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

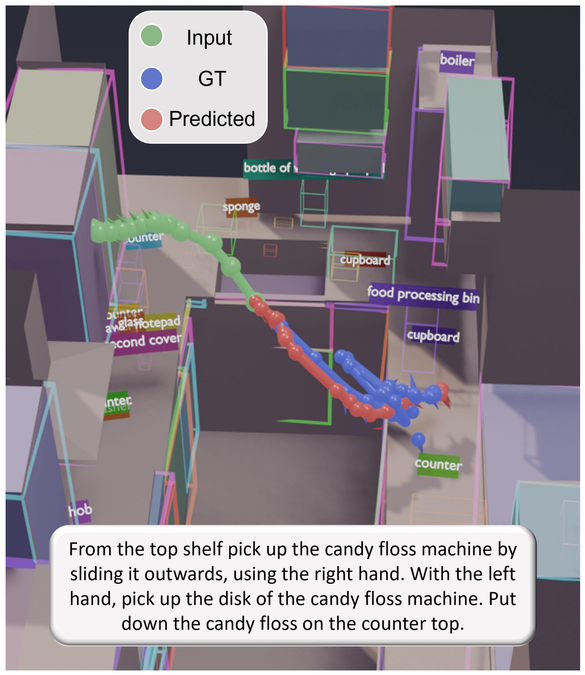
Huajian Zeng[1,4,*], Abhishek Saroha[1,2,*], Daniel Cremers[1,2], Xi Wang[1,2,3]

[1]Technical University of Munich (TUM), [2]Munich Center for Machine Learning (MCML),
[3]ETH Zürich, [4]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

3DV 2026

Why Object Trajectory

Synthesizing controllable 6-DOF object manipulation trajectories in 3D environments is essential for enabling robots to interact with complex scenes.



Problem Formulation

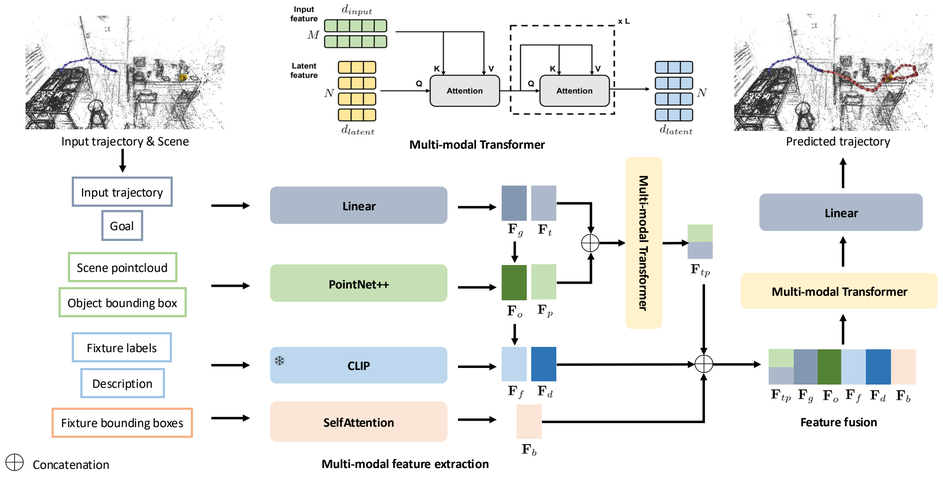
We learn a conditional distribution over future object motion:

$$P(\hat{X}_{H+1:T} \mid X_{1:H}, G, S)$$

where $X_{1:H}$ is the observed 6-DOF history, G is the target end pose, and S contains scene point clouds and fixture bounding boxes.

The model predicts the remaining future trajectory that is both scene-consistent and goal-directed

GMT Architecture



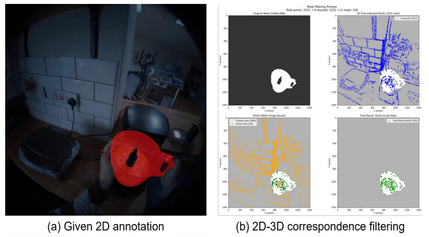
Key Design

Object-centric generation: predict object motion directly instead of modeling it as a byproduct of human motion.
Tailored conditioning: combine local geometry, fixture semantics, language description, and goal pose.
Geometry-first fusion: hierarchical transformer fusion improves spatial feasibility, long-horizon stability, and goal alignment.

Egocentric Video Preprocessing

Webpage Video

Paper Huajian

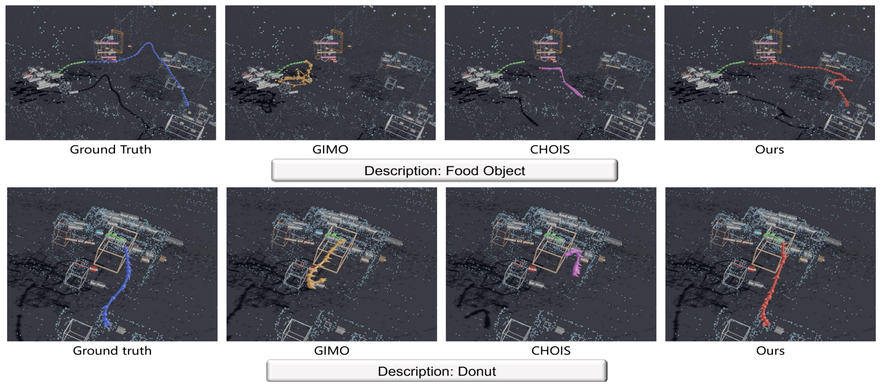


Surrounding Objects Reconstruction

Experiments

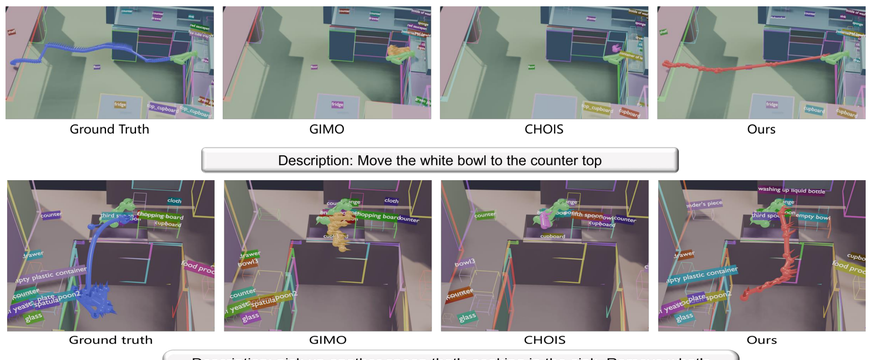
- Controlled Scenarios (Aria Digital Twin Dataset [1])

Method	ADE[m] ↓	FDE[m] ↓	FD[m] ↓	AC[m] ↑	CR[%] ↓
GIMO [3]	0.982	1.401	1.511	0.140	19.6
CHOIS [4]	0.853	1.062	1.209	0.283	**9.3**
GMT (Ours)	**0.366**	**0.072**	**0.438**	**0.402**	13.1



- Realistic Scenarios (HD-EPIC Dataset [2])

Method	ADE[m] ↓	FDE[m] ↓	FD[m] ↓	AC[m] ↑	CR[%] ↓
GIMO [3]	0.411	0.654	0.780	0.002	11.8%
CHOIS [4]	0.446	0.589	0.760	0.009	12.0%
GMT (Ours)	**0.283**	**0.034**	**0.391**	**0.037**	**10.3%**



References

[1] Pan X, Charron N, Yang Y, et al. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. CVPR 2023.
[2] Perrett T, Darkhalil A, Sinha S, et al. Hd-epic: A highly-detailed egocentric video dataset. CVPR 2025.
[3] Zheng Y, Yang Y, Mo K, et al. Gimo: Gaze-informed human motion prediction in context. ECCV 2022.
[4] Li J, Clegg A, Mottaghi R, et al. Controllable human-object interaction synthesis. ECCV 2024.