GMT: Goal-Conditioned Multimodal Transformer for 6-DOF Object Trajectory Synthesis in 3D Scenes

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

Abstract

Synthesizing controllable 6-DOF object manipulation trajectories in 3D environments is essential for enabling robots to interact with complex scenes, yet remains challenging due to the need for accurate spatial reasoning, physical feasibility, and multimodal scene understanding. Existing approaches often rely on 2D or partial 3D representations, limiting their ability to capture full scene geometry and constraining trajectory precision. We present GMT, a multimodal transformer framework that generates realistic and goal-directed object trajectories by jointly leveraging 3D bounding box geometry, point cloud context, semantic object categories, and target end poses. The model represents trajectories as continuous 6-DOF pose sequences and employs a tailored conditioning strategy that fuses geometric, semantic, contextual, and goal-oriented information. Extensive experiments on synthetic and real-world benchmarks demonstrate that GMT outperforms state-of-the-art human motion and human-object interaction baselines, such as CHOIS and GIMO, achieving substantial gains in spatial accuracy and orientation control. Our method establishes a new benchmark for learning-based manipulation planning and shows strong generalization to diverse objects and cluttered 3D environments.

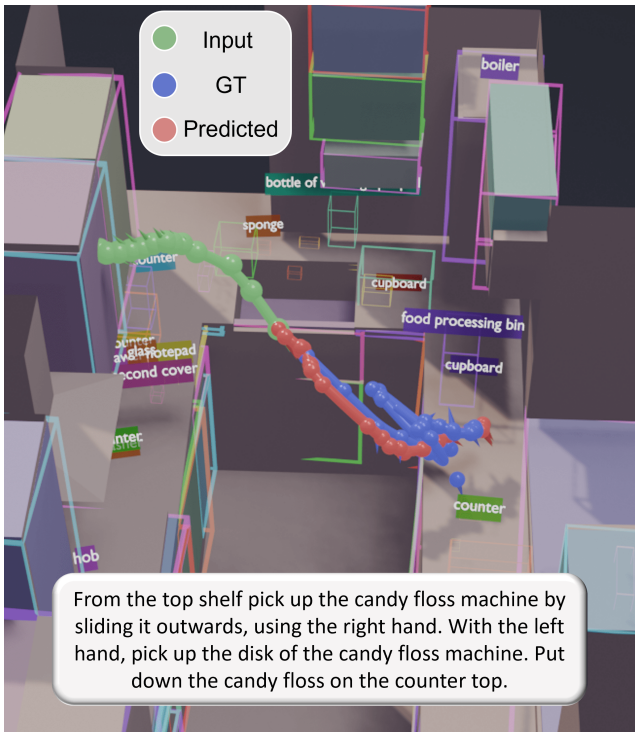


Figure 1. Given an observed trajectory, scene context, and action description, our model predicts plausible future 6-DOF object trajectories. The generated trajectories are more efficient than natural human motions.

1. Introduction

Generating realistic and controllable 6-DOF object manipulation trajectories in 3D environments is a central challenge in robotics and computer vision [2, 47]. In manipulation tasks, the object trajectory is often closely aligned with the end-effector trajectory of the robot. Given such a trajectory in Cartesian space, inverse kinematics (IK) [5] can be used to compute the corresponding joint configurations, thereby converting the end-effector path into a full sequence of robot arm motions. This trajectory-centric formulation decouples perception from control [21, 43], allowing flexible integration of downstream planners or controllers and facilitating generalization across tasks and platforms.

However, synthesizing such trajectories in cluttered 3D scenes remains challenging. First, accurate 3D perception is difficult: depth sensors suffer from noise, occlusion, and sparsity [40, 55, 59], where object centers may lie far from any captured surface points [41]. Second, generated trajectories must respect spatial constraints and physical plausibility, avoiding collisions, maintaining stability, and aligning with object affordances [3, 15, 53]. Third, goal-conditioned generation requires integrating geometry, semantics, context, and target poses. Traditional planners [25, 27] face high computational cost in high-dimensional spaces, while most learning-based ap-

* These authors contributed equally.

proaches [4, 28] predict low-level actions end-to-end, limiting explicit trajectory control or physical constraint injection.

These challenges highlight the need for generative models that can capture long-range dependencies, integrate heterogeneous modalities, and enforce structured constraints during synthesis. Recent advances in transformers [49] and diffusion-based generative models [19, 61] have demonstrated precisely these capabilities, excelling at modeling complex spatial-temporal structures in high-dimensional spaces. Nevertheless, applications have focused primarily on human motion [17, 60] or human-object interaction. In particular, Human-Object Interaction (HOI) models typically treat objects as passive entities, with object trajectories induced indirectly by human motion [14, 30]. These approaches [32, 57] inject HOI behaviors into simulators and rely on reinforcement learning to refine them into executable policies. While effective for human-centric skill transfer, this pipeline restricts the generative model's flexibility and hinders cross-embodiment generalization: the learned policy is tied to specific morphologies and training simulators rather than abstract object dynamics.

Our work takes a different perspective: we shift the focus from human-centric HOI to *object-centric trajectory generation*. By directly modeling 6-DOF object trajectories, we treat objects as primary dynamic entities conditioned on scene and goal constraints. This design allows generated trajectories to serve as a universal intermediate representation: through IK, they can be instantiated by arbitrary robotic embodiments, enabling cross-platform transfer without simulator-dependent policy learning.

In this work, we address these gaps with a multimodal transformer framework for controllable 6-DOF object trajectory synthesis in 3D scenes. Our model jointly leverages geometric, semantic, contextual, and goal information to produce spatially consistent and physically plausible trajectories that can be directly executed on robotic systems via IK.

In summary, our main contributions are:

- **GMT, a multimodal transformer architecture** for 6-DOF object trajectory generation, unifying scene geometry, semantics, and task goals within a single framework.
- **A tailored fusion strategy** integrating: (i) geometric conditioning via feature propagation from scene point clouds to 3D bounding box corners; (ii) semantic conditioning via hierarchical category embeddings [42]; (iii) contextual conditioning through global scene features; and (iv) goal conditioning via learnable end-pose embeddings.
- **Extensive experiments** on challenging 3D manipulation benchmarks, achieving state-of-the-art performance over strong human motion baselines such as CHOIS [30] and GIMO [62], with substantial gains in spatial accuracy and orientation control.

2. Related Work

Object trajectory synthesis lies at the intersection of computer vision, motion modeling, and 3D scene understanding. While these areas have achieved notable progress, synthesizing controllable 6-DOF object motion in cluttered environments remains underexplored. Below we review related directions and position our work accordingly.

2.1. Video Prediction & Dynamics Learning

Video prediction models aim to forecast object dynamics directly in pixel space. Early methods such as PredNet [31] and ConvLSTM [45] learned short-term temporal dependencies, while Interaction Networks [1] and Visual Interaction Networks [50] introduced relational reasoning between objects. More recent efforts leverage transformers for long-horizon forecasting [51, 52] or diffusion models for stochastic video generation [18, 20].

These works highlight the importance of modeling dynamics but operate in image space, which struggles with depth ambiguity, occlusion, and 3D consistency. Video generative models such as Sora [34, 65] achieve impressive visual fidelity and exhibit emergent properties like object permanence, yet they often lack explicit physical understanding and fail to support planning or decision-making. Similarly, frameworks treating videos as "world models" are hindered by the absence of explicit state-action structure and limited controllability [29, 54]. In contrast, we generate explicit 6-DOF object trajectories in 3D space, enabling precise control over motion and direct interaction with the environment. Our approach focuses on synthesizing physically plausible trajectories that respect spatial constraints, rather than predicting pixel-level dynamics.

2.2. Human Motion & Interaction Synthesis

Human motion synthesis has advanced rapidly, spanning text-conditioned generation [16, 48], scene-aware prediction [62], and diffusion-based motion priors [17, 60]. Human-object interaction models further integrate semantics and contact reasoning: CHOIS [30] generates synchronized HOI from language prompts, while differentiable simulation [14] enforces physical plausibility.

Recently, diffusion-based methods have advanced HOI synthesis: CG-HOI [11] explicitly models human-object contact in a joint diffusion framework, improving physical coherence; InterDiff [56] introduces physics-informed correction within diffusion steps for long-term HOI predictions; HOI-Diff [37] utilizes a dual-branch diffusion model plus affordance correction to generate diverse yet coherent human-object motions from text prompts.

Despite these strengths, all of these approaches remain fundamentally human-centric—modeling object motion only as a response to human behavior. In contrast, our

work shifts the focus to object-centric trajectory generation, treating objects as primary dynamic entities conditioned on scene and goal constraints. This enables trajectories to be executed via inverse kinematics across robots of varying morphology, rather than being limited to humanoid embodiments.

2.3. Scene Understanding & Geometric Reasoning

Effective motion synthesis requires efficient scene representation. Point cloud methods provide detailed geometry but impose computational constraints. PointNet++ [40] addresses some limitations through hierarchical feature learning on point sets in metric spaces, but still faces computational challenges in dense environments. Voxel representations [33, 63] improve efficiency but sacrifice resolution needed for precise manipulation. Recent fully sparse approaches like VoxelNeXt [8] eliminate sparse-to-dense conversion requirements while maintaining detection performance.

Recent work suggests that coarser representations can be sufficient for many tasks [10]. 3D-BoNet [58] demonstrates that direct bounding box regression can be more computationally efficient than existing approaches by eliminating post-processing steps such as non-maximum suppression, feature sampling, and clustering. This key insight, that high fidelity is not always necessary, suggests that bounding boxes provide sufficient geometric information for trajectory synthesis while enabling real-time performance.

Multimodal fusion architectures enable flexible combination of geometric and semantic information. Perceiver [23] provides a scalable blueprint, while Perceiver IO [22] extends this with flexible querying mechanisms for structured inputs and outputs. SUGAR [7] demonstrates effective multimodal pre-training for robotics through joint cross-modal knowledge distillation. The key insight from robotics applications [46] is that fusion must respect constraint hierarchies: hard geometric constraints should dominate soft semantic preferences to ensure physically valid output. Our framework builds on these insights by using 3D bounding boxes as a compact yet expressive representation, and enforcing a fusion hierarchy where hard geometric constraints dominate semantic cues.

Furthermore, spatial reasoning in cluttered environments increasingly benefits from hybrid symbolic geometric approaches, where discrete scene graphs capture semantic relations while continuous modules preserve metric precision [24]. This dual representation allows agents to reason over both affordances and spatial feasibility, bridging perception and action.

3. Methodology

Our goal is to synthesize controllable 6-DOF object trajectories in 3D scenes, conditioned on observed motion, scene context, and a target goal state. The central challenge is to fuse heterogeneous modalities: geometry, semantics, and dynamics into a single representation that preserves physical plausibility and goal consistency. Naively concatenating features or relying on a single modality (e.g., raw point clouds) caused unstable training and implausible motion (e.g., interpenetration, drifting). We design a multimodal transformer with three key insights: (1) *spatial feature propagation* is a compact yet stable spatial abstraction compared to dense point features; (2) *vision language semantics* (CLIP) transfer behavior patterns across action description or categories better than one-hot labels; and (3) *hierarchical fusion* that prioritizes hard geometric constraints over softer semantic cues significantly reduces collision and goal drift. An overview is shown in Fig. 2.

3.1. Problem Formulation

We formulate trajectory prediction as a conditional sequence modeling problem.

Let the input history trajectory be denoted as $\mathbf{X}_{1:H} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_H\} \in \mathbb{R}^{H\times 9}$, where each $\mathbf{x}_i = (\mathbf{p}_i, \mathbf{r}_i)$ consists of a 3D position $\mathbf{p}_i \in \mathbb{R}^3$ and a 6D continuous rotation representation $\mathbf{r}_i \in \mathbb{R}^6$ [64]. Each object trajectory is associated with a fixed object category c and size $s \in \mathbb{R}^3$.

The scene context is represented as $\mathbf{S} = (\mathbf{P}, \mathbf{B})$, where $\mathbf{P} \in \mathbb{R}^{N\times 3}$ denotes the scene point cloud with N points, $\mathbf{B} = \{(l_k, b_k)\}_{k=1}^{M}$ is defined as the set of M semantic fixture bounding boxes, with each l_k being the semantic label, and $b_k = (\mathbf{c}_k, \mathbf{s}_k, \mathbf{r}_k) \in \mathbb{R}^{12}$ containing the 3D center, size, and 6D rotation representation. The goal condition $\mathbf{G} \in \mathbb{R}^9$ specifies the target object state. The output trajectory to be predicted is $\hat{\mathbf{X}}_{H+1:T} = \{\hat{\mathbf{x}}_{H+1}, \hat{\mathbf{x}}_{H+2}, \ldots, \hat{\mathbf{x}}_H\}$, where $T - H$ is the prediction horizon. In our particular setup, we use 30% of the trajectory as the input history to predict the remaining 70%. Thus, the task is to learn a conditional distribution $P(\hat{\mathbf{X}}_{H+1:T} \mid \mathbf{X}_{1:H}, \mathbf{G}, \mathbf{S})$, which generates physically plausible and semantically consistent future trajectories, aligned with the specified goal and conditioned on multimodal scene context.

3.2. Multimodal Scene Encoding

Generating plausible object trajectories requires a comprehensive understanding of both spatial and semantic context. Specifically, the model must capture (1) object motion patterns, (2) spatial arrangements, and (3) environmental constraints, including collision and interaction dynamics. To this end, we construct dedicated representations for each scene modality.

Trajectory Feature. We observe that using trajectory geometry alone often leads to overfitting, as the model fails to capture category-specific motion patterns. To address this, we couple trajectory embeddings with semantic category features.

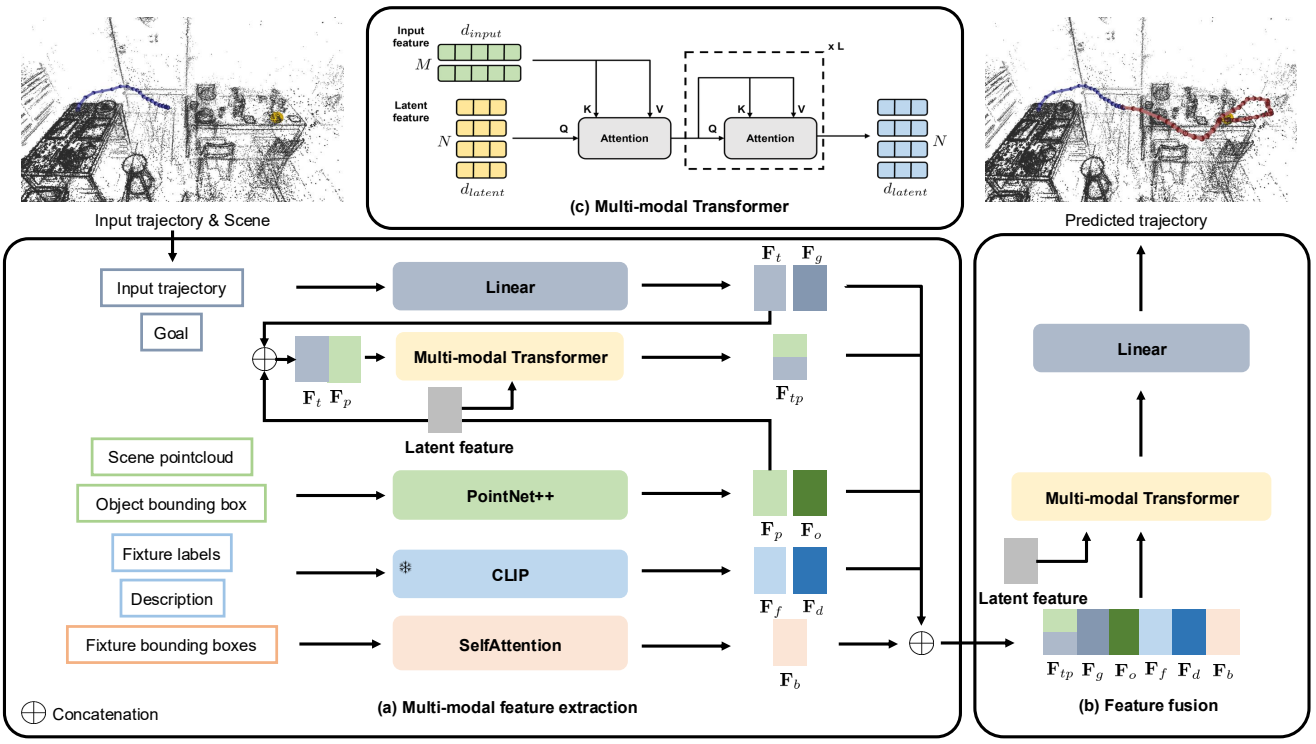


Figure 2. **Pipeline overview.** Given an observed trajectory and scene context, our model predicts future 6-DOF object trajectories conditioned on a specified goal state. We encode (a) trajectory dynamics, (b) local geometry propagated from the scene point cloud to the object's bounding box, (c) semantic fixture boxes and labels, (d) natural language description of the action (e) a goal descriptor. A multi-modal transformer performs hierarchical fusion that emphasizes geometric feasibility before semantic preferences. The fused latent is fed directly to the prediction head (no separate decoding stage), which we found more stable for long-horizon control.

The temporal motion context $\mathbf{F}_t$ is obtained by passing the observed trajectory sequence through a linear layer, yielding an embedding suitable for multimodal fusion.

Spatial Feature. To account for environmental constraints such as floors, walls, or tabletops, the model must capture both global and local spatial cues. Directly concatenating a global scene feature $\mathbf{F}_o$ is insufficient, as it fails to distinguish spatially distinct regions (e.g., floor vs. tabletop) and introduce irrelevant noise (e.g., clutter on the ground). Instead, we encode the raw point cloud $\mathbf{P}$ using PointNet++ [40], producing both a global feature $\mathbf{F}_o$ and local features $\mathbf{F}_l$. To provide trajectory features with awareness of their local surroundings, we propagate per-point local features from the scene point cloud to the object's bounding box at each observed timestep. Specifically, we interpolate features from the k nearest neighbors using inverse-distance weighting [40]:

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

where $\mathbf{c}_t$ denotes the center of the object bounding box at time t, and $\mathbf{p}_i$ and $\mathbf{f}_i$ are the coordinates and features of the i-th point, respectively. By repeating this process over all observed timesteps $t = 1, \ldots, H$, we obtain a temporal sequence of local geometric features $\mathbf{F}_p = \{\mathbf{F}_p^1, \mathbf{F}_p^2, \ldots, \mathbf{F}_p^H\}$, which are subsequently incorporated in the fusion stage to enrich trajectory representations with spatial context.

Furthermore, relying solely on point-cloud features is insufficient to fully capture the interaction constraints between the moving object and nearby static fixtures. The spatial extents of these fixtures can be reliably obtained using modern instance segmentation approaches [6, 26, 44]. Direct concatenation of their embeddings, however, fails to model inter-object relationships and may result in physically implausible predictions (e.g., a chair penetrating a tabletop). To more explicitly encode such interaction constraints, we apply a multi-head self-attention module over the set of fixture bounding boxes:

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

Semantic Feature. Geometry alone is insufficient to distinguish objects with similar sizes but different affordances (e.g., desk vs. bed). Semantic information is incorporated by embedding fixture labels l_k and the natural language de-

scription of the action/object category name d involving the object with a frozen CLIP encoder [42], followed by projection into the feature space:

$$\mathbf{F}_f = \text{Proj}(\text{CLIP}(l_k)), \mathbf{F}_d = \text{Proj}(\text{CLIP}(d)) \quad (3)$$

To mitigate semantic noise, only the K nearest fixtures (based on center distance) are retained before applying attention, as distant objects contribute little and increase variance.

Goal Feature. The target object state $\mathbf{G}$, representing the desired future position and orientation, is encoded via a linear layer and projected into the same feature space:

$$\mathbf{F}_g = \text{Proj}(\text{Linear}(\mathbf{G})). \quad (4)$$

Note that this goal plays two complementary roles. First, it serves as a high-level intention variable that disambiguates between otherwise plausible futures under the same history and scene, e.g., "place the object on the tabletop" versus "place it back on the floor". Second, it provides a controllable knob at inference time: given a fixed observed trajectory $\mathbf{X}_{1:H}$ and scene $\mathbf{S}$, different choices of $\mathbf{G}$ induce qualitatively different yet physically valid futures. By making $\mathbf{G}$ a first-class conditioning signal, our formulation bridges trajectory prediction and goal-directed planning, enabling the synthesis of future motions that are not only plausible and scene-consistent but also explicitly aligned with user-specified targets.

3.3. Multimodal Feature Fusion

Multi-modal Transformer. Naively concatenating features across modalities can cause scale imbalance, leading the model to over-rely on certain inputs. To achieve balanced and flexible integration, we adopt a Transformer-based fusion module inspired by Perceiver IO [23]. This design introduces a learnable latent array that acts as an information bottleneck, ensuring scale normalization across modalities and enabling modality-agnostic fusion.

Given a collection of modality-specific input tokens $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, where $\mathbf{X} \in \mathbb{R}^{M \times d_{\text{in}}}$, the fusion module maintains a learnable latent array $\mathbf{Z}_0 \in \mathbb{R}^{N \times d_{\text{latent}}}$, with $N \ll M$ and d_{latent} denoting the latent feature dimension.

The fusion process is implemented via stacked cross-attention and latent self-attention blocks. At each layer ℓ, the latent array is first updated by attending to the input tokens:

$$\mathbf{Z}'_\ell = \text{CrossAttn}(\mathbf{Z}_{\ell-1}, \mathbf{X}) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_K}}\right)V, \quad (5)$$

where $Q = \mathbf{Z}_{\ell-1}W_q$, $K = \mathbf{X}W_k$, and $V = \mathbf{X}W_v$ are linearly projected query, key, and value matrices, respectively.

Next, latent self-attention and a feed-forward network are applied to propagate information among latent slots:

$$\mathbf{Z}_\ell = \text{FFN}\left(\text{SelfAttn}(\mathbf{Z}'_\ell)\right), \quad (6)$$

where SelfAttn follows the same formulation as Eq. 5, but operates only on latent tokens.

After L layers, the final latent representation $\mathbf{Z}_L$ serves as the fused multimodal embedding. Unlike Perceiver IO [23], our architecture does not include a second decoding stage; instead, $\mathbf{Z}_L$ is directly used as input to the trajectory prediction head.

Scene-aware Trajectory Augmentation. To endow trajectory features with spatial awareness, we fuse $\mathbf{F}_t$ with the propagated local geometry $\mathbf{F}_p^t$ through the multimodal transformer:

$$\mathbf{F}_{tp} = \text{MultiTrans}\big(\text{Concat}(\mathbf{F}_t, \mathbf{F}_p^t)\big). \quad (7)$$

Direct concatenation without attention underperformed, indicating that attention-based alignment is necessary to resolve frame and scale ambiguities.

Semantic Geometric Fusion. While the trajectory features are enhanced with 3D spatial information, incorporating semantic cues is crucial for guiding trajectory generation, as the prediction network must understand object semantics and their relations (e.g., a chair should not move through a table). To this end, we project the semantic features $\mathbf{F}_f$ and $\mathbf{F}_d$, the spatial features $\mathbf{F}_b$, the global point cloud feature $\mathbf{F}_o$, and the goal state feature $\mathbf{F}_g$ into the same dimension as the fused trajectory feature $\mathbf{F}_{tp}$ via linear layers. These features are then concatenated to form a comprehensive multimodal representation:

$$\mathbf{F}_{\text{fuse}} = \text{Concat}(\mathbf{F}_{tp}, \mathbf{F}_f, \mathbf{F}_d, \mathbf{F}_b, \mathbf{F}_o, \mathbf{F}_g). \quad (8)$$

Finally, the predicted trajectory $\hat{\mathbf{X}}_{H:T}$ is conditioned on the fused feature $\mathbf{F}_{\text{fuse}}$ through the multimodal transformer described above.

3.4. Training Objective

Our training objective is designed to supervise both the future prediction accuracy and the reconstruction quality of the observed motion. The total loss comprises four terms: translation loss, orientation loss, reconstruction loss, and destination loss for both translation and orientation. Each term captures a specific aspect of the prediction quality.

Given the model's prediction $\hat{\mathbf{X}} \in \mathbb{R}^{T \times 9}$ and the ground truth trajectory $\mathbf{X} \in \mathbb{R}^{T \times 9}$, where each frame contains 3D position and 6D rotation representation, we first split each sequence into history and future segments based on predefined input and predict ratios. Let T_{hist} and T_{fut} denote the number of historical and future steps, respectively.

To supervise the *future prediction*, we compute an L_1 loss between the predicted and ground truth values for both

translation and rotation components:

$$\mathcal{L}_{\text{trans}} = \frac{1}{T_{\text{fut}}} \sum_{t \in \text{future}} \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_1 , \tag{9}$$

$$\mathcal{L}_{\text{ori}} = \frac{1}{T_{\text{fut}}} \sum_{t \in \text{future}} \|\hat{\mathbf{r}}_t - \mathbf{r}_t\|_1 , \tag{10}$$

where $\mathbf{p}_t \in \mathbb{R}^3$ and $\mathbf{r}_t \in \mathbb{R}^6$ denote the ground truth position and rotation at timestep t, and $\hat{\mathbf{p}}_t$, $\hat{\mathbf{r}}_t$ are their corresponding predictions.

To preserve fidelity in the *observed segment*, a reconstruction loss is applied to the historical frames:

$$\mathcal{L}_{\text{rec}} = \frac{1}{T_{\text{hist}}} \sum_{t \in \text{history}} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_1 , \tag{11}$$

where $\hat{\mathbf{x}}_t$ is the full 9D predicted pose at timestep t.

Additionally, we introduce a *destination loss* to explicitly constrain the model's final predicted pose to match the last valid ground truth frame:

$$\mathcal{L}_{\text{dest}} = \|\hat{\mathbf{x}}_{T_{\text{end}}} - \mathbf{x}_{T_{\text{end}}}\|_1 , \tag{12}$$

The final loss is a weighted sum of the above terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{dest}}\mathcal{L}_{\text{dest}}, \tag{13}$$

where $\lambda_{\text{trans}}, \lambda_{\text{ori}}, \lambda_{\text{rec}}, \lambda_{\text{dest}}$ are hyperparameters controlling the contribution of each loss component.

4. Experiments

We start this section by describing the experimental setup, including datasets and metrics. We then present results on two datasets, followed by an ablation study to analyze the contributions of different components in our method.

4.1. Experiments Setup

Baseline. To provide comparative insights, we adapt two state-of-the-art human motion prediction methods for the object manipulation trajectory generation task:

- **GIMO** [62]: A transformer-based model originally designed for egocentric human-object interaction forecasting. It leverages a unified Perceiver-inspired architecture to fuse geometry, object category, and semantic scene context for predicting 6-DOF human motion. We repurpose GIMO to predict object trajectories by replacing the human body input with object-specific motion and geometry. Since our object-centric prediction task does not contain gaze information, we disable the gaze branch and remove all gaze-related modules in both training and inference.
- **CHOIS** [30]: This generative framework produces human-object interaction sequences conditioned on object geometry, sparse object waypoints, and textual instructions. In our implementation, the human-motion branch is deactivated, preserving only the object trajectory prediction component. The waypoint conditioning is restricted to the initial 30% of the input and goal for fair comparison. To ensure compatibility with our deterministic prediction framework, we disable the diffusion-based sampling mechanism and utilize only the transformer backbone for direct single-step prediction.

Metrics. We evaluate predicted object trajectories using the following quantitative metrics to assess accuracy, temporal consistency, and physical plausibility:

- **Average Displacement Error** (ADE): Measures the mean L2 distance between the predicted and ground truth object positions across all future time steps.
- **Final Displacement Error** (FDE): L2 distance between the predicted and ground truth position at the final future time step.
- **Fréchet Distance** [12] (FD): Measures the maximum deviation between two trajectories over time by considering the best possible alignment along the temporal axis. It is sensitive to both spatial proximity and temporal consistency. A smaller Fréchet distance indicates that the predicted trajectory closely follows the shape and timing of the ground truth, whereas a large value indicates temporal mismatch or outlier deviations.
- **Angular Consistency** (AC): Measures how well the directional dynamics of the predicted trajectory align with the reference sequence. The positional differences between consecutive frames are represented as direction vectors, and the mean cosine similarity between these vectors quantifies the preservation of orientation trends and motion smoothness. Higher values indicate better temporal coherence and directional stability.
- **Collision Rate** (CR): The fraction of predicted trajectories that result in collisions with surrounding fixtures, computed based on the intersection of predicted bounding boxes and static scene geometry. Lower collision rates indicate better physical plausibility and spatial awareness of the model.

4.2. Controlled Idealized Scenarios

Dataset. We first evaluate our method on the Aria Digital Twin (ADT) dataset [35], which offers high-fidelity recordings of human-object interactions in a fully controlled 3D simulation environment. The sequences are captured under noise-free conditions with complete visibility and accurate tracking, providing an ideal setting to assess the upper-bound performance of our model under perfect geometric and semantic observations. To align with our task objective, we exclude all ADT sequences involving recognition-only

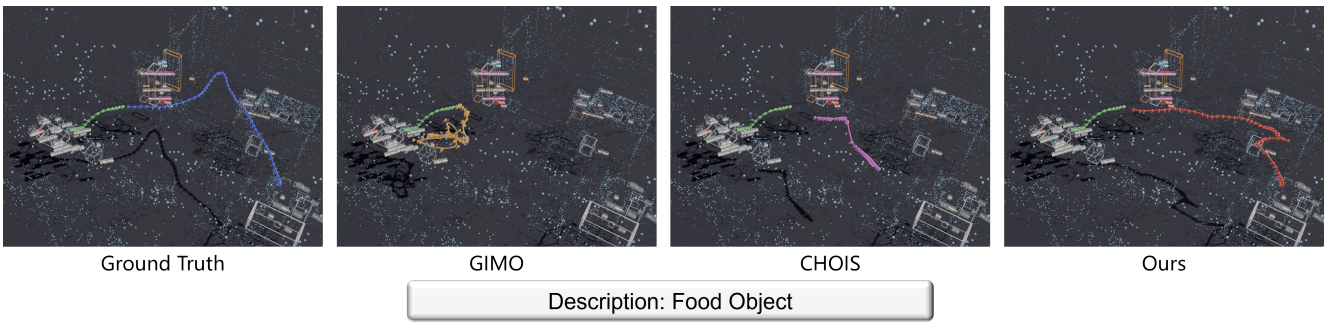


Figure 3. **Qualitative results on the ADT dataset.** The green trajectory represents the input history across all experiments. Only our model produces trajectories that both reach the target and avoid collisions, while also achieving shorter path lengths compared to the ground-truth natural trajectories. Adaptive GIMO fails due to the absence of gaze information, whereas CHOIS accumulates errors over time, ultimately leading to failure.

Method	ADE[m] ↓	FDE[m] ↓	FD[m] ↓	AC[m] ↑	CR ↓
GIMO [62]	0.982	1.401	1.511	0.140	19.6%
CHOIS [30]	0.853	1.062	1.209	0.283	**9.3%**
GMT (Ours)	**0.366**	**0.072**	**0.438**	**0.402**	13.1%

Table 1. Quantitative results on the ADT dataset. Our method outperforms both baselines in all metrics except collision rate. Note that collision rate is only meaningful when FDE is also low; otherwise, it may decrease due to trivial predictions such as static forecasts.

Method	ADE[m] ↓	FDE[m] ↓	FD[m] ↓	AC[m] ↑	CR ↓
GIMO [62]	0.411	0.654	0.780	0.002	11.8%
CHOIS [30]	0.446	0.589	0.760	0.009	12.0%
GMT (Ours)	**0.283**	**0.034**	**0.391**	**0.037**	**10.3%**

Table 2. Quantitative results on the HD-EPIC dataset. Ours achieves the best performance across all metrics, demonstrating superior trajectory prediction accuracy and robustness in real-world scenarios.

interactions (i.e., without significant object displacement). We train our model and the baselines on a randomly selected subset of 228 sequences and evaluate them on the remainder.

Results. As shown in Tab. 1, our method consistently outperforms both adapted baselines across all evaluation metrics except for the collision rate, with significant gains in spatial accuracy (ADE/FDE). Notably, it achieves the lowest Fréchet distance and highest angular consistency, indicating superior alignment with the ground-truth in both position and orientation. While a lower collision rate can be trivially achieved by predicting static trajectories, only our model attains both low FDE and a low collision rate, demonstrating its ability to generate plausible and physically meaningful object motions. An illustration is shown in Fig. 3 and more in the supplementary.

4.3. Realistic Challenging Scenarios

Dataset. To assess the robustness of our model in real-world settings, we evaluate it on the HD-EPIC dataset [38]. HD-EPIC contains 41 hours of egocentric videos of human-object interactions recorded in natural household environments. A key challenge in leveraging the HD-EPIC dataset for our task is the sparsity of its annotations. The dataset provides object positions for pickup and drop events, but does not provide dense, frame-by-frame object trajectory sequences between these points. To bridge this gap, we use the interacting hand as a proxy for computing the object's motion. Our key assumption is that between the moments of pickup and release, the hand and object are physically coupled and move together. By tracking the hand's motion using Project Aria's Machine Perception Service (MPS) [13], we can accurately infer the object's trajectory during this unannotated phase.

Our process begins by identifying the primary hand interacting with the object. At the start of each sequence, we use the dataset's initial 3D object position to select the hand with the minimum Euclidean distance to the annotated object position. For all subsequent frames until the drop-off, this choice is propagated by identifying which hand is in physical contact, as determined by using a pretrained Hands23 [9] detector. A sliding window filter is then applied to this sequence of hand selections to ensure better temporal consistency and remove flickering.

Finally, we process the motion of this consistently tracked hand to create a stable trajectory. We compute a 6D rotation for the hand's orientation by constructing an orthonormal basis from its normal vector of the wrist-palm using singular value decomposition (SVD). This refined 6-DOF hand trajectory is then directly transferred to the annotated target object, yielding a clean and realistic manipulation sequence.

Results. Tab. 2 summarizes the quantitative results. Our

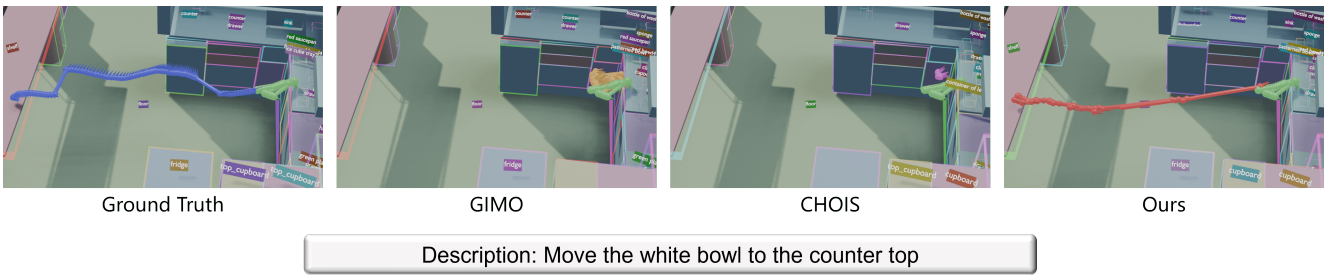


Figure 4. **Qualitative results on the HD-EPIC dataset.** Across all examples, the green points indicate the input history. Our model generates trajectories that are more efficient than the ground truth, while all baselines remain stuck in repetitive motions.

model consistently and significantly outperforms all baselines across all metrics. Although the HD-EPIC dataset poses greater challenges due to occlusions and sensor noise, its substantially larger scale ($\sim$20$\times$ trajectories than ADT) and more diverse scenes help compensate for these difficulties and allow the model to maintain strong performance. Examples can be found in Fig. 4 and in the supplementary.

4.4. Ablation Study

Variant	ADE[m] ↓	FDE[m] ↓	FD[m] ↓	AC[m] ↑	CR ↓
w/o pointcloud	0.364	**0.062**	0.493	0.384	18.7%
w/o semantic	**0.360**	0.080	0.505	0.391	14.0%
w/o goal	0.531	0.593	0.729	0.311	13.3%
First frame	0.554	0.258	0.696	0.298	83.2%
GMT (Full)	0.366	0.072	**0.438**	**0.402**	**13.1%**

Table 3. Ablation study on the ADT dataset. The best results are highlighted in bold.

To analyze the contributions of different components in our method, we conduct an ablation study on the ADT dataset. We first evaluate the effect of each input modality by removing geometric and semantic information individually. Next, we examine the influence of input trajectory length by reducing the number of observed frames to using only the first frame. Finally, we assess the importance of goal conditioning by removing the goal input from the model.

Results. Table 3 reports the ablation results on the ADT dataset. Removing geometric features leads to a noticeable degradation in ADE and Fréchet distance, indicating that local spatial structure is important for accurate trajectory prediction. Excluding semantic information similarly worsens overall performance, though the model retains reasonable final-position accuracy. In contrast, removing goal conditioning results in a substantial drop across all metrics, confirming that explicit goal specification is essential for producing coherent long-horizon trajectories. The "first frame" variant performs the worst, with extremely high collision rates and large deviations from the target, demonstrating that dynamic context is crucial for stable motion generation. Overall, the full model achieves the best Fréchet distance, angular similarity, and lowest collision rate, highlighting the complementary contributions of geometric, semantic, and goal-related features.

5. Limitations, Future Work and Conclusion

In this work, we introduced a trajectory-centric framework that predicts realistic and controllable 6-DOF object trajectories in complex 3D environments. By combining geometric representations, semantic cues, and goal conditioning, our model bridges perception and control through a flexible and generalizable formulation. A key insight of our approach is that object trajectories themselves serve as an effective intermediate representation, enabling cross-embodiment execution via inverse kinematics and simplifying the integration of downstream planners. This design allows us to achieve accurate spatial reasoning and efficient trajectory synthesis while maintaining broad applicability.

Despite these advances, our work has several limitations. First, the model assumes well-aligned scene context and object annotations, which may not hold in cluttered or noisy real-world scenarios. Second, the provided goal condition plays a decisive role in guiding the trajectory generation process; however, such information is often unavailable or difficult to obtain in real applications.

Future work will address these limitations by incorporating goal inference mechanisms that estimate plausible target states from visual observations and contextual cues (e.g. VLM). Furthermore, integrating reinforcement learning or closed-loop feedback could improve adaptation to unseen conditions and support long-horizon planning. Additionally, introducing post refinement based on collision optimization is also a feasible direction. We also see potential in exploring broader cross-embodiment transfer, where a single predicted object trajectory can guide manipulation across robots with different morphologies.

Acknowledgments. This research was partially funded by the German Federal Ministry of Education and Research through the ExperTeam4KI funding program for UDance (Grant No. 16IS24064).

References

[1] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016. 2

[2] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science Robotics*, 4(27), 2019. 1

[3] Jessica Borja-Diaz, Oier Mees, Gabriel Kalweit, Lukas Hermann, Joschka Boedecker, and Wolfram Burgard. Affordance learning from play for sample-efficient policy learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6372–6378. IEEE, 2022. 1

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2

[5] Samuel R Buss. Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods. *IEEE Journal of Robotics and Automation*, 2004. Available online: https://www.math.ucsd.edu/~sbuss/ResearchWeb/ikmethods/. 1

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 4

[7] Shizhe Chen, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Sugar: Pre-training 3d visual representations for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18049–18060, 2024. 3

[8] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21674–21683, 2023. 3

[9] 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, 3

[10] Qian Deng, Le Hui, Jin Xie, and Jian Yang. Sketchy bounding-box supervision for 3d instance segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8879–8888, 2025. 3

[11] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 2

[12] 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. 6

[13] 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. 7, 3

[14] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13190–13200, 2022. 2

[15] James J Gibson. The theory of affordances. *Perceiving, acting, and knowing*, 1:67–82, 1977. 1

[16] 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

[17] 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

[18] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in neural information processing systems*, 35:27953–27965, 2022. 2

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

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[21] Brian Ichter, James Harrison, and Marco Pavone. Learning sampling distributions for robot motion planning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7087–7094. IEEE, 2018. 1

[22] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 3

[23] 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. 3, 5

[24] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 3

[25] Mrinal Kalakrishnan, Sachin Chitta, Evangelos Theodorou, Peter Pastor, and Stefan Schaal. Stomp: Stochastic trajectory optimization for motion planning. In *2011 IEEE International Conference on Robotics and Automation*, pages 4569–4574, 2011. 1

[26] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 4

[27] J.J. Kuffner and S.M. LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Proceedings 2000*

ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065), pages 995–1001 vol.2, 2000. 1

[28] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. 2

[29] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025. 2

[30] 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

[31] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 2

[32] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems*, 37:2161–2184, 2024. 2

[33] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. Ieee, 2015. 3

[34] OpenAI. Sora: Creating video from text, 2024. Accessed: 2025-08-17. 2

[35] 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*, pages 20133–20143, 2023. 6, 1

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Neurips*, 2019. 1

[37] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2878–2888, 2025. 2

[38] 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. 7, 1

[39] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1028–1039, 2025. 1

[40] 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. 1, 3, 4

[41] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 5, 1

[43] Nathan Ratliff, Matt Zucker, J Andrew Bagnell, and Siddhartha S Srinivasa. Chomp: Gradient optimization techniques for efficient motion planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 489–494. IEEE, 2009. 1

[44] 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. 4

[45] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 2

[46] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 3

[47] Bruno Siciliano and Oussama Khatib. *Springer Handbook of Robotics*. Springer, 2016. 1

[48] 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

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[50] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. *Advances in neural information processing systems*, 30, 2017. 2

[51] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 2

[52] Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2318–2328, 2021. 2

[53] Paula Wulkop, Halil Umut Özdemir, Antonia Hüfner, Jen Jen Chung, Roland Siegwart, and Lionel Ott. Learning affordances from interactive exploration using an object-level map. *arXiv preprint arXiv:2501.06047*, 2025. 1

[54] Eric Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025. 2

[55] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6814–6824, 2023. 1

[56] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 2

[57] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025. 2

[58] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *Advances in neural information processing systems*, 32, 2019. 3

[59] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1

[60] 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

[61] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024. 2

[62] 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

[63] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 3

[64] 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

[65] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 2

GMT: Goal-Conditioned Multimodal Transformer for 6-DOF Object Trajectory Synthesis in 3D Scenes

Supplementary Material

In this supplementary material, we provide detailed descriptions of our implementation, dataset processing procedures, and additional experimental results. The document is organized as follows. We first outline the implementation details in Sec. 6, then the overall dataset processing workflow in Sec. 7, followed by the dataset-specific pre-processing steps applied to the Aria Digital Twin (ADT) [35] and HD-EPIC [38] datasets. Next, in Sec. 8, we present several failure cases arising from complex motion patterns, along with additional qualitative results.

6. Implementation Details

Our model is implemented in PyTorch [36] and is trained on a single NVIDIA A6000 GPU. The training process utilizes the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 5×10^{-4}, and exponential learning rate decay with a factor of 0.99.

For data usage, 90% of the available sequences are used for training, with the remaining 10% split equally between validation and testing. Each training sample consists of a full scene point cloud, semantic fixture bounding boxes and labels, and a 6-DOF object trajectory. Trajectories are uniformly resampled to 200 frames, and the first 30% of the sequence is used as the input history. layers dimension

The multimodal Transformer encoder consists of 6 layers with 8 attention heads per layer. The input trajectory is embedded in a 128-dimensional feature space; the global scene point cloud feature has 128 dimensions, and per-point features have 64 dimensions. Semantic fixture bounding boxes are projected to 128 dimensions, while CLIP [42] embeddings for object categories and semantic labels are linearly projected to the same dimension. The goal state feature is also projected to 128 dimensions to ensure compatibility in the fusion space. The latent array within the Transformer is 256-dimensional.

During training, the loss function combines translation, orientation, reconstruction, and destination losses, with weights λ_{trans}, λ_{ori}, λ_{rec}, and λ_{dest} set to 1.0.

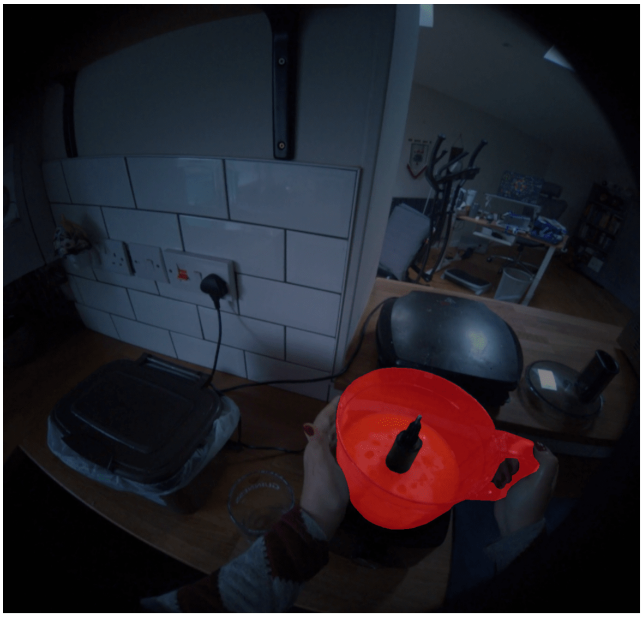
7. Dataset Processing

General Processing. We begin by processing the trajectory data. Since trajectories in ADT are overly dense, we downsample them by retaining one point every five frames to improve computational efficiency while preserving essential motion cues. To support training with multiple batches, we fix the predicted trajectory length to 200. Trajectories shorter than 200 frames are padded, whereas longer ones are truncated. An attention mask ensures that only valid trajectory points contribute to the training objective. We further apply two filtering rules to discard unrealistic motion: an object is considered to be moving only when its velocity exceeds 0.05 m/s, and segments are labeled static when an object remains still for more than three consecutive frames. These criteria help preserve only meaningful trajectories for both training and evaluation.

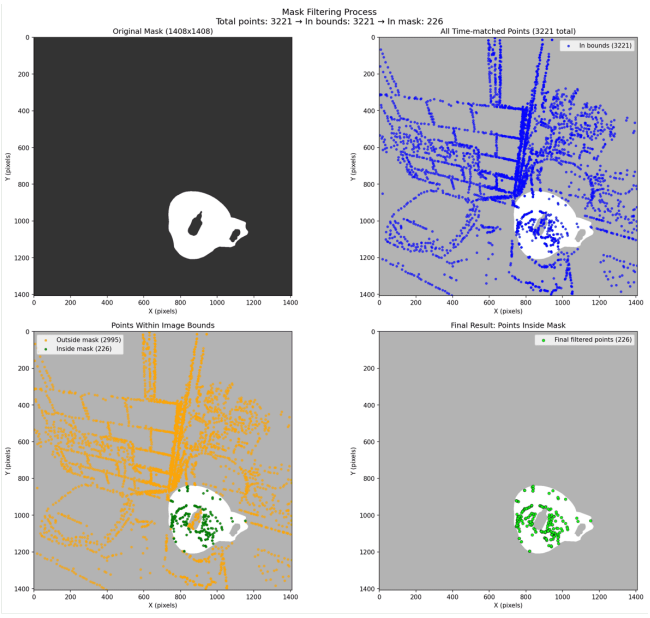
Next, instead of using the full scene point cloud, we extract only the local region around each trajectory to reduce computational overhead. For every trajectory point, we define a spherical neighborhood with a radius of 1 m and collect all points within this region, ensuring that only relevant scene geometry is retained. Similarly, fixture information is extracted only from regions near the trajectory, as distant geometry has limited influence on downstream tasks. These extracted fixture features are then incorporated into the model input. We now describe dataset-specific processing procedures for ADT and HD-EPIC.

Aria Digital Twin Dataset. Since ADT does not include action-level semantic annotations, we use only object categories as descriptive information. Certain trajectories are removed because they do not meet the requirements of our robotic manipulation setting. For example, trajectories in which an object is held or moved within an extremely limited spatial region for extended periods are excluded, as they lack meaningful interaction patterns and do not reflect practical robotic behaviors.

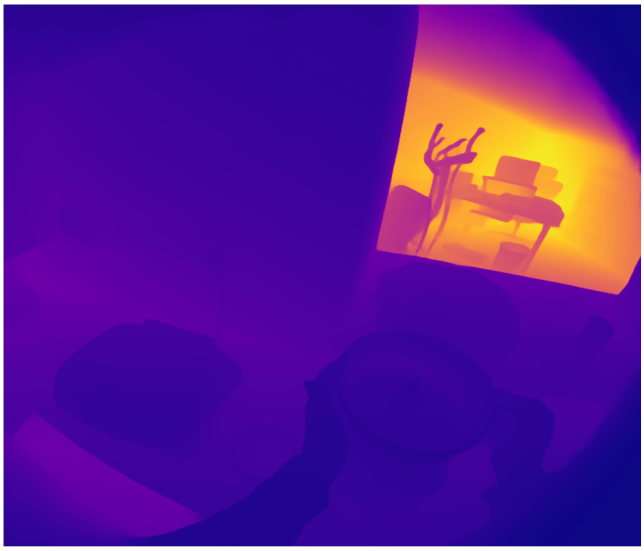
HD-EPIC Dataset. In contrast to ADT, the HD-EPIC dataset does not provide bounding boxes or semantic labels for scene objects. Large static structures such as countertops and drawers are manually reconstructed in Blender and aligned with the scene point cloud to serve as static bounding boxes. For small manipulable objects such as coffee machines and knives, HD-EPIC provides start/end timestamps of object motion, 2D masks, and 3D object centers. Using this information, we first align the timestamps with SLAM data and obtain sparse 2D–3D correspondences using MPS data collected with Aria glasses. We then estimate monocular depth using UniK3D [39], perform linear depth alignment with the correspondences to recover the true scale, and reconstruct 3D object bounding boxes. These *dynamic bounding boxes*, together with static ones, are used for model training. Fig. 5 illustrates the full processing pipeline.



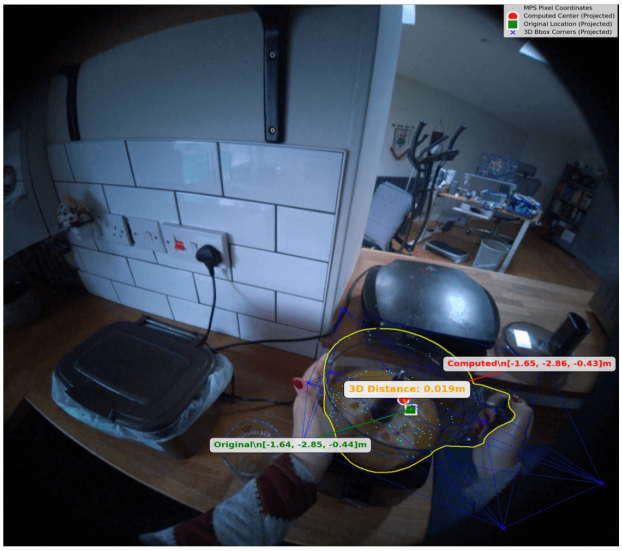
(a) Given 2D annotation



(b) 2D-3D correspondence filtering



(c) metrics depth estimation by Unik3D



(d) 3D bbox caculation

Figure 5. **3D bounding box reconstruction in the HD-EPIC dataset.** (a): input RGB frame with object mask. (b): mask filtering and sparse 2D-3D correspondences from SLAM and MPS data. (c): monocular depth estimation from UniK3D. (d): final 3D bounding box recovered after depth alignment and scaling. This pipeline enables accurate localization of small objects (e.g., bowls, cups) in cluttered scenes.

Since the dataset provides only pick-up and drop-off annotations for each object, we generate dense object trajectories using hand-tracking. We compute the transformations of the manipulating hand over time and apply these transformations to the provided initial object position, producing a complete trajectory for each object.

The hand trajectory extraction pipeline integrates multiple perception systems to process egocentric video. The pipeline begins with a bootstrap stage that establishes hand–object correspondence by computing the 3D Euclidean distance between detected hand positions and the annotated object center in world coordinates to identify the primary

manipulating hand. Subsequent hand positions are obtained using Project Aria's MPS [13]. For frames in which both hands are confidently detected, we leverage Hands23 [9] to disambiguate which hand is physically interacting with the object. Hands23 infers binary contact states for each hand, enabling reliable determination of the manipulating hand even when both hands appear in view. When Hands23 outputs are ambiguous (e.g., both hands detected in contact), the system maintains temporal consistency by defaulting to the initially selected primary hand. Temporal coherence is further enforced via a sliding-window filter (window size = 3), which suppresses spurious frame-to-frame switching.

For orientation estimation, we construct a 6D rotation representation [64] derived from the geometric structure of the hand. Specifically, we use Singular Value Decomposition (SVD) to compute the hand coordinate frame, where the primary axis aligns with the wrist-to-palm vector and the palm normal defines the facing direction, both obtained from MPS. This 6D parameterization ensures continuity across the rotation manifold and avoids the singularities present in Euler angles and quaternions.

We demonstrate the effectiveness of the reconstructed object trajectories in Fig. 6, where the recovered 3D object location is projected onto the input RGB frames for visual verification.

8. Additional Qualitative Results

To complement the results presented in the main paper, we provide additional qualitative examples for both the ADT and HD-EPIC datasets. Figures 9 and 10 illustrate representative failure cases and their likely causes, while Figs. 7 and 8 present additional successful predictions.

Figure 6. **Object Position computation using hand-tracking** We demonstrate the object positions. depicted as Yellow along with the orientation and the hand that is interacting with the object at the particular frame sampled at intervals during the entire period of the moving object.

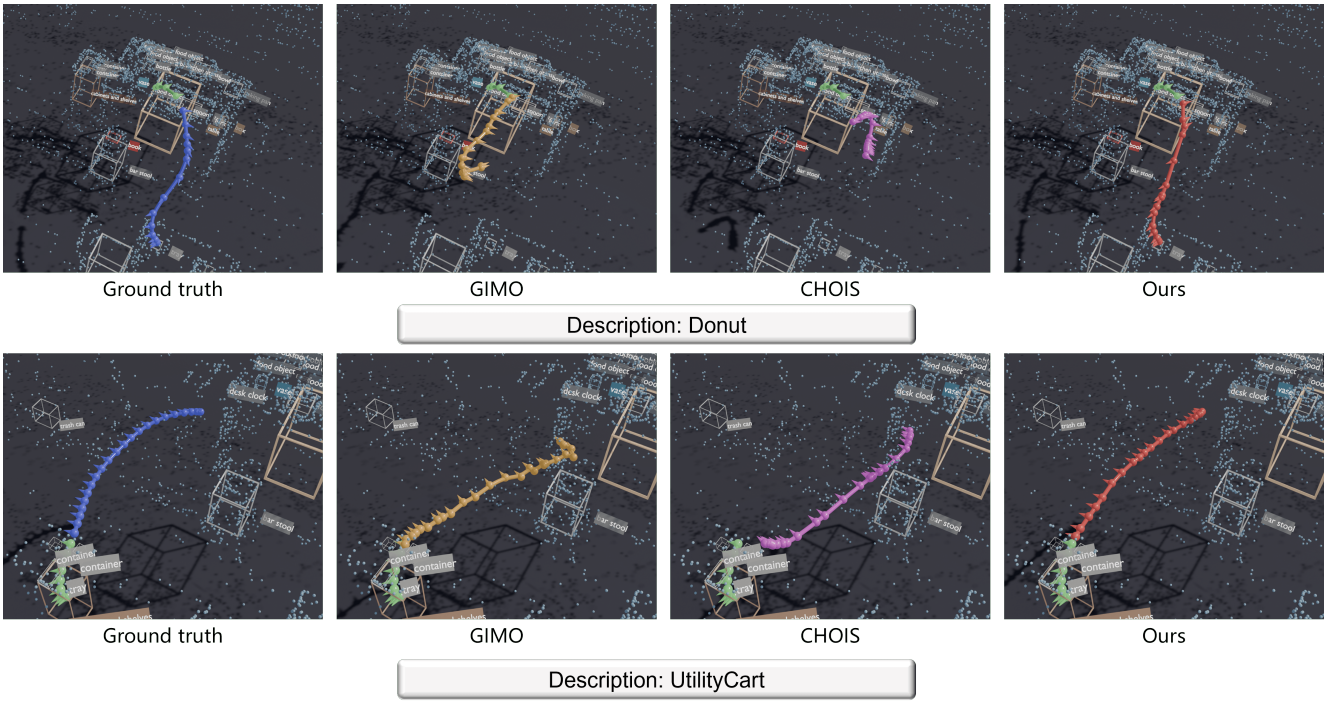
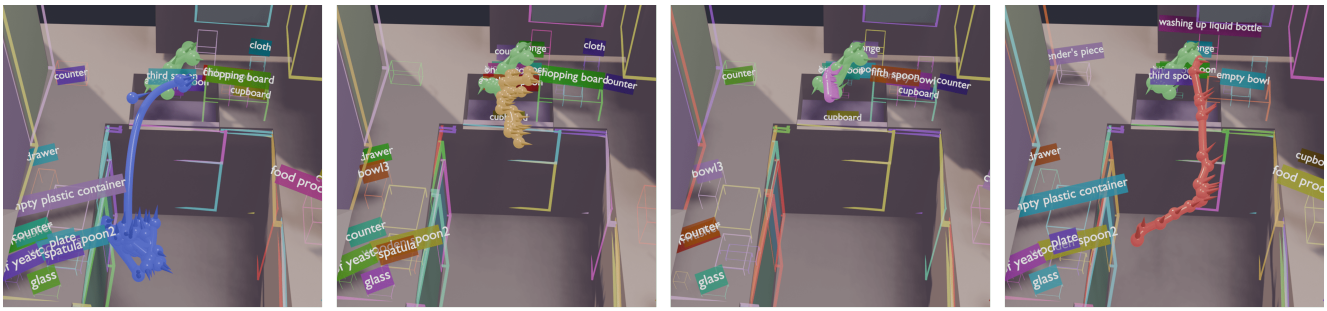
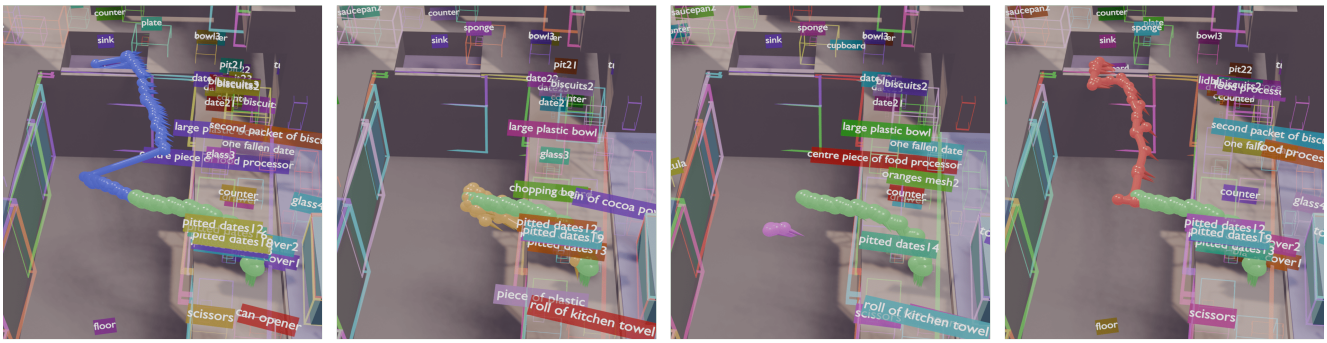


Figure 7. **Qualitative results in the ADT dataset.**



Description: pick up another spoon that's soaking in the sink. Remove what's stuck on the spoon by brushing the finger over it while soaking it in the water. Open the top cutlery drawer. Put the knife into a slot in the cutlery drawer.



Description: pick up an empty bowl from the countertop using the left hand pick up an empty plate from the countertop using the right hand this action is mostly off screen put the plate inside the sink using the right hand put the bowl inside the sink using the left hand this action is mostly off screen

Figure 8. **Qualitative results in the HD-EPIC dataset.**

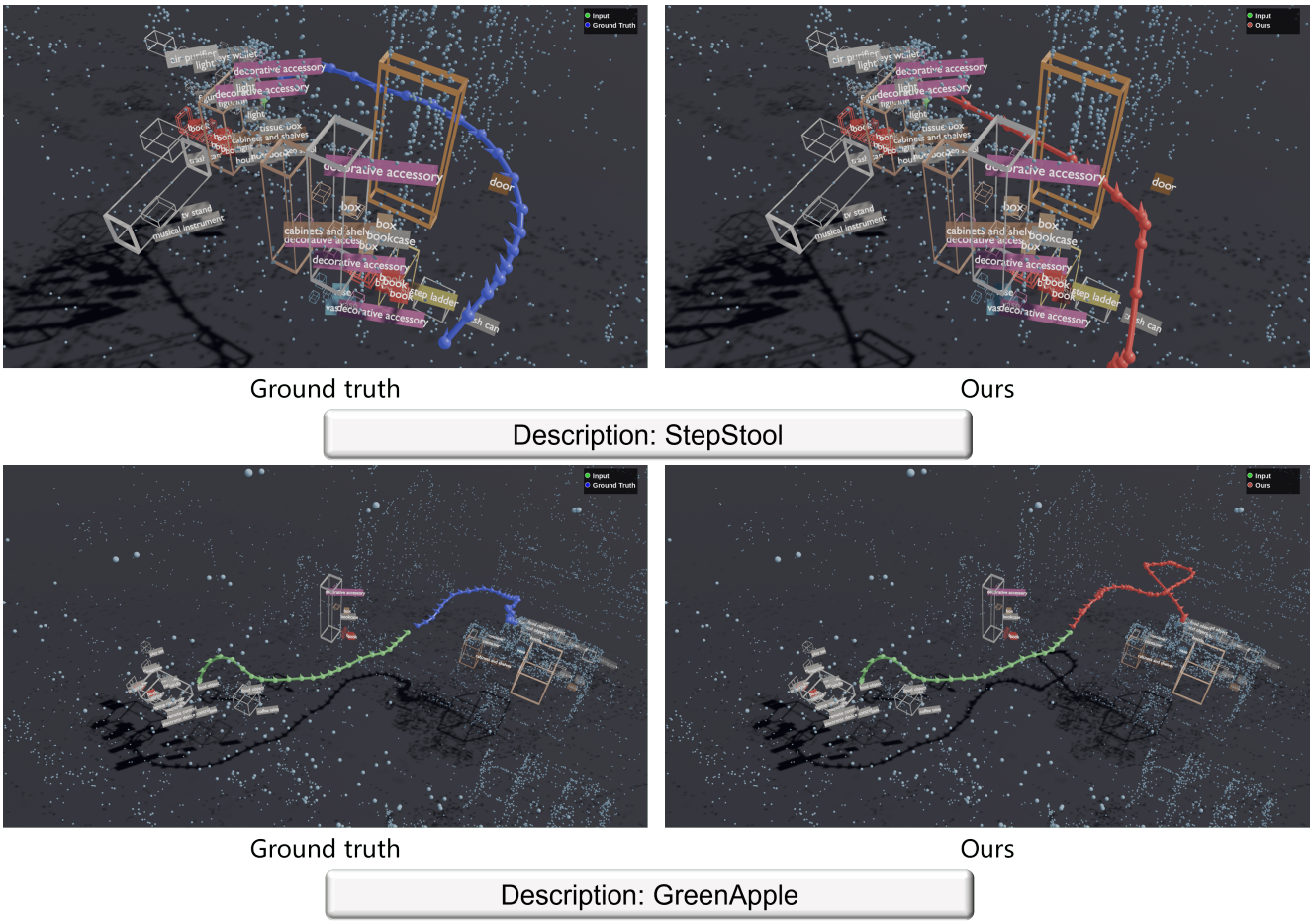


Figure 9. **Failure Cases in the ADT dataset.** We can observe that sometimes, despite being goal-conditioned, the generated trajectory may be longer than the ground truth trajectory and may overshoot the destination.

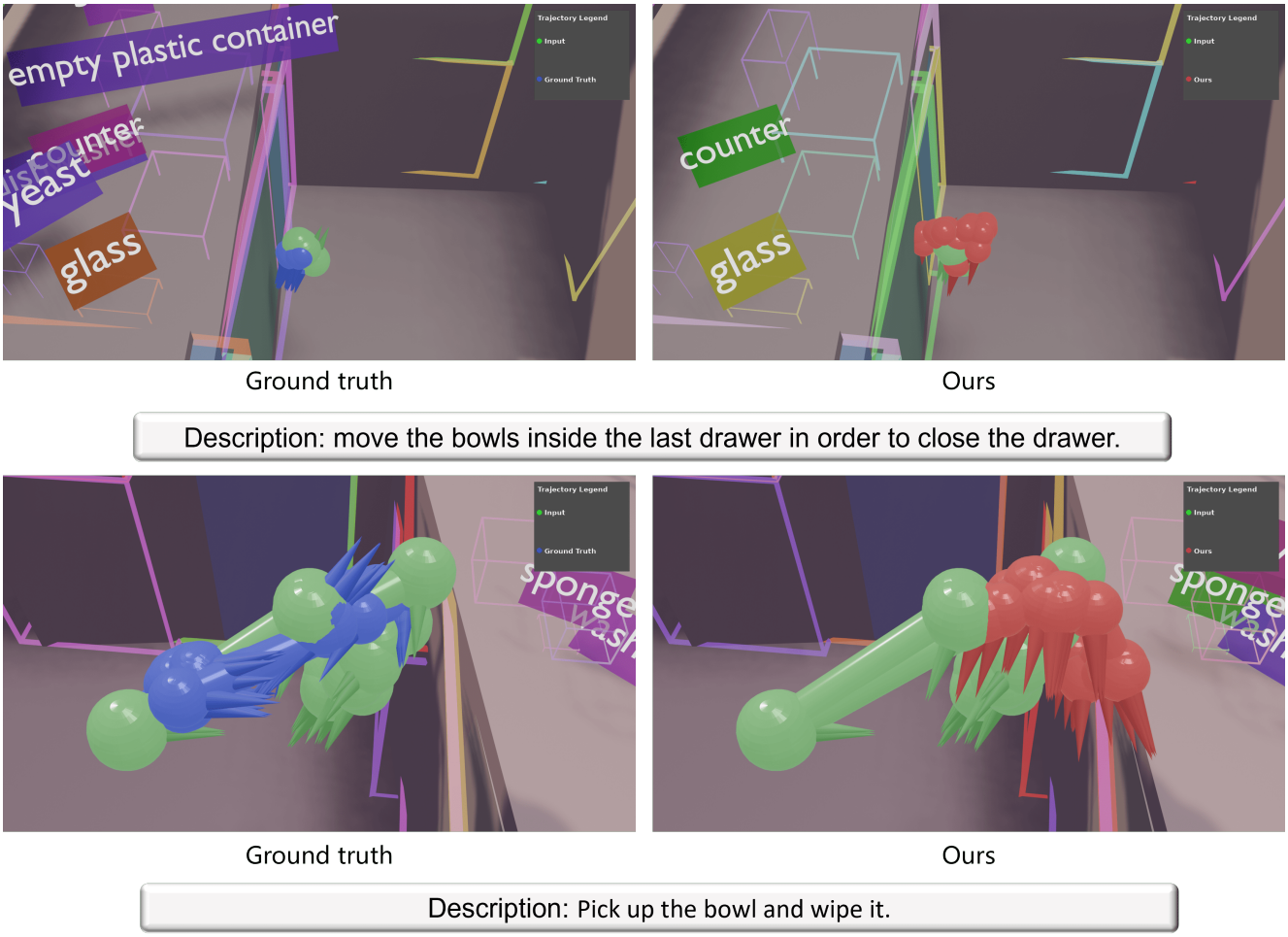


Figure 10. **Failure cases in the HD-EPIC dataset.** Our method suffers from adding redundant motion for inputs don't have significant change in their positions. Such motion os observed for small object trajectories that are often interacted with for a very small duration.