This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EgoFlow: Gradient-Guided Flow Matching for Egocentric 6DoF Object Motion Generation

Abhishek Saroha[1,2] Huajian Zeng[4] Xingxing Zuo[4] Daniel Cremers[1,2] Xi Wang[1,2,3]
[1] TU München [2] MCML [3] ETH Zürich [4] MBZUAI

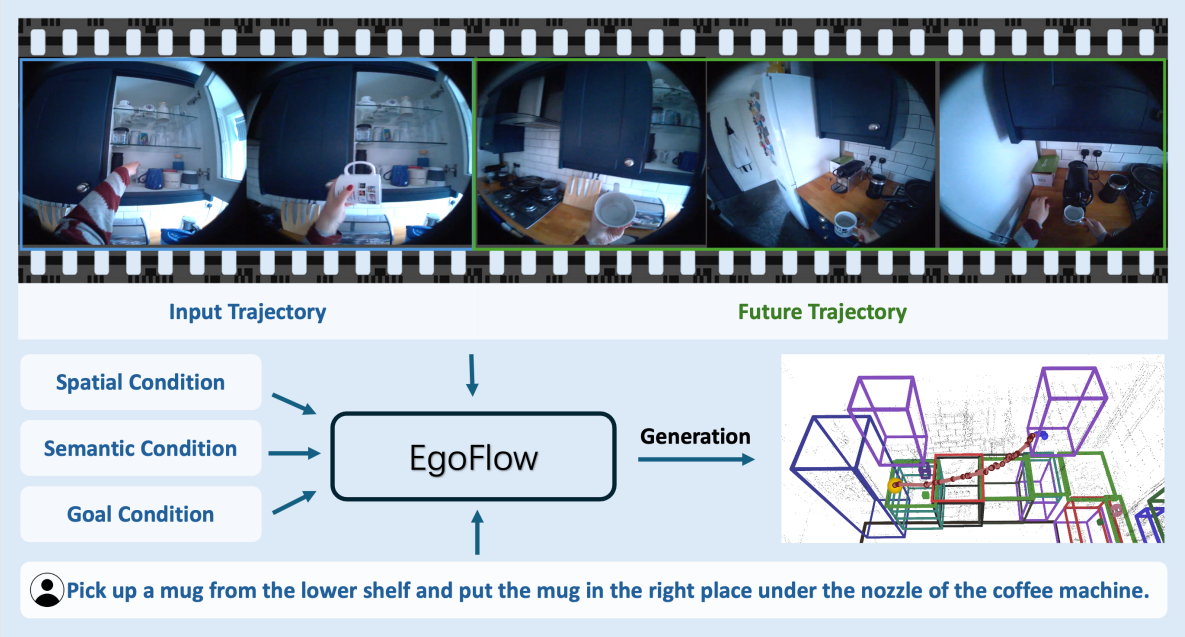


Figure 1. **EgoFlow: a method for object trajectory generation from egocentric videos.** Given a textural command and the surrounding environment, EgoFlow generates physically valid 6DOF object trajectories that respect spatial constraints across diverse environments by learning from egocentric videos.

Abstract

Understanding and predicting object motion from egocentric video is fundamental to embodied perception and interaction. However, generating physically consistent 6DoF trajectories remains challenging due to occlusions, fast motion, and the lack of explicit physical reasoning in existing generative models. We present EgoFlow, a flow-matching framework that synthesizes realistic and physically plausible trajectories conditioned on multimodal egocentric observations. EgoFlow employs a hybrid Mamba–Transformer–Perceiver architecture to jointly model temporal dynamics, scene geometry, and semantic intent, while a gradient-guided inference process enforces differentiable physical constraints such as collision avoidance and motion smoothness. This combination yields coherent and controllable motion generation without post-hoc filtering or additional supervision. Experiments on real-world datasets HD-EPIC, EgoExo4D, and HOT3D show that EgoFlow outperforms diffusion-based and transformer baselines in accuracy, generalization, and physical realism, reducing collision rates by up to 79%, and strong generalization to unseen scenes. Our results highlight the promise of flow-based generative modeling for scalable and physically grounded egocentric motion understanding. Project page: https://abhi-rf.github.io/egoflow/.

1. Introduction

Recent advances in augmented reality devices [6] have enabled large-scale egocentric datasets with fine-grained spatial and semantic annotations [1, 8, 22, 23, 25, 29], providing new opportunities for data-driven learning of object motion trajectories in realistic environments. These egocentric videos capture a rich source of information about object interactions, providing valuable cues about how objects move, are interacted, and evolve over time. Such data offers dense,

first-person evidence for embodied perception and interaction in robotics, where understanding and predicting object dynamics is essential for planning and interaction.

Object motion generation models the temporal evolution of object movements, revealing how humans interact with their surroundings from a first-person viewpoint. However, learning to generate object trajectories from these egocentric videos poses unique challenges. First, egocentric scenes are highly diverse and cluttered, with objects frequently occluded. In addition, egocentric videos have limited field of view and rapid camera motions lead to blurring content. These factors require models to reason over complex geometric layouts and semantic cues to generate accurate trajectories. Second, long-horizon prediction demands consistent temporal reasoning, as small spatial errors can accumulate and lead to unrealistic motion patterns over time. Third, ensuring physical plausibility without explicit physics supervision is challenging; generated trajectories must remain collision-free and dynamically smooth while reflecting realistic motion patterns.

To address these challenges, we propose a generative model EgoFlow that learns scene-conditioned continuous flows for 6DoF object trajectory generation. Rather than using stochastic diffusion, our method leverages flow matching [20, 21] to learn deterministic transport fields in $\mathbb{R}^9$, enabling efficient trajectory synthesis. Building upon the multimodal conditioning framework introduced in GMT [55], EgoFlow further incorporates a hybrid architecture that combines bidirectional Mamba state-space models [9, 58] with Transformer blocks and a Perceiver-based cross-attention encoder [17], allowing scalable sequence modeling and effective fusion of geometric, semantic, and goal-oriented context. Finally, we introduce a gradient-guided sampling strategy that injects differentiable physical costs, specifically SDF-based collision penalties, rotation and velocity smoothness, into the generation loop to improve physical plausibility without requiring explicit constraint supervision.

We evaluate EgoFlow on large-scale egocentric benchmarks to verify both fidelity and generalization. Specifically, we train and test on HD-EPIC [25] to assess physical consistency under realistic scene contexts. Following a similar protocol as in previous work [51], we perform zero-shot transfer from Ego-Exo4D [8] to HOT3D [1] to evaluate robustness in unseen environments. EgoFlow achieves substantial gains over existing generative models, particularly in terms of spatial accuracy and physical feasibility, reducing collision rates by up to 79%. EgoFlow also shows strong generalization to unseen scenes. Our code is publicly available to facilitate future research.

In summary, our contributions are:

- A scene-conditioned flow matching method for 6DoF object motion generation, which learns continuous flows in Euclidean space $\mathbb{R}^9$ for efficient and physically consistent motion synthesis.
- A hybrid architecture that integrates bidirectional Mamba state-space models with Transformer and Perceiver components, enabling scalable sequence modeling and effective fusion of geometric and semantic scene features.
- A gradient-guided sampling strategy that incorporates differentiable physical costs into the generation process, enhancing the physical plausibility of generated trajectories without explicit supervision.
- Comprehensive experiments on egocentric datasets, demonstrating substantial improvements over existing models and strong generalization to unseen scenes.

2. Related Work

Motion and Trajectory Generation. Modeling 3D motion from visual or multimodal data has been widely studied across human motion [12, 13, 38, 56], object dynamics [19, 51], and scene-conditioned motion prediction [59]. Early approaches often relied on short-term kinematic priors or autoregressive prediction, which limited their ability to capture long-range dependencies and global temporal coherence. Recent progress in diffusion-based generative modeling has enabled high-fidelity motion synthesis conditioned on language, geometry, or scene context [19, 43, 57], showing that large-scale motion datasets can be effectively modeled as continuous spatiotemporal distributions. However, existing methods remain constrained to short temporal horizons and often overlook explicit physical consistency such as collision avoidance or trajectory smoothness. In contrast, our framework focuses on generating long-horizon, physically consistent object trajectories.

Egocentric Motion Understanding. Egocentric perception provides a natural framework for studying object and scene dynamics from a first-person viewpoint. Large-scale benchmarks such as EgoExo4D [8], HOT3D [1], and HD-EPIC [25] enable fine-grained analysis of 3D trajectories, spatial context, and visual cues observed during everyday activities. These datasets have motivated new research in egocentric motion forecasting, trajectory synthesis, and cross-view understanding, where models learn to infer plausible 3D trajectories directly from wearable sensors or video streams. For example, EgoScaler [51] generates object trajectories from egocentric visual inputs, while EgoChoir [50] explores scene affordance reasoning from egocentric observations. Despite their success, most existing approaches focus on short-term motion generation or limited geometric reasoning. Our work advances this line of research by introducing a continuous flow generative model capable of synthesizing long-horizon, physically consistent trajectories conditioned on egocentric scene context.

Generative Models and Mamba. Diffusion-based generative models [14, 34, 35] have established a new foundation

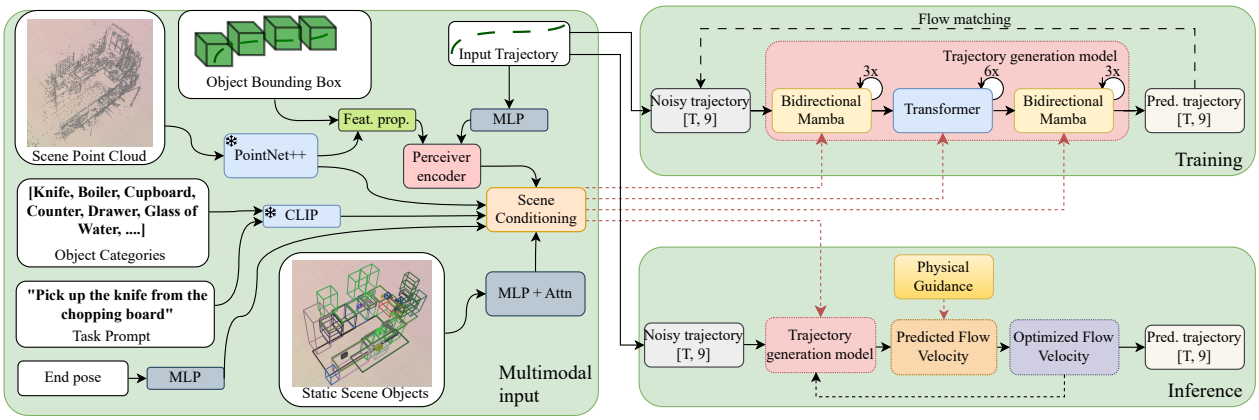


Figure 2. **EgoFlow overview.** Given a 3D scene, a task prompt, and a task goal, our method first fuses multimodal inputs through a scene conditioning block (Sec. 3.2). The fused features are used as conditioning for trajectory generation. We use input trajectories as the source samples of our flow matching model (Sec. 3.3), which maps the generated trajectories to the target distribution, the ground-truth trajectories, through a hybrid architecture (Sec. 3.4). We integrate physical guidance at inference to ensure physical plausible and collision-free trajectories (Sec. 3.5).

for high-quality synthesis across images, audio, and motion. Their continuous extensions using stochastic differential equations provide a principled view of data generation as iterative refinement along a noise-to-data path. More recently, flow matching [20, 40] has emerged as a deterministic alternative, learning continuous velocity fields that transport samples from noise to the target manifold. This formulation offers comparable expressiveness while enabling faster and more stable inference.

Parallel to these algorithmic advances, architectural innovations have improved the scalability of sequence modeling. The Mamba architecture [4, 9] introduces a selective state-space mechanism [10, 11, 33] that supports linear-time inference and constant memory cost, making it suitable for very long sequences. Variants such as Vision Mamba [61] and DiM [7, 37] demonstrate strong performance in large-scale visual and generative tasks.

Building on these advances, our framework integrates flow matching with Mamba-based sequence modeling to achieve scalable, coherent, and physically consistent trajectory generation from complex egocentric scenes.

Concurrent Work. ObjectForesight [36] predicts future 6DoF object trajectories from egocentric video using diffusion. FlowHOI [54] generates hand-object interaction sequences via two-stage flow matching conditioned on egocentric observations and 3D scene reconstructions. EgoMAN [2] and EgoVerse [26] provide large-scale egocentric datasets with 6DoF hand trajectories and diverse manipulation demonstrations, complementary to our work. GMT [55] addresses goal-conditioned 6DoF object trajectory synthesis using a multimodal transformer and shares the input scene conditioning design with our method.

3. Methodology

Object trajectory generation requires capturing both the diversity of feasible motions and the physical regularities that govern real-world interactions. To address this dual challenge, the proposed framework formulates trajectory synthesis as a continuous transport process in $\mathbb{R}^9$, where each frame is represented as position $\mathbb{R}^3$ and continuous 6D rotation [60]. Motion evolution is, therefore, learned as a deterministic flow field rather than a stochastic diffusion. A hybrid Mamba-Transformer backbone further provides structured temporal reasoning and multimodal scene understanding, supporting long-horizon prediction under complex spatial and semantic constraints. Physical consistency during inference is ensured through gradient-guided optimization that dynamically aligns generated trajectories with collision and stability constraints. Together, these components form a unified generative framework capable of producing smooth, physically grounded 6DoF trajectories (see Fig. 2).

3.1. Problem Formulation

We begin by formally defining the object trajectory generation problem. Given a 3D scene with known spatial and semantic context, we aim to generate physically plausible 6DoF object trajectories that satisfy motion and collision constraints. Each trajectory frame is represented as $\mathbf{x}_t = [\mathbf{p}_t; \mathbf{r}_t] \in \mathbb{R}^9$ where $\mathbf{p}_t \in \mathbb{R}^3$ is the 3D position and $\mathbf{r}_t \in \mathbb{R}^6$ is the continuous 6D rotation representation [60]. The input conditions include a point cloud of the scene $`P \in \mathbb{R}^{N\times 3}$, M oriented bounding boxes of fixtures $\mathcal{B} = \{(\mathbf{c}_i, \mathbf{s}_i, \mathbf{r}_i)\}_{i=1}^{M}$ representing centers, sizes, and rotations, object category $\mathcal{C}$, target end pose $\mathbf{x}_T \in \mathbb{R}^9$ for goal conditioning, text prompt describing the task, and observed trajectory history $\mathbf{x}_{1:H} \in \mathbb{R}^{H\times 9}$. Using a radius based selection, the number of fixatures M is typi-

cally around 50. The goal is to generate future trajectory $\mathbf{x}_{H+1:T} \in \mathbb{R}^{(T-H)\times 9}$ that captures plausible object motion while avoiding collisions. In practice, we consider the 30% of the trajectory as input and predict the remaining 70%.

3.2. Multimodal Conditioning

Following the multimodal conditioning design of GMT [55], we first build a unified conditioning representation $\mathbf{u}$ by encoding the following multimodal input.

(i) Trajectory dynamics. The observed history $\mathbf{x}_{1:H}$ is linearly projected to a temporal embedding $\mathbf{F}_{\text{traj}}$ that preserves local kinematics while remaining compatible with cross-modal fusion.

(ii) Local geometric context. The raw scene point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ is encoded by PointNet++ [27] to produce a global descriptor $\mathbf{F}_g$ and per-point features $\mathbf{F}_p$. To provide geometry precisely "at" the manipulated object at each time t, per-point features are propagated from nearby points to the object's oriented box center $\mathbf{c}_t$ using inverse-distance weighting over the k nearest neighbors:

$$\mathbf{F}p = \sum_{t=1}^{H} \frac{\sum_{i=1}^{k} w_i(\mathbf{c}_t)\mathbf{f}_i}{\sum_{i=1}^{k} w_i(\mathbf{c}_t)}, \quad w_i(\mathbf{c}_t) = \frac{1}{|\mathbf{c}_t - \mathbf{p}_i|^2}, \quad (1)$$

where $\mathbf{p}_i$ and $\mathbf{f}_i$ are the coordinates and features of the i-th neighbor point.

(iii) Fixture layout. Static fixtures $\mathcal{B}$ are embedded as tokens from box geometry $(\mathbf{c}_i, \mathbf{s}_i, \mathbf{r}_i)$. Geometric tokens interact via a lightweight self-attention layer to capture pairwise relations (e.g., "counter above cabinet"). To reduce noise, only the M nearest fixtures to the object are retained before attention:

$$\mathbf{F}_b = \text{SelfAttn}\big(\{b_k\}_{k=1}^{M}\big). \quad (2)$$

(iv) Category and task prompt. Object category $\mathcal{C}$ and task description are embedded by CLIP and projected by MLP to $\mathbf{F}_s$, providing behavior priors that complement purely geometric cues (e.g., "drawer" vs. "box" implies different feasible orientations).

(v) Goal descriptor. The target pose $\mathbf{x}_T \in \mathbb{R}^9$ is mapped by MLP to a goal token $\mathbf{F}_{goal}$ that conditions long-horizon planning and mitigates drift from the intended destination.

All conditioning features are concatenated and projected to form the final conditioning vector $\mathbf{u}$:

$$\mathbf{u} = \text{MLP}([\mathbf{F}_{\text{traj}}, \mathbf{F}_p, \mathbf{F}_g, \mathbf{F}_b, \mathbf{F}_s, \mathbf{F}_{\text{goal}}]). \quad (3)$$

This unified conditioning vector is shared across all model stages but modulated differently at each stage to enable hierarchical fusion of spatial, semantic, and task-specific cues.

3.3. Flow Matching for Trajectory Generation

The trajectory distribution is modeled using flow matching [20, 21], a class of continuous normalizing flows that learn deterministic data transport between probability distributions. In contrast to diffusion-based models, flow matching offers two distinct advantages: First, it learns direct velocity fields through simple regression without requiring stochastic differential equation solvers or score estimation. Second, it produces smoother probability paths, enabling efficient generation with fewer integration steps. We extend this formulation to handle 6DoF trajectories conditioned on history observation.

Training. Given a clean trajectory $\mathbf{x}_0 \sim p_{\text{data}}$ and noise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$, we define a linear interpolation:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, t \in [0, 1]. \quad (4)$$

The model learns a time dependent velocity field $\mathbf{v}_\theta$ conditioned on the scene context $\mathcal{S}$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t\sim \text{U}[0,1],\, \mathbf{x}_0, \mathbf{x}_1} \left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathcal{S}) - (\mathbf{x}_1 - \mathbf{x}_0)\|_1 \right], \quad (5)$$

where $\mathbf{v}_\theta$ predicts the instantaneous flow direction at intermediate states. We apply equal L1 loss weights to the position and rotation components of 6DoF trajectory, treating the continuous 6D rotation [60] as an embedding in $\mathbb{R}^6$.

Sampling At inference, we draw an initial sample $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ and integrate the learned velocity field backward using Euler's method [31]:

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - \Delta t \cdot \mathbf{v}\theta(\mathbf{x}_t, t, \mathcal{S}), \quad (6)$$

with $\Delta t = 1/20$ for 20 steps. This deterministic integration produces smooth and physically consistent trajectories within the learned manifold.

3.4. Hybrid Architecture

A hybrid architecture is used to combines efficient sequence modeling with expressive multimodal reasoning to support long-horizon trajectory generation under complex scene constraints. The model integrates three stages: temporal encoding, cross-modal reasoning, and trajectory refinement.

Stage 1: Temporal Context Encoding. The first stage employs three bidirectional Mamba layers [58] to capture long range temporal dependencies with linear computational complexity. Each layer updates the trajectory embedding $\mathbf{h}_t$ through feature-wise linear modulation (FiLM) [24]: $\mathbf{h}'_t = \gamma(\mathbf{u}_t) \odot \mathbf{h}_t + \beta(\mathbf{u}_t)$, where $\odot$ denotes element-wise (Hadamard) multiplication, and γ and β are learnable affine transformations of the conditioning vector. Bidirectional processing combines forward and backward Mamba passes, enabling the model to reason over both past and future temporal context, crucial for generating collision-free, dynamically consistent motion.

Stage 2: Cross-Modal Attention. The second stage consists of six Transformer blocks adapt from the Perceiver architecture [17]. First, self-attention captures temporal dependencies among trajectory tokens. Next, cross-attention allows trajectory features to query the full multimodal conditioning $\mathbf{u}_t$, enabling selective integration of relevant scene and task information. Finally, a position-wise feedforward network (FFN) refines the fused representation. Note that unlike FiLM's affine modulation, cross-attention provides learned attention weights that enable the model to focus on pertinent conditioning information (e.g., nearby obstacles, goal location) for each trajectory segment, enhancing multimodal reasoning.

Stage 3: Trajectory Refinement. The final stage applies another stack of three bidirectional Mamba layers, again conditioned via FiLM. This stage refines the fused representation from Stage 2, integrating temporal coherence with the enriched multimodal context. The output is projected to the $\mathbb{R}^9$ velocity space, comprising both linear and angular velocity components via a linear head, producing the final velocity prediction $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathcal{S})$ for flow matching.

3.5. Gradient-Guided Sampling

Models trained solely on collision-free demonstrations cannot explicitly reason about physical constraints, failing to avoid obstacles in novel scene configurations at test time. We address this through gradient-guided sampling [16, 18, 47] that enforces constraints via differentiable optimization during inference. This approach avoids train-test distribution mismatch: the flow matching model learns motion distributions from data, while physical constraints are imposed only during generation when needed.

At each integration timestep t, before computing $\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - \Delta t \cdot \mathbf{v}_\theta$, we refine the predicted velocity through K=50 gradient descent steps with learning rate α=0.1 and in each step k:

$$\mathbf{v}_\theta^{(k+1)} = \mathbf{v}_\theta^{(k)} - \alpha \nabla_{\mathbf{v}} \mathcal{J}(\mathbf{x}_t - \Delta t \cdot \mathbf{v}^{(k)}), \tag{7}$$

where $\mathcal{J}$ combines three differentiable costs computed on the integrated trajectory state.

Collision avoidance. We compute signed distance fields between trajectory positions and static fixtures $\mathcal{B}$ using analytic formulas for oriented bounding boxes [28]. The minimum distance from the position $\mathbf{p}_j$ to all M fixtures is:

$$d(\mathbf{p}_j) = \min_{i=1}^{M} \text{SDF}(\mathbf{p}_j, \mathbf{c}_i, \mathbf{s}_i, \mathbf{r}_i). \tag{8}$$

The cost penalizes violations of a safety margin ϵ=5cm:

$$\mathcal{J}_{\text{coll}} = \sum_{j=H+1}^{T} \max(0, \epsilon - d(\mathbf{p}_j)). \tag{9}$$

Rotational consistency. Physically plausible trajectories require smooth angular motion with consistent velocity directions, which is enforced in the rotation space by measuring cosine similarity between consecutive rotation changes:

$$\mathcal{J}_{\text{rot}} = \sum_{j=H+1}^{T-1} \left(1 - \frac{\langle \Delta\mathbf{r}_j, \Delta\mathbf{r}_{j-1} \rangle}{\|\Delta\mathbf{r}_j\| \|\Delta\mathbf{r}_{j-1}\|}\right), \quad \Delta\mathbf{r}_j = \mathbf{r}_{j+1} - \mathbf{r}_j, \tag{10}$$

which penalizes abrupt changes in rotation direction.

Translational smoothness. We enforce smooth velocity profiles by penalizing linear acceleration, which prevents abrupt velocity changes as per the following:

$$\mathcal{J}_{\text{vel}} = \sum_{j=H+1}^{T-2} \|\mathbf{a}_j\|, \quad \mathbf{a}_j = \mathbf{v}_{j+1} - \mathbf{v}_j, \quad \mathbf{v}_j = \frac{\mathbf{p}_{j+1} - \mathbf{p}_j}{\Delta t}. \tag{11}$$

The total cost is $\mathcal{J} = \mathcal{J}_{\text{coll}} + \lambda_{\text{rot}}\mathcal{J}_{\text{rot}} + \lambda_{\text{vel}}\mathcal{J}_{\text{vel}}$ and its full differentiability enables gradient-based refinement toward physically plausible trajectories. Further analysis on the formulation and complexity of the gradient-guided sampling can be found in the Supp. Mat.

3.6. Implementation Details

Training. We train using AdamW with learning rate 10^{-4}, batch size 32, for 100 epochs. The implementation uses PyTorch Lightning for efficiency.

Inference. Gradient guidance applies 50 gradient descent steps per integration step with learning rate $\alpha = 0.1$ and collision margin $\epsilon = 0.05$m. Smoothness weights are tuned per dataset: $\lambda_{\text{rot}} = \lambda_{\text{vel}} = 2.0$ for HD-EPIC, $\lambda_{\text{rot}} = \lambda_{\text{vel}} = 5.0$ for HOT3D. Architectural hyperparameters (layer counts, hidden dimensions, attention heads) are detailed in the supplementary material.

4. Experiments

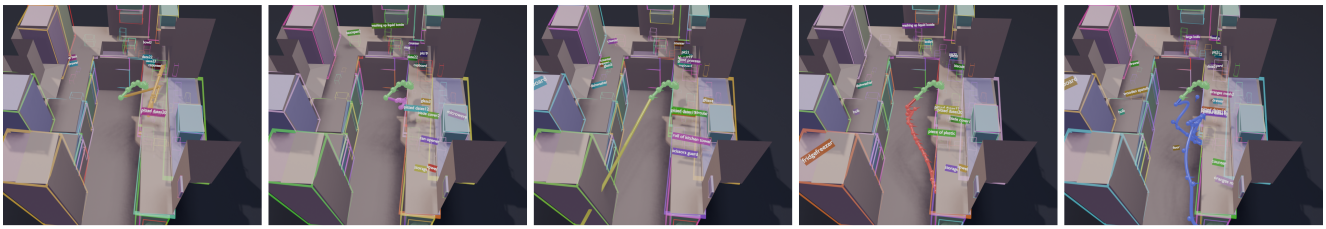
We present the evaluation of our proposed method by first describing baselines and evaluation metrics used in our experiments. Next, we compare our method with baselines on various benchmarks. Finally, we analyze the results and discuss the implications of our findings.

4.1. Experiments Setup

Baselines. Given our task setting, there are no directly comparable existing methods. Following the baseline adaptation protocol of GMT [55], we select three relevant approaches and adapt them to our setting, and additionally evaluate four recent diffusion/flow-based methods retrained on our data:

- **EgoScaler** [51]: A vision-language-based generative framework that synthesizes 6DoF object manipulation trajectories from textual action descriptions and egocentric visual inputs. We adopt their **PointLLM** [45] variant and retrain under our dataset configuration.

Task Prompt: Pick up a new packet of biscuits from the upper shelf. Open the packet and throw the crumpled plastic packaging into the waste bin



GIMO CHOIS EgoScaler EgoFlow GT

Figure 3. **HD-Epic Qualitative Result.** The trajectory in green in each image is the history followed by the respective prediction by the various baselines and the ground truth. We can see that not ony our method generates a plausible trajectory to the end goal, it also takes a rather more natural and smooth path to the target pose.

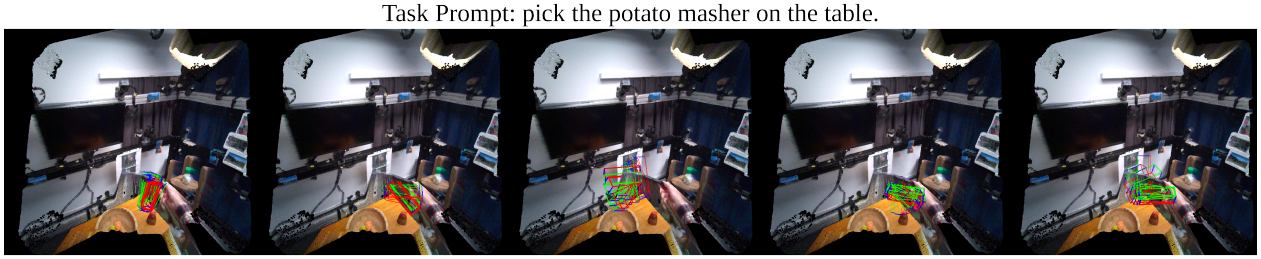
Task Prompt: place the birdhouse toy on the table.



Task Prompt: pick the cellphone on the table.



Task Prompt: pick the potato masher on the table.



GIMO CHOIS EgoScaler EgoFlow Ground Truth

Figure 4. **Hot3D Qualitative Results.** We compare EgoFlow against the established baselines. Our method shows better generalization to unseen conditions and produce geometrically coherent and physically plausible 6DoF trajectories.

- **GIMO** [59]: A Perceiver-based transformer for egocentric human motion forecasting. We replace the human body representation with our object-centric descriptor $\mathbf{x}_t = [\mathbf{p}_t; \mathbf{r}_t] \in \mathbb{R}^9$ and add goal conditioning.
- **CHOIS** [19]: A generative framework for human-object interaction conditioned on object geometry, waypoints, and text. We retain only the object trajectory branch and modify the waypoint conditioning to first 30% of input.
- **M2Diffuser** [48]: A diffusion-based model that generates scene-conditioned whole-body mobile manipulation trajectories from 3D scans, with differentiable physical cost functions enforced during inference.
- **DP3** [53]: A visuomotor imitation learning method that generates robot actions via diffusion conditioned on compact 3D representations extracted from point clouds.
- **SPOT** [15]: An object-centric framework that first generates SE(3) object pose trajectories via diffusion, then uses them to condition a robot action policy, enabling cross-embodiment transfer.
- **ManiFlow** [46]: A visuomotor imitation learning policy

Table 1. **Quantitative Results: HD-EPIC** We compare model performance on various metrics on the HD-EPIC dataset. We can observe that our method performs the best against the baselines, while adding guidance sampling significantly reduces its collision rate. The results are averaged over 3 runs for EgoFlow.

Model	ADE↓	FDE↓	Frechet↓	Geodesic↓	Coll.↓
GIMO [59]	0.285	0.509	0.210	**0.725**	23.5%
CHOIS [19]	0.471	0.755	0.262	1.255	18.7%
Egoscaler [51]	1.330	1.494	0.315	1.614	35.8%
ManiFlow [46]	1.214	1.573	0.609	1.692	35.5%
DP3 [53]	1.298	1.713	0.460	1.734	28.1%
SPOT [15]	1.440	1.795	0.469	1.740	28.7%
M2Diffuser [48]	0.601	0.442	0.476	1.788	8.5%
EgoFlow	**0.279**	**0.102**	**0.197**	1.141	**2.5%**

that combines flow matching with consistency training for efficient action generation from visual, language, and proprioceptive inputs.

Evaluation Metrics. To quantitatively evaluate the quality of the predicted object trajectories, we employ a set of metrics designed to capture positional accuracy, temporal coherence, and physical realism:

- **Average Displacement Error (ADE)**: Computes the mean Euclidean (L2) distance (in metres) between predicted and ground-truth positions over all future time steps, reflecting the overall accuracy of the prediction.
- **Final Displacement Error (FDE)**: Measures the Euclidean distance (in metres) between the predicted and ground-truth positions at the last prediction step, emphasizing the accuracy of the final position.
- **Fréchet Distance** [5]: Quantifies the maximum deviation between two trajectories while accounting for optimal temporal alignment. This metric jointly evaluates spatial and temporal consistency. Smaller values indicate closer adherence of the predicted trajectory to the ground truth in both shape and timing.
- **Geodesic Distance (GD)** [41]: Quantifies the instantaneous angular deviation between two rotations in radians, providing a measure of rotational error in 3D space.
- **Collision Rate**: Represents the proportion of trajectories intersecting with static scene elements, determined via bounding box overlap. Lower values suggest more physically plausible and spatially consistent predictions.

4.2. Realistic Environments

Dataset. We evaluate our framework on the HD-EPIC dataset [25], which provides 41 hours of egocentric recordings captured across nine household kitchens using Project Aria glasses, accompanied by 3D digital twins, dense narrations, and temporally aligned action annotations. However, HD-EPIC offers only sparse object-level annotations at pickup and placement events, making it unsuitable for direct use in continuous trajectory generation.

Preprocessing. We reconstruct continuous 6DoF object trajectories from the sparse annotations by leveraging the interacting hand as a physical proxy. Under the rigid-body coupling assumption during manipulation, we extract 6DoF hand trajectories via Project Aria's Machine Perception Service (MPS) [6] and transfer them to the object reference frame. Hand identity is initialized by proximity to the annotated object position and propagated using contact predictions from Hands23 [3], with a sliding-window filter to ensure temporal consistency. Hand orientation is recovered via a 6D rotation representation derived from the wrist-palm coordinate frame using SVD. We refer readers to GMT [55] and the Supp. Mat. for full preprocessing details.

Table 2. **Cross-dataset evaluation.** Following the setup of [51], we compare EgoFlow against the baselines on HOT3D dataset after training on Ego-Exo4D, thus demonstrating our superior performance on unseen scenes and cross-dataset generalization.

Model	ADE↓	FDE↓	GD↓
GIMO [59]	0.299	0.436	2.06
CHOIS [19]	0.513	0.571	2.46
SPOT [15]	1.018	1.082	2.535
DP3 [53]	1.019	1.096	2.541
M2Diffuser [48]	1.079	1.157	2.525
Egoscaler [51]	0.351	0.540	**0.856**
EgoFlow	**0.265**	**0.027**	1.49

Results. Tab. 1 reports quantitative comparisons on HD-EPIC. Our approach outperforms all baselines across most metrics, demonstrating both spatial and physical consistency. Notably, our model achieves a collision rate of only 2.5%, a substantial improvement over all baselines, demonstrating the effectiveness of differentiable physical constraint enforcement during inference. While GIMO attains the lowest GD, suggesting slightly better orientation alignment, our method maintains a comparable score while ensuring significantly fewer collisions and better endpoint accuracy. These results validate that our framework effectively balances geometric precision and physical feasibility, enabling reliable object trajectory generation in complex real-world scenes. An illustration is shown in Fig. 3 and more in the supplementary.

4.3. Zero-shot Scenarios

Datasets. Following the same setup as EgoScaler [51], we evaluate our framework under zero-shot conditions on two complementary egocentric datasets: Ego-Exo4D [8] and HOT3D [1]. Ego-Exo4D is a large-scale multimodal benchmark capturing both egocentric and exocentric viewpoints of human activities, comprising 1,286 hours of video from 740 participants across 123 real-world environments. We extract sequences involving explicit object motion and align them temporally to obtain dense 6DoF trajectories. HOT3D provides high-precision 3D ground-truth annotations for egocentric object motion, with approximately 833

Table 3. **Input and Guidance Ablation.** Ablation analysis on HD-EPIC assessing input conditioning modalities (top) and gradient guidance costs (bottom).

Model	ADE↓	FDE↓	Frechet↓	Geodesic↓	Coll.↓
w/o $\mathcal{P}$	0.305	0.110	0.205	**1.121**	2.9%
w/o Action	0.330	0.147	0.213	1.168	3.1%
w/o $\mathbf{x}_T$	0.386	0.619	0.239	1.261	3.1%
w/o History	0.405	0.207	0.275	1.293	3.4%
w/o $\mathcal{J}_{\text{vel}}$	0.312	0.128	0.211	1.150	3.5%
w/o $\mathcal{J}_{\text{rot}}$	0.312	0.128	0.211	1.150	3.6%
w/o $\mathcal{J}_{\text{coll}}$	0.310	0.123	0.210	1.142	11.6%
Ours (Full)	**0.278**	**0.102**	**0.197**	1.141	**2.5%**

Table 4. **Architecture Ablation.** Study of different layer configurations (Mamba-Transformer-Mamba layers) on HD-EPIC.

M-T-M Config	ADE↓	FDE↓	Frechet↓	Geodesic↓	Coll.↓
0-12-0	0.440	**0.072**	0.292	1.271	2.9%
1-10-1	0.312	0.108	0.211	**1.138**	3.1%
5-2-5	0.305	0.113	0.207	1.164	3.7%
6-0-6	0.327	0.137	0.206	1.149	4.6%
Ours (3-6-3)	**0.279**	0.102	**0.197**	1.141	**2.5%**

minutes of synchronized multi-view recordings from 19 participants manipulating 33 rigid objects. Following the preprocessing protocol of EgoScaler, we uniformly sample frames at 20 fps to ensure temporal consistency, resulting in 27,788 training samples from Ego-Exo4D and 1,652 test samples from HOT3D. Further dataset statistics and preprocessing details are provided in the supplementary material.

Results. Tab. 2 summarizes the quantitative comparison on the HOT3D zero-shot test set. While Egoscaler [51] attains the lowest GD, its position errors remain substantially higher, suggesting limited spatial consistency despite better orientation alignment. In contrast, our model achieves balanced performance across both translational and rotational metrics, confirming its ability to produce geometrically coherent and physically plausible 6DoF trajectories under unseen conditions, as qualitatively shown in Fig. 4.

4.4. Ablation Study

We conduct an ablation study to validate the effectiveness of various input conditioning and guidance components. As shown in Tab. 3, removing any individual component consistently degrades performance across most evaluation metrics, confirming their complementary roles.

Input Conditioning. Excluding the scene point cloud or end position guidance leads to the largest increase in ADE/FDE, indicating that accurate spatial grounding and goal conditioning are crucial for reliable motion generation. Similarly, removing history inputs significantly harms temporal coherence, resulting in higher geodesic and Frechet distances. The action label also provides beneficial semantic context, helping the model disambiguate motion intent.

Guidance Cost Functions. Among the guidance terms, eliminating velocity smoothness or rotation consistency slightly increases trajectory deviation and collision rates, demonstrating their contribution to producing physically plausible and stable motions. Notably, omitting collision avoidance results in a sharp rise in collision rate (from 2.5% $\rightarrow$ 11.6%), implying that incorporating this term reduces collisions by about 79%, thus emphasizing its importance for generating physically feasible trajectories.

Layer Configurations. In Tab. 4, we investigate the impact of layer distribution across the three stages of our Motion Generation Model. We compare hybrid Mamba-Transformer configurations against pure architectures: a Transformer-only model (0-12-0) and a Mamba-only model (6-0-6). Our balanced 3-6-3 configuration achieves the best overall performance across all metrics.

5. Conclusion

In this work, we introduced EgoFlow, a generative framework that unifies flow matching in $\mathbb{R}^9$ with gradient-guided inference for physically grounded egocentric trajectory generation. By integrating long-horizon temporal modeling through a hybrid Mamba–Transformer–Perceiver architecture and embedding differentiable physical priors directly into the sampling process, EgoFlow achieves a rare balance between realism, controllability, and efficiency.

Extensive experiments across egocentric benchmarks demonstrate that EgoFlow consistently outperforms diffusion-based and transformer-only baselines in both spatial accuracy and physical plausibility, while generalizing effectively to unseen domains without fine-tuning. The results suggest that flow-based motion generation can serve as a strong alternative to diffusion models for continuous 3D prediction tasks.

Looking forward, we see EgoFlow as a step toward holistic embodied scene understanding, where perception, motion, and interaction are modeled in a unified generative framework. Future directions include extending our approach to deformable objects, integrating closed-loop sensory feedback, and coupling EgoFlow with robot policy learning to enable goal-directed, physically consistent action generation.

Limitations Our current formulation handles diverse objects including semi-deformable items (clothes, paper), but doesn't model volumetric deformations or shape changes during its motion. We also assume static environments with known geometry, which limits applicability in scenes with moving agents or dynamic clutter. Addressing these limitations presents promising directions for future work.

Acknowledgments This research was partially funded by the German Federal Ministry of Education and Research through the ExperTeam4KI fund-

ing program for UDance (Grant No. 16IS24064).

References

[1] 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 Computer Vision and Pattern Recognition Conference*, pages 7061–7071, 2025. 1, 2, 7

[2] Mingfei Chen, Yifan Wang, Zhengqin Li, Homanga Bharadhwaj, Yujin Chen, Chuan Qin, Ziyi Kou, Yuan Tian, Eric Whitmire, Rajinder Sodhi, Hrvoje Benko, Eli Shlizerman, and Yue Liu. Flowing from reasoning to motion: Learning 3d hand trajectory prediction from egocentric human interaction videos, 2025. 3

[3] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. *Advances in Neural Information Processing Systems*, 36:30453–30465, 2023. 7

[4] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 3

[5] Efrat, Guibas, Sariel Har-Peled, and Murali. New similarity measures between polylines with applications to morphing and polygon sweeping. *Discrete & Computational Geometry*, 28(4):535–569, 2002. 7

[6] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1, 7

[7] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, Youqiang Zhang, and Junshi Huang. Dimba: Transformer-mamba diffusion models. *arXiv preprint arXiv:2406.01159*, 2024. 3

[8] 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*, pages 19383–19400, 2024. 1, 2, 7

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 2, 3

[10] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[11] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 3

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 2

[13] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se(3) pose trajectory diffusion for object-centric manipulation, 2025. 6, 7

[16] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 5

[17] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2, 5

[18] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022. 5

[19] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024. 2, 6, 7

[20] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2, 3, 4

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2, 4

[22] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024. 1

[23] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 1

[24] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[25] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913, 2025. 1, 2, 7

[26] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, et al. EgoVerse: An egocentric human dataset for robot learning from around the world, 2026. 3

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4

[28] Inigo Quilez. Distance functions. https://iquilezles.org/articles/distfunctions/, 2008. Accessed: 2025-11-10. 5

[29] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar Parkhi. Egoblur model, 2023. 1

[30] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1

[31] James C Robinson. *An introduction to ordinary differential equations*. Cambridge University Press, 2004. 4

[32] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *arXiv preprint arXiv:2309.06440*, 2023. 4

[33] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 3

[34] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[36] Rustin Soraki, Homanga Bharadhwaj, Ali Farhadi, and Roozbeh Mottaghi. Objectforesight: Predicting future 3d object trajectories from human videos, 2026. 3

[37] Yao Teng, Yue Wu, Han Shi, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Dim: Diffusion mamba for efficient high-resolution image synthesis. *arXiv preprint arXiv:2405.14224*, 2024. 3

[38] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2

[39] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012. 4

[40] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023. 3

[41] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3109–3118, 2015. 7

[42] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023. 4

[43] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11176–11186, 2025. 2

[44] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 1

[45] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024. 5

[46] Ge Yan, Jiyue Zhu, Yuquan Deng, Shiqi Yang, Ri-Zhao Qiu, Xuxin Cheng, Marius Memmel, Ranjay Krishna, Ankit Goyal, Xiaolong Wang, and Dieter Fox. Maniflow: A dexterous manipulation policy via flow matching. *arXiv preprint arXiv:*, 2025. 6, 7

[47] Sixu Yan, Zeyu Zhang, Muzhi Han, Zaijin Wang, Qi Xie, Zhitian Li, Zhehan Li, Hangxin Liu, Xinggang Wang, and Song-Chun Zhu. M 2 diffuser: Diffusion-based trajectory optimization for mobile manipulation in 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 5

[48] Sixu Yan, Zeyu Zhang, Muzhi Han, Zaijin Wang, Qi Xie, Zhitian Li, Zhehan Li, Hangxin Liu, Xinggang Wang, and Song-Chun Zhu. M2diffuser: Diffusion-based trajectory optimization for mobile manipulation in 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 6, 7

[49] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1

[50] Yuhang Yang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37: 54529–54557, 2024. 2

[51] 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. 2, 5, 7, 8

[52] Kevin Zakka. Mink: Python inverse kinematics based on MuJoCo, 2025. 4

[53] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 6, 7

[54] Huajian Zeng, Lingyun Chen, Jiaqi Yang, Yuantai Zhang, Fan Shi, Peidong Liu, and Xingxing Zuo. Flowhoi: Flow-based semantics-grounded generation of hand-object interactions for dexterous robot manipulation. *arXiv preprint arXiv:2602.13444*, 2026. 3

[55] Huajian Zeng, Abhishek Saroha, Daniel Cremers, and Xi Wang. GMT: Goal-conditioned multimodal transformer for 6-dof object trajectory synthesis in 3d scenes. In *International Conference on 3D Vision (3DV)*, 2026. 2, 3, 4, 5, 7

[56] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 2

[57] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8521–8531, 2024. 2

[58] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 2, 4

[59] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 2, 6, 7

[60] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 3, 4

[61] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3