline**: an off-the-shelf future video generator (Luma AI Ray3)

For the video-generation baseline, we feed three context frames and let the model synthesize a short future clip. Because running this pipeline is computationally expensive, we apply it to 20 randomly selected videos with clear object visibility, recover 6-DoF motion from the generated frames

	ADE↓	FDE↓	DES↓	ARE↓	FRE↓	RES↓
Epic-Kitchens						
ObjectForesight-DiT	**0.019**	**0.035**	0.005	**7.98°**	13.93°	1.86°
ObjectForesight-AR	0.067	0.074	**0.002**	9.48°	**12.58°**	**0.93°**
vs. Video Generation						
ObjectForesight-DiT	**0.029**	**0.059**	**0.008**	**7.29°**	**13.98°**	**1.77°**
Luma AI Ray3	0.084	0.149	0.020	12.86°	20.90°	2.62°
HOT3D-Clips						
ObjectForesight-DiT	**0.026**	**0.042**	**0.001**	**7.14°**	**11.44°**	**1.50°**
ObjectForesight-AR	0.055	0.082	0.007	9.80°	14.95°	1.55°

Table 1. **Quantitative evaluation of 8-step 3D trajectory forecasting on Epic-Kitchens and HOT3D-Clips.** Lower values indicate better performance. Our diffusion-based model (**ObjectForesight-DiT**) either outperforms or is comparable to the autoregressive variant (**ObjectForesight-AR**) across translation and rotation metrics on both datasets. In the video generation comparison on Epic-Kitchens, ObjectForesight-DiT also substantially outperforms a state-of-the-art video generation baseline (Luma AI Ray3), highlighting the benefits of explicit 3D reasoning compared to image-space synthesis.

Step Name	**Number**
Action Segments	76,885 vids
Selected Vids (hands, $\leq$ 10s)	72,046 vids
SAM2 Tracks	229,102 tracks
Filtered Tracks	112,057 tracks
TRELLIS Models	71,296 models
Objects with Pose Tracks	59,174 tracks
Pre-Filtering Trajectories	3,065,568 trajectories
Post-Filtering Trajectories	2,073,109 trajectories

Table 2. Statistics of the curated dataset of 3D object trajectories from human videos of daily activities. We will release this dataset for the community.

using our curation pipeline, and compute the same trajectory metrics.

We report **ADE** (mean Euclidean error across all timesteps), **FDE** (final-step Euclidean error), **DES** (slope of the per-timestep Euclidean distance error), **ARE** (average rotation error), **FRE** (final rotation error), and **RES** (slope of the per-timestep rotation error). Additional ablations, such as reducing history frames (which increases uncertainty and prediction diversity), are included in the appendix due to space constraints.

4.4. Qualitative Results

Fig. 4 shows that ObjectForesight predicts smooth, physically consistent 6-DoF trajectories that respect scene geometry across a wide range of manipulation scenarios. The model captures realistic interactions, including lifting, rotating, and placing objects, and generates coherent futures in terms of 3D object motion conditioned on the observed context. Fig. 5 presents trajectories recovered from videos generated by Luma AI Ray3. Although these videos exhibit appearance artifacts and provide no explicit 3D constraints, our curation pipeline can still recover approximate object motion from the rendered frames. The recovered trajectories, however, are typically less stable than ObjectForesight's direct predictions, underscoring the advantage of explicitly modeling 3D dynamics rather than inferring them post-hoc from generated videos.

4.5. Quantitative Results

Table 1 reports performance across all 6-DoF trajectory metrics on Epic-Kitchens and HOT3D-Clips. On Epic-Kitchens, ObjectForesight-DiT achieves the best overall translation and average rotation accuracy (e.g., over 3$\times$ lower ADE than ObjectForesight-AR), while the autoregressive variant shows slightly better error-growth trends on a subset of slope/final-rotation metrics. On HOT3D-Clips, ObjectForesight-DiT outperforms ObjectForesight-AR across all metrics, indicating stronger generalization beyond Epic-Kitchens. In the video-generation comparison, ObjectForesight-DiT substantially outperforms Ray3, reinforcing the benefit of predicting motion directly in SE(3) rather than inferring it from synthesized frames.

5. Discussion

We introduce the task of forecasting future 3D object motion directly from passive human videos, framing object dynamics prediction as an object-centric, SE(3) trajectory modeling problem grounded in realistic manipulation behavior. Our large-scale dataset—constructed through automated segmentation, tracking, monocular reconstruction, and pose alignment—provides millions of metrically grounded trajectories that capture how objects move across diverse everyday interactions. Building on this foundation, **ObjectForesight** integrates monocular geometry, recent motion, and local scene structure within a diffusion-based transformer, enabling multimodal prediction of physically consistent futures. Experiments demonstrate strong

performance across translation, rotation, and temporal stability metrics, with **ObjectForesight** outperforming autoregressive and video-generation baselines and generalizing to new objects and environments.

While the current formulation focuses primarily on rigid objects and short-horizon predictions, it suggests several promising directions for future research. One limitation is the rigidity assumption, which restricts the model's ability to capture interactions involving flexible, articulated, or deformable objects. Future work can extend our object-centric representation to more expressive parameterizations—such as articulated kinematic models, learned deformation fields, or neural implicit surfaces—to handle richer categories of everyday objects. Overall, our results establish a foundation for scalable, object-centric 3D dynamics modeling and point toward richer, more general predictive models of physical interaction.

References

[1] 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. 3

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025. 7

[3] 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. 2

[4] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *ECCV*, 2024. 2, 3

[5] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. *arXiv*, 2019. 2

[6] Kaihua Chen, Deva Ramanan, and Tarasha Khurana. Using diffusion priors for video amodal segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 22890–22900, 2025. 4

[7] 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. *arXiv preprint arXiv:2512.16907*, 2025. 2

[8] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. *ECCV*, 2018. 2

[10] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. *CVPR*, 2013. 2

[11] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009. 2

[12] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 2

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. 2

[14] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. *CVPR*, 2022. 2

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. *CVPR*, 2017. 2

[16] 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. *CVPR*, 2022. 2

[17] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ArXiv*, abs/1912.01603, 2019. 3

[18] Adam W Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Suya You, et al. Alltracker: Efficient dense point tracking at high resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5253–5262, 2025. 2

[19] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[20] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and J. Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. *CVPR*, 2020. 2

[21] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. *CVPR*, 2019. 2

[22] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018. 2

[23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 2

[24] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. *ICCV*, 2017. 2

[25] Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. *ArXiv*, abs/1911.12247, 2019. 3

[26] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025. 3

[27] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. *ECCV*, 2018. 2

[28] 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 (CVPR)*, 2022. 2

[29] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. *CVPR*, 2020.

[30] Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian image understanding: Unfolding the dynamics of objects in static images. In *CVPR*, 2015.

[31] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. *CVPR*, 2019. 2

[32] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 2

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023. 2, 5

[34] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 5

[35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 6

[37] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 2

[38] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2

[39] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 3

[40] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European conference on computer vision*, pages 835–851. Springer, 2016. 3

[41] Bowen Wen, Wei Yang, Jan Kautz, and Stanley T. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17868–17879, 2023. 4

[42] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 2, 5

[43] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025. 2, 4

[44] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv*, 2018. 2

[45] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: Advancing 3d point tracking with explicit camera motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6726–6737, 2025. 2, 4

[46] Tomoya Yoshida, Shuhei Kurita, Taichi Nishimura, and Shinsuke Mori. Generating 6dof object manipulation trajectories from action description in egocentric vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17370–17382, 2025. 3

[47] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 2, 3

[48] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2018. 3

[49] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4

[50] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *CVPR*, 2017. 2

A. Additional Details of the Data Curation Pipeline

We elaborate on the stages of the data curation pipeline summarized in Sec. 3.2, focusing on the heuristics, constraints, and cross-stage checks that improve data quality for pose trajectories.

A.1. Presence Filtering and Initialization

To mitigate false segmentations from EgoHOS, we aggregate interaction signals over each clip. We apply run-length smoothing to binary hand/object presence indicators using a threshold proportional to the clip length. This process fills brief detection gaps and eliminates false short-duration positives. The resulting smoothed signals serve two critical functions: they act as execution gates to ensure downstream modules run only when targets are reliably present, and they guide the SAM initialization towards temporally stable windows, reducing error propagation without introducing long-term drift.

A.2. Robust 2D Tracking

We augment the standard SAM2 tracking pipeline with a multi-stage regularization protocol that promotes temporal stability and suppresses duplicate object instances.

Point Sampling Strategy. To initialize and guide the model, we employ a robust sampling strategy. Positive points are sampled from the segmented object mask. To prevent mask leakage into the surrounding context, we explicitly sample negative points from three regions: detected hand masks, other object masks (if present), and a dilated background band surrounding the target object's mask.

Temporal Stability and Consensus. To mitigate per-frame segmentation noise, we construct a short-window consensus mask. When individual frame proposals are noisy, this consensus serves as a high-confidence positive prior. Furthermore, we apply mild morphological opening and closing to eliminate isolated speckles and smooth boundaries.

Trajectory Linking and De-duplication. We associate object components across frames using greedy Intersection over Union (IoU) matching. To handle brief occlusions or detection failures, we permit a small gap tolerance in the temporal sequence. Tracks that fail to meet a minimum length requirement are discarded as noise. Simultaneously, we perform de-duplication within each video clip. If a new object proposal overlaps with an existing active track above a defined IoU threshold within a short temporal window, it is rejected. This ensures that the system maintains unique, distinct identifiers for each object instance.

Initialization. When multiple seeds are available, we prioritize candidates with the largest temporally stable area. Propagation is executed bidirectionally to maximize tracking duration.

A.3. Quality Filtering and Selection

We implement a two-stage filtering protocol to ensure only viable candidates reach the reconstruction stage.

Manipulation Gate. We employ a strict video-level gate using InternVL3 to filter out static or irrelevant objects. This module operates on object-highlighted visual summaries derived from the input track, rather than raw frames. Only tracks exhibiting active manipulation are retained.

Clean-View Selection. For the remaining valid tracks, we categorize frames into *Partial/Invalid* (occluded, blurred, or insufficient resolution) and *Clean* (unambiguous shape). Only clean frames are selected for geometry estimation. Input crops include a context margin to preserve local semantic cues.

A.4. Reconstruction Preparation

We prepare the data for 3D reconstruction through a sequence of filtering and completion steps.

Frame Selection and Background Removal. For the TRELLIS model, we select optimal "clean" frames based on foreground area size, excluding statistical outliers to maximize geometric consistency. Background clutter is masked out to isolate the object on a neutral canvas, enhancing texture and shape recovery.

Amodal Mask Generation. Separately, we use Diffusion-VAS to generate amodal masks. The segmentation masks contain holes or cutouts wherever the object is blocked by hands or other interactions. Diffusion-VAS corrects this by estimating the complete, physical shape of the object, filling in the missing regions. This ensures that we recover the full object silhouette, which is essential for accurate pose estimation and tracking in later steps of the pipeline.

A.5. Pose Estimation and Tracking

We adapt FoundationPose to recover robust 6DoF object trajectories, utilizing camera intrinsics, extrinsics, and dense depth maps provided by SpaTrackerV2. We add the following specific safeguards:

Scale Estimation and Locking. To handle monocular scale ambiguity, we lock the mesh diameter after the initial depth-to-mesh alignment. Subsequent residuals are normalized by this fixed diameter to ensure consistent error scoring across objects of varying sizes.

Initialization Stress-Test. To prevent tracking failures from the start, we do multi-view initialization of object pose. Each potential initial frame undergoes a brief "refine-and-validate" optimization loop that jointly minimizes depth alignment error and maximizes silhouette consistency. Initial views yielding high depth alignment errors or silhouette inconsistencies are rejected.

Bidirectional Tracking and Re-registration. Tracking proceeds bidirectionally (forward and backward) from the optimal seed, with the estimator explicitly re-centered at the

Table 3. **Translation Error Analysis.** Comparison of ADE and FDE across models trained with different horizon lengths (H). Lower is better. Missing values (-) indicate the model cannot predict to that horizon.

	Eval @ $H = 4$		Eval @ $H = 8$		Eval @ $H = 16$		Eval @ $H = 32$	
Train H	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
4	0.016	0.023	-	-	-	-	-	-
8	**0.009**	**0.015**	**0.017**	**0.030**	-	-	-	-
16	0.015	0.020	0.022	0.034	0.034	0.055	-	-
32	0.018	0.022	0.023	0.031	**0.032**	**0.049**	**0.050**	**0.083**

Table 4. **Rotation Error Analysis.** Comparison of ARE and FRE across models trained with different horizon lengths (H). Lower is better.

	Eval @ $H = 4$		Eval @ $H = 8$		Eval @ $H = 16$		Eval @ $H = 32$	
Train H	ARE↓	FRE↓	ARE↓	FRE↓	ARE↓	FRE↓	ARE↓	FRE↓
4	4.81°	7.41°	-	-	-	-	-	-
8	**3.50°**	**6.05°**	**6.53°**	11.47°	-	-	-	-
16	4.63°	6.98°	7.39°	11.82°	11.77°	18.85°	-	-
32	5.95°	7.45°	7.87°	**11.13°**	**11.37°**	**17.77°**	**18.07°**	**29.77°**

anchor frame before each pass. To detect and correct drift, we compute a suite of complementary consistency terms at every step:

- **Silhouette Metrics:** We monitor Intersection over Union (IoU) with specific penalties for *overflow* (mesh projection exceeding the mask) and *underfill* (mesh projection failing to cover the mask).
- **Geometric Residuals:** We track the error between the rendered mesh depth and the observed sensor depth.
- **Motion Monitors:** We apply conservative thresholds on rotation and translation deltas to flag physically implausible jumps.

Re-registration is triggered by compounded evidence from these metrics, allowing the system to curb drift under heavy occlusions or rapid egocentric motion.

B. Ablation Studies

In this section, we conduct a series of ablation studies to evaluate the contribution of different values of context length (C) and prediction horizon (H) and validate the design choices of our proposed framework.

B.1. Ablation Study on Context Length

To determine the optimal temporal receptive field for our method, we conducted an ablation study on the number of input context frames C. We evaluated the model's performance by varying $C \in \{1, 2, 3, 5, 10\}$ while maintaining a fixed prediction horizon of $H = 8$. To ensure a consistent evaluation benchmark across all configurations, the validation set was constructed using the maximum context length ($C = 10$). For models trained with shorter contexts, we trimmed the input sequences accordingly, ensuring that all models predicted the exact same target frames based on the appropriate historical window. The results of this experiment are summarized in Table 5.

Table 5. **Ablation studies on the number of context frames.** We evaluate the impact of context length C on pose prediction accuracy with a fixed prediction horizon of $H = 8$.

C	ADE↓	FDE↓	ARE↓	FRE↓
1	0.026	0.038	7.97°	12.36°
2	0.021	0.033	7.61°	12.14°
3	**0.018**	**0.032**	**7.03°**	12.42°
5	0.025	0.035	7.69°	**11.68°**
10	0.027	0.038	8.09°	12.21°

As illustrated in Table 5, we observe that increasing the context information initially improves prediction accuracy. The performance improves significantly as C increases from 1 to 3, with $C = 3$ achieving the lowest error rates across the majority of metrics, including an ADE of 0.018 and an ARE of 7.03°. This suggests that a context of three frames provides sufficient historical information to effectively capture the object's immediate trajectory and rotational dynamics.

However, increasing the context length beyond this point ($C = 5$ and $C = 10$) results in a performance degradation. For instance, at $C = 10$, the ADE regresses to 0.027, and the ARE increases to 8.09°. We attribute this decline to two primary factors. First, longer context sequences are more susceptible to accumulated noise, which can distract the model from the most relevant recent motion cues. Sec-

ond, an excessively long history may cause the model to overfit to past trajectories, hindering its ability to generalize to dynamic changes in pose movements or sudden shifts in direction. Consequently, we adopt $C = 3$ as the default setting for our main method.

B.2. Ablation Study on Prediction Horizon

We further analyze the impact of the prediction horizon H by training separate models with $H \in \{4, 8, 16, 32\}$ and a fixed input context length $C = 3$. To ensure a fair comparison, the validation set is constructed using the maximum horizon ($H = 32$); for models with shorter output capabilities, we crop the ground truth sequences to match their respective prediction lengths (4, 8, or 16 frames). This setup allows us to evaluate how training on different temporal lengths affects performance at various evaluation horizons.

Table 3 and Table 4 summarize the results for translation (ADE/FDE) and rotation (ARE/FRE) errors, respectively. Columns indicate the evaluation horizon used, while rows represent the model's training configuration.

The results highlight a clear trade-off between short-term precision and long-term capability. Interestingly, the model trained with $H = 8$ outperforms the model trained with $H = 4$ when evaluated at the shorter horizon of $H = 4$ (e.g., ADE decreases from 0.0161 to 0.0095). This suggests that training on a slightly longer horizon encourages the network to learn more robust motion dynamics, acting as a form of regularization that benefits short-term accuracy.

However, blindly increasing the training horizon is not always beneficial. The model trained with $H = 32$ exhibits significantly higher error rates at shorter horizons ($H = 4, 8$) compared to the $H = 8$ model. This degradation likely stems from the optimization difficulty; the loss function for $H = 32$ is averaged over a long sequence where errors naturally accumulate, potentially diluting the gradients for earlier frames. Conversely, for long-term predictions ($H = 16$ and $H = 32$), the model explicitly trained on the larger horizon ($H = 32$) yields the superior performance. This is expected, as models trained with shorter horizons optimize for immediate accuracy and lack the supervisory signal required to maintain trajectory consistency over extended periods. Without the long-term loss component, these models suffer from severe error accumulation (drift) when extrapolating beyond their training window. The $H = 32$ model, by contrast, learns to model global temporal dependencies, effectively trading off some short-term precision for long-term stability.