GRAIL: Generating Humanoid Loco-Manipulation from 3D Assets and Video Priors

Tianyi Xie[1,2,†], **Haotian Zhang**[1,†], **Jinhyung Park**[1,†], **Zi Wang**[1,†], **Bowen Wen**[1], **Jiefeng Li**[1], **Xueting Li**[1], **Qingwei Ben**[1], **Haoyang Weng**[1], **Yufei Ye**[1], **David Minor**[1], **Tingwu Wang**[1], **Chenfanfu Jiang**[2], **Sanja Fidler**[1], **Jan Kautz**[1], **Linxi Fan**[1], **Yuke Zhu**[1], **Zhengyi Luo**[1,‡], **Umar Iqbal**[1,‡], **Ye Yuan**[1,‡]

[1] NVIDIA [2] UCLA

[†] Co-First Authors [‡] Project Leads

https://research.nvidia.com/labs/dair/grail/

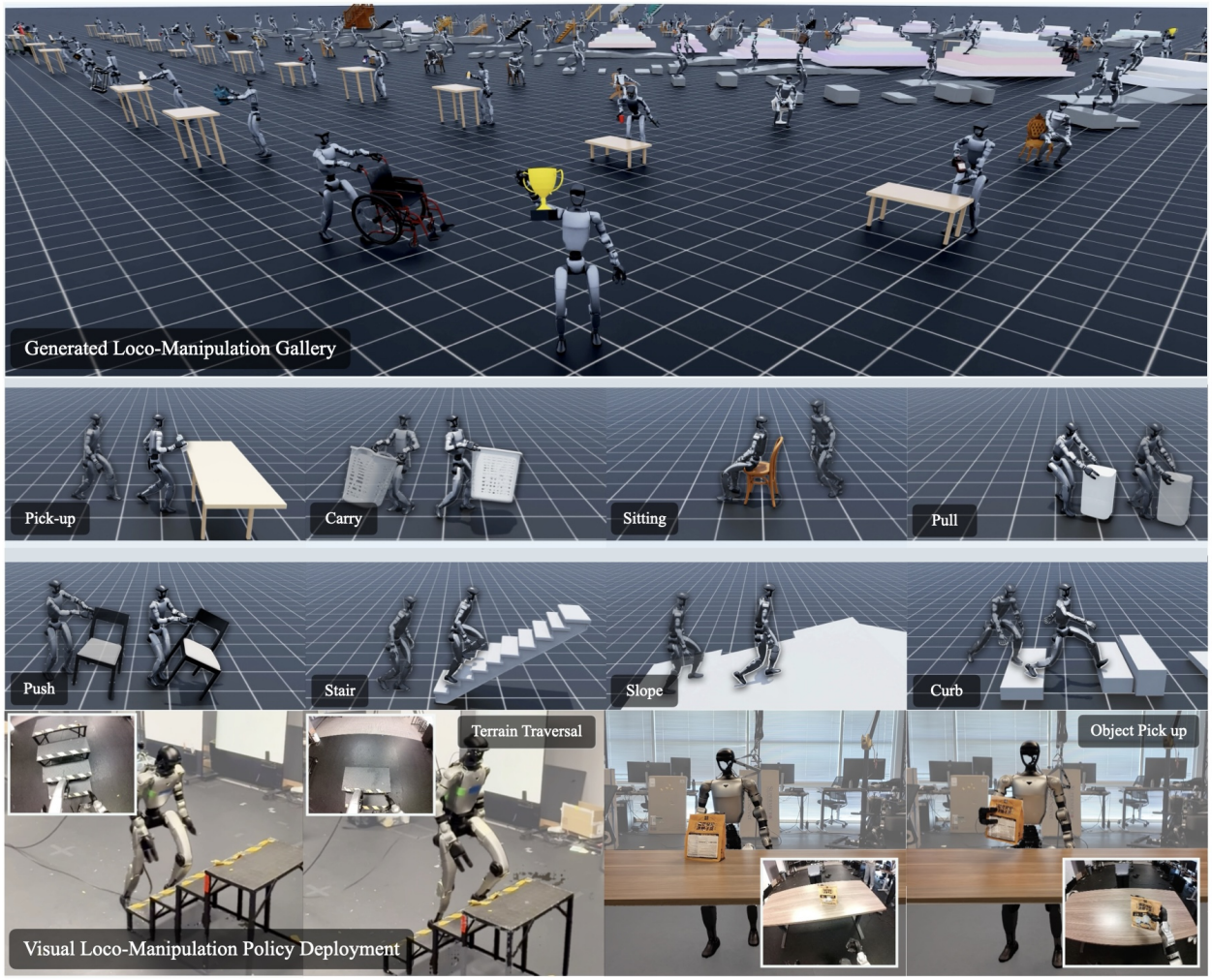

Figure 1: **From Fully Digital Data Generation to Real-World Deployment.** GRAIL generates humanoid loco-manipulation data from 3D assets and video priors without physical scene rebuilds or robot teleoperation. *Top:* simulated humanoids execute generated references spanning pick-up, whole-body manipulation, sitting, and terrain traversal. *Bottom:* egocentric visual policies trained only on GRAIL-generated data are deployed on a Unitree G1 for stair-climbing and object pick-up.

Abstract

Scaling humanoid loco-manipulation requires robot-compatible demonstrations across diverse objects, whole-body motions, and scene geometries, but teleoperation and motion capture are difficult to scale because each collection depends on physical setups, instrumented actors, and robot operation. We present GRAIL, a digital generation pipeline that remains fully virtual until deployment: it composes 3D assets, simulator-ready scenes, and priors from video foundation models (VFMs) to synthesize interactions without rebuilding physical environments or teleoperating the robot. Rather than reconstructing unconstrained in-the-wild videos, GRAIL starts from fully specified 3D configurations in which object geometry, camera parameters, metric scale, environment depth, and a robot-proportioned character are known before video generation and reused during reconstruction. This privileged setup better conditions 4D recovery, allowing model-based object tracking, human motion estimation, and interaction-aware optimization to reconstruct metric 4D human-object interaction (HOI) trajectories with reduced depth ambiguity and morphology mismatch. We retarget the recovered motions to a humanoid robot and train complementary task-general trackers: an object-aware latent adaptor for manipulation and a scene-aware tracker for terrain traversal. GRAIL produces over 20,000 sequences spanning pick-up, whole-body manipulation, sitting, and terrain traversal. Using only GRAIL-generated data, we train egocentric visual policies through a sim-to-real pipeline and deploy them on a Unitree G1 humanoid, achieving 84% real-world success on diverse object pick-up and 90% success on stair-climbing.

1. Introduction

Humanoid loco-manipulation requires policies that coordinate whole-body balance, object contact, and scene-aware locomotion across a broad distribution of objects and terrain geometries. Scaling the corresponding demonstration data is challenging because each trajectory must be both physically plausible and executable by the target robot. Teleoperation (Aldaco et al., 2024; Ben et al., 2025; Khazatsky et al., 2024; Ze et al., 2025a,b) and motion capture (Lu et al., 2025; Taheri et al., 2020) provide high-quality demonstrations, but they are difficult to scale: each new object or terrain layout can require human-operated robot demonstrations, instrumented actors, and physical scene reconfiguration. Reconstructing robot-ready 4D trajectories from in-the-wild videos (Hou et al., 2023; Kim et al., 2023; Petrov et al., 2023; Wang et al., 2022; Xie et al., 2022, 2026; Zhang et al., 2020) offers broad visual coverage, but requires inferring camera, scale, object geometry, human shape, contacts, and world-space motion from ambiguous monocular observations. Recent advances in 3D asset generation and video foundation models suggest an alternative route: instead of recovering the entire 3D interaction from an uncontrolled video, can we first specify the 3D scene and then use video generative priors to synthesize diverse interactions for humanoid policy learning?

We introduce GRAIL, a humanoid-centric data-generation pipeline that remains fully digital until real-world deployment. It uses video foundation models (VFMs) as interaction priors inside a simulator-ready 3D asset pipeline: rather than reconstructing uncontrolled videos into ambiguous 4D scenes, GRAIL first specifies the object, scene geometry, camera, scale, and robot-proportioned character, then recovers the interaction within this known metric frame. This design addresses two bottlenecks of prior data sources: it avoids repeated physical collection required by teleoperation and motion capture, and it produces robot-trackable trajectorie s already aligned with simulation for downstream sim-to-real policy training.

The recovered 4D HOI trajectories are retargeted to a Unitree G1 and converted into task-general tracking policies built on a pretrained whole-body controller (Luo et al., 2025). Rather than fitting one controller per sequence or per object, we pool related trajectories so the trackers cover families of manipulation and scene-interaction behaviors. This stage uses two complementary specializations: an object-aware latent adaptor that augments the frozen whole-body controller with manipulation by modulating its latent tokens and emitting hand actions, and a scene-aware tracker that fine-tunes the controller together with a height-map encoder for

terrain-conditioned whole-body control. With this pipeline, we generate a large-scale dataset of over 20,000 humanoid loco-manipulation sequences spanning pick-up, whole-body manipulation, sitting, and terrain traversal. Using only this generated data, we train egocentric visual policies (He et al., 2025a) with visual domain randomization and camera alignment as a closed-loop sim-to-real validation; deployed on a Unitree G1, the resulting RGB-based policies perform autonomous loco-manipulation on pick-up and stair-climbing tasks.

In summary, GRAIL makes the following contributions: (i) a fully digital humanoid-centric data-generation framework that uses VFMs as interaction priors inside a fully specified 3D asset pipeline, producing over 20,000 physically plausible loco-manipulation sequences; (ii) an interaction-aware 4D HOI reconstruction stack that exploits known geometry, metric scale, camera parameters, environment depth, and a robot-proportioned character; (iii) complementary task-general trackers for reconstructed 4D HOI, pairing object-aware latent adaptation for manipulation with scene-aware height-map conditioning for terrain traversal and sitting; and (iv) an end-to-end sim-to-real validation of GRAIL-generated data through egocentric visual policies deployed on a Unitree G1, achieving 84% pick-up success and 90% stair-climbing success in the real world.

2. Related Work

Human-Object Interaction Generation and Reconstruction. Synthesizing human-object interactions (HOI) requires reasoning about human motion, object affordances, and physical contact. Existing data sources rely on motion capture (Bhatnagar et al., 2022; Fan et al., 2023; Huang et al., 2022; Jiang et al., 2023; Kim et al., 2025b; Li et al., 2023a; Lu et al., 2025; Taheri et al., 2020; Zhang et al., 2023b; Zhao et al., 2024) or RGB-based reconstruction (Hou et al., 2023; Kim et al., 2023; Petrov et al., 2023; Wang et al., 2022; Xie et al., 2022, 2026; Zhang et al., 2020), but remain expensive, category-limited, or underconstrained. Learning-based methods synthesize HOI from affordance, language, or vision-language priors (Dang et al., 2025; Diller and Dai, 2024; Dwivedi et al., 2025; Jiang et al., 2024; Kulkarni et al., 2024; Li et al., 2024; Li and Dai, 2024; Li et al., 2023b; Peng et al., 2025; Wu et al., 2025; Xu et al., 2023, 2024; Ye et al., 2023; Zhang et al., 2023a, 2025; Zheng et al., 2023; Zhou et al., 2022), but physical realism and temporal coherence remain challenging. VFM-based pipelines such as DAViD (Kim et al., 2025a), ZeroHSI (Li et al., 2026), and related methods (Lou et al., 2025) use generated videos as priors for 4D HOI recovery, yet typically leave camera, scale, character morphology, object geometry, or environment structure to be inferred after generation. GRAIL instead specifies the 3D scene before generation and reuses it during reconstruction, yielding robot-compatible 4D HOI trajectories grounded by known metric scale, environment geometry, and a robot-proportioned character, facilitating downstream sim-to-real policy learning.

Human Video as Humanoid Supervision. Human video has become an increasingly important supervision source for humanoids: large-scale mining, retargeting, and robotized-video generation provide broad pose-control or pretraining data (Mao et al., 2024; Yang et al., 2025), while VideoMimic (Allshire et al., 2025), HumanX (Wang et al., 2026), and related systems (Shi et al., 2026; Weng et al., 2025; Yin et al., 2025; Yu et al., 2025) train interaction policies from third-person, monocular, or egocentric videos. In parallel, physics-based control and whole-body imitation (Fu et al., 2024; He et al., 2024, 2025b; Luo et al., 2023, 2025; Peng et al., 2018, 2021), residual adaptors (Zhao et al., 2025), and multimodal controllers (He et al., 2026; Jiang et al., 2026) show how robot-ready references can be converted into executable policies. The shared bottleneck is data: videos still require recovering metric motion, contacts, object state, and scene geometry, while teleoperation, motion capture, wearable interfaces, and generated robot-video demonstrations (Ben et al., 2025; Khazatsky et al., 2024; Lu et al., 2025; Nai et al., 2026; Patel et al., 2025; Taheri et al., 2020; Ze et al., 2025a,b) remain limited by human effort, retargeting, platform dependence, or morphology mismatch. GRAIL addresses this upstream bottleneck by using generated video for behavioral priors while keeping geometry, scale, camera, environment, and target morphology known, producing robot-compatible 4D references for task-general tracking policies.

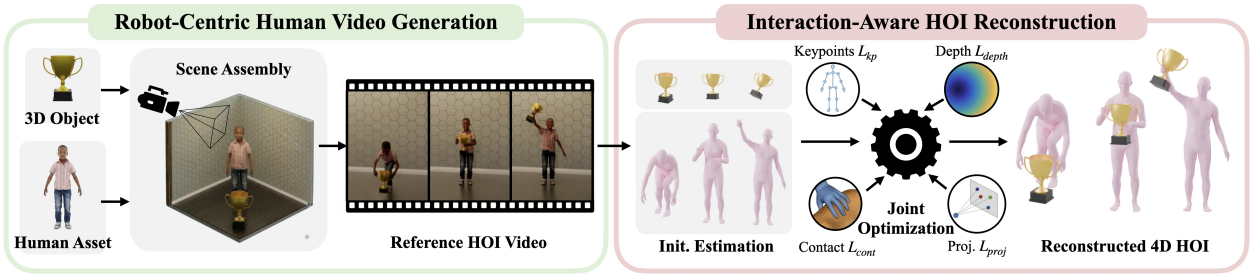

Figure 2: **Asset-Conditioned 4D HOI Generation.** Given a 3D object asset, we render a fully specified 3D scene with a character prefitted to the target humanoid and known camera parameters, synthesize a static-camera interaction video via a VFM conditioned on the rendered frame, and reconstruct metric 4D human-object motion by jointly refining initial human and object trajectory estimates with keypoint, depth, and contact losses anchored to the privileged 3D configuration.

3. Method

Given a 3D object asset $\mathcal{M}^{\mathcal{O}}$, GRAIL produces humanoid loco-manipulation demonstrations comprising humanoid kinematic motion $\{\boldsymbol{\Theta}_t^{\mathcal{R}}\}_{t=1}^T$, object kinematic motion $\{\boldsymbol{\Theta}_t^{\mathcal{O}}\}_{t=1}^T$, and robot actions $\{\boldsymbol{a}_t^{\mathcal{R}}\}_{t=1}^T$. Our data generation pipeline proceeds in three stages. First, we assemble a fully specified 3D configuration with a character prefitted to the target robot, render an initial frame, and feed it into a VFM to synthesize a reference HOI video $\{I_t\}_{t=1}^T$ (Sec. 3.1). Second, leveraging the known 3D configuration, we reconstruct coherent 4D HOI trajectories $\{(\widetilde{\boldsymbol{\Theta}}_t^{\mathcal{H}}, \widetilde{\boldsymbol{\Theta}}_t^{\mathcal{O}})\}_{t=1}^T$ through human pose estimation, object tracking, and joint optimization (Sec. 3.2). Third, we retarget the reconstructed motions to the target humanoid and train task-general tracking policies across each task family (Sec. 3.3). Using the generated data, we further train egocentric visual policies through a sim-to-real pipeline and deploy them on a real Unitree G1 for pick-up and stair-climbing (Sec. 3.4).

3.1. Robot-Centric Human Video Generation

Although one could generate robot videos directly, current VFMs have stronger priors over human manipulation, and human body and hand reconstruction tools are more mature than robot reconstruction tools. We therefore synthesize human interaction videos using a character asset prefitted to the target humanoid, which facilitates retargeting the recovered motion to the robot. To assemble the 3D configuration, we construct candidate environments using Infinigen (Raistrick et al., 2023) and position the human asset in a rest pose alongside the object. We use rigid body simulation (Macklin et al., 2016) to settle the object into a stable, collision-free initial configuration $\boldsymbol{\Theta}_1^{\mathcal{O}}$. We then render the first frame using Blender with known camera intrinsics $\boldsymbol{C}_K \in \mathbb{R}^{3\times3}$ and extrinsics $\boldsymbol{C}_E = (\boldsymbol{r}^{\mathcal{C}}, \boldsymbol{t}^{\mathcal{C}})$. The generated environment serves two purposes: realistic visual context for VFM generation and a ground-truth point cloud for metric-scale depth alignment during reconstruction (Sec. 3.2). A VLM (OpenAI, 2024) generates an interaction prompt from the rendered frame, and a VFM (*e.g.*, Kling (Kuaishou, 2025)) then synthesizes the reference HOI video $\{I_t\}_{t=1}^T$ under a static-camera setting that preserves the known camera parameters $(\boldsymbol{C}_K, \boldsymbol{C}_E)$ for reconstruction.

3.2. Interaction-Aware HOI Reconstruction

Given the generated interaction video, we recover the 4D HOI trajectory $\{(\widetilde{\boldsymbol{\Theta}}_t^{\mathcal{H}}, \widetilde{\boldsymbol{\Theta}}_t^{\mathcal{O}})\}_{t=1}^T$ in two steps: independent initial estimation of human and object motion, followed by interaction-aware joint optimization that anchors the trajectories to the privileged 3D configuration.

3.2.1. Initial Motion Estimation

We first estimate human and object motion independently, yielding initial world-space trajectories $\{\widehat{\boldsymbol{\Theta}}_t^{\mathcal{H}}\}_{t=1}^T$ and $\{\widehat{\boldsymbol{\Theta}}_t^{\mathcal{O}}\}_{t=1}^T$.

Human Motion Estimation. For the human body, GENMO (Li et al., 2025) provides per-frame SMPL-X (Pavlakos et al., 2019) pose parameters from the generated video in camera space; the body shape is held fixed at the prefitted character morphology from Sec. 3.1 rather than re-estimated, so GENMO only contributes per-frame pose parameters. The camera-space motion is then transformed into world coordinates using the known camera extrinsics C_E. For the hands, WiLoR (Potamias et al., 2025) refines per-frame MANO (Romero et al., 2017) parameters for the left and right hands independently; missing detections due to partial occlusion or detection failure are filled via temporal linear interpolation and smoothed with a Savitzky-Golay filter (Savitzky and Golay, 1964) to suppress per-frame jitter. The smoothed hand poses are integrated into the SMPL-X body through wrist inverse-kinematic (IK) alignment, preserving the WiLoR-predicted finger configuration.

Object Pose Tracking. For the object, we fine-tune FoundationPose (Wen et al., 2024) on its proposed synthetic dataset for 5 epochs with the depth channels zeroed at both training and inference to adapt to our RGB-only setup; at inference, the 6-DoF tracker is initialized from the known first-frame pose $\mathbf{\Theta}_1^{\mathcal{O}}$ and propagates the object pose across all frames. FoundationPose requires known object geometry, texture, and camera parameters, all available in our pipeline by construction, which ensures accurate object tracking. We additionally validate tracking quality by comparing predicted poses against SAM2 (Ravi et al., 2024) segmentation masks and discard sequences with inconsistent geometry (Sec. A.3).

3.2.2. Joint Optimization

Directly combining the independent reconstructions often produces misaligned interactions (floating contacts, penetration, and depth-scale drift). We therefore jointly refine both trajectories through a global optimization over all frames, holding the hand poses fixed for stability. Rather than optimizing full trajectories directly, we optimize residual motion parameters $\{\Delta\mathbf{\Theta}_t^{\mathcal{H}}\}_{t=1}^T$ and $\{\Delta\mathbf{\Theta}_t^{\mathcal{O}}\}_{t=1}^T$; the final poses are $\mathbf{\Theta}_t^{\mathcal{H}} = \widehat{\mathbf{\Theta}}_t^{\mathcal{H}} \oplus \Delta\mathbf{\Theta}_t^{\mathcal{H}}$ and $\mathbf{\Theta}_t^{\mathcal{O}} = \widehat{\mathbf{\Theta}}_t^{\mathcal{O}} \oplus \Delta\mathbf{\Theta}_t^{\mathcal{O}}$, with $\oplus$ denoting residual translation and rotation updates and the 6D rotation representation (Zhou et al., 2019) used for continuous parameterization. The full refinement objective is:

$$L = \lambda_{\text{kp}} L_{\text{kp}} + \lambda_{\text{proj}} L_{\text{proj}} + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{cont}} L_{\text{cont}} + \lambda_{\text{reg}} L_{\text{reg}}, \tag{1}$$

yielding the optimized trajectories $\{(\tilde{\mathbf{\Theta}}_t^{\mathcal{H}}, \tilde{\mathbf{\Theta}}_t^{\mathcal{O}})\}_{t=1}^T$.

Keypoint Alignment. To keep the optimized human trajectory aligned with the generated video, we minimize the distance between projected and detected 2D body and hand keypoints:

$$L_{\text{kp}} = \frac{1}{T} \sum_{t=1}^{T} \left\| \mathcal{K}^{\mathcal{H}}(\mathbf{\Theta}_t^{\mathcal{H}}) - p_t \right\|, \tag{2}$$

where $p_t \in \mathbb{R}^{J\times 3}$ are 2D keypoints obtained from body and hand keypoint estimators (Potamias et al., 2025; Xu et al., 2022), and $\mathcal{K}^{\mathcal{H}}(\cdot)$ projects the SMPL-X parameters using the known camera.

Object Projection Alignment. Since FoundationPose provides image-aligned object poses, we regularize the optimized object pose to preserve that alignment:

$$L_{\text{proj}} = \sum_{t=1}^{T} \left\| \mathcal{P}(V_t^{\mathcal{O}}) - \mathcal{P}(\widehat{V}_t^{\mathcal{O}}) \right\|, \tag{3}$$

where $\mathcal{P}(\cdot)$ is the camera projection function, and $V_t^{\mathcal{O}}$ and $\widehat{V}_t^{\mathcal{O}}$ are object vertices under the optimized and initial poses.

Depth Alignment. Leveraging the known 3D configuration, we first estimate a depth map with MoGe-2 (Wang et al., 2025) and align it to the ground-truth background depth rendered from the environment, recovering metric-scale depth. We then segment human and object regions with SAM2 (Ravi et al., 2024) and unproject

them into per-frame point clouds $\mathbf{P}_t^{\mathcal{H}}$ and $\mathbf{P}_t^{\mathcal{O}}$. The depth-alignment loss encourages the reconstructed meshes to match these point clouds:

$$L_{\text{depth}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{CD}(V_t^{\mathcal{H},\text{vis}}, \mathbf{P}_t^{\mathcal{H}}) + \mathcal{CD}(V_t^{\mathcal{O},\text{vis}}, \mathbf{P}_t^{\mathcal{O}}), \tag{4}$$

where $V_t^{\mathcal{H},\text{vis}}$ and $V_t^{\mathcal{O},\text{vis}}$ are visible mesh vertices and $\mathcal{CD}$ is bidirectional Chamfer distance.

Contact Alignment. To encourage physically plausible contact, we query a VLM (OpenAI, 2024) on uniformly sampled video frames to predict per-frame contact labels (*e.g.*, left or right hand) and propagate each label to its surrounding interval. Using these labels, we identify the relevant SMPL-X vertices $V_t^{\mathcal{H},\text{cont}}$ via SMPL-X part segmentation and apply the contact loss only to frames where contact is detected. Since the image-space losses already enforce projection consistency, the contact loss only needs to resolve depth discrepancies, so we restrict it to object vertices whose projected positions overlap with the contact body region and penalize only their depth offset:

$$L_{\text{cont}} = \frac{1}{|\mathcal{T}_c|}\sum_{t\in\mathcal{T}_c} \mathcal{CD}_z(V_t^{\mathcal{H},\text{cont}}, V_t^{\mathcal{O},\text{cont}}), \quad V_t^{\mathcal{O},\text{cont}} = \mathcal{F}(V_t^{\mathcal{O}}, V_t^{\mathcal{H},\text{cont}}), \tag{5}$$

where $\mathcal{T}_c$ is the set of frames where contact is detected. The filter $\mathcal{F}$ projects both vertex sets to screen space and keeps the object vertices whose projections fall within a distance threshold of the contact body vertices, and $\mathcal{CD}_z(\cdot,\cdot)$ is a depth-only bidirectional Chamfer distance that penalizes the positional difference along the viewing direction. L_{cont} is disabled for terrain-only sequences without hand-object interaction.

Regularization. The regularization term decomposes as $L_{\text{reg}} = L_{\text{foot}} + L_{\text{vel}} + L_{\text{smooth}}$. L_{foot} leverages per-frame foot contact labels from GENMO (Li et al., 2025) to penalize foot vertex displacement during detected contact frames, suppressing foot skating. L_{vel} regularizes the optimized pelvis velocity to match GENMO's global-space velocity estimate, suppressing the depth-direction oscillations that camera-space estimates exhibit under depth-scale ambiguity. L_{smooth} penalizes the first- and second-order temporal finite differences of the human and object mesh vertex positions for temporal coherence.

3.3. Task-General Loco-Manipulation Tracking

This robot-proportioned reconstruction allows GMR (Araújo et al., 2025) to retarget the SMPL-X motion $\{\widetilde{\boldsymbol{\Theta}}_t^{\mathcal{H}}\}_{t=1}^T$ to the Unitree G1 with reduced morphology mismatch, better preserving hand-object and body-scene contacts. The result is a kinematic reference motion $\{\widetilde{\boldsymbol{q}}_t\}_{t=1}^T$ in the robot's joint space, while the reconstructed object trajectory $\{\widetilde{\boldsymbol{\Theta}}_t^{\mathcal{O}}\}_{t=1}^T$ provides the reference object pose. We then train tracking policies built on SONIC (Luo et al., 2025), a pretrained whole-body controller, to convert these retargeted 4D HOI trajectories into robot-action data. Rather than fitting a controller per sequence or per object, we train task-general policies across each task family; as related trajectories are added, existing policies provide initialization for fine-tuning, amortizing adaptation across the pool. As outlined in Fig. 3, we instantiate this stage with two complementary specializations: an *object-aware* latent adaptor trained on object-manipulation trajectories and a *scene-aware* tracker trained on terrain traversal and chair-sitting trajectories. The object-aware adaptor adds hand actions and modulates the latent tokens fed to the controller's frozen action decoder, enabling manipulation while preserving the locomotion prior; the scene-aware tracker fine-tunes the controller with a height-map encoder, improving terrain-conditioned whole-body control for traversal and scene interaction.

Object-Aware Tracking. For object-manipulation 4D HOI trajectories, we extend the pretrained whole-body controller with an object-aware adaptor policy π_ϕ that modulates its latent token space and emits hand actions, giving the frozen controller object-manipulation capability while preserving its pretrained locomotion behavior. The controller encodes kinematic motion targets into a discrete latent token $\boldsymbol{z}_t = \mathcal{E}(\widetilde{\boldsymbol{q}}_t)$ via finite scalar quantization and decodes them into joint-level actions through $\mathcal{G}(\boldsymbol{z}_t)$. We keep its encoder, quantizer, and

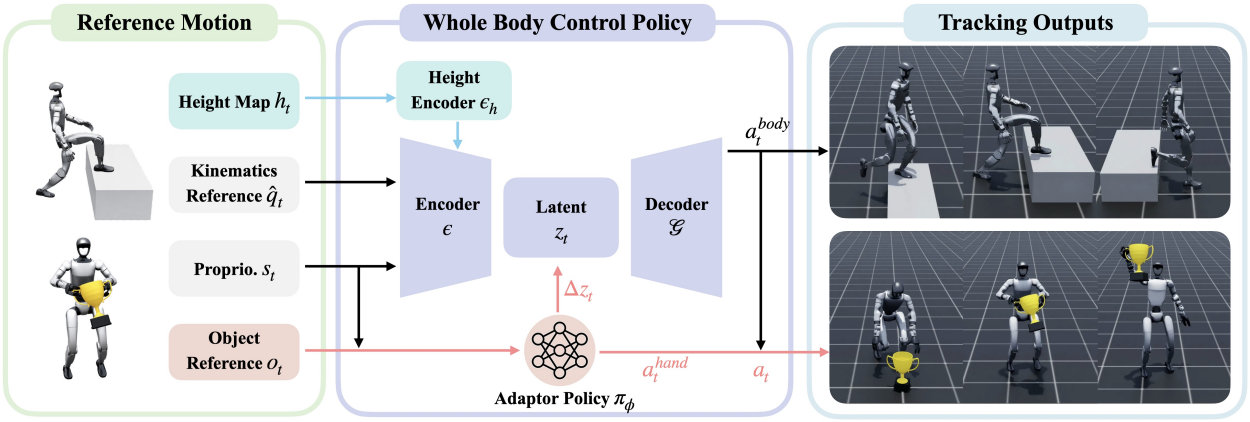

Figure 3: **Task-General Tracking via Complementary Controller Adaptation.** Retargeted 4D HOI trajectories are converted into robot-action data by adapting different parts of a pretrained whole-body controller. Object-manipulation trajectories are used to train an object-aware adaptor π_ϕ while the controller remains frozen: the adaptor observes proprioception $\boldsymbol{s}_t$ and an object reference $\boldsymbol{o}_t$, injects a latent residual $\Delta \boldsymbol{z}_t$, and emits hand actions $\boldsymbol{a}_t^{\text{hand}}$. Terrain traversal and chair-sitting trajectories are used to fine-tune the controller with a scene-aware height encoder ϵ_h: the encoder maps the local height map h_t into terrain context for scene-conditioned whole-body control.

decoder frozen, then train only π_ϕ to inject manipulation-specific residuals. The adaptor observes proprioception $\boldsymbol{s}_t$ and an object reference $\boldsymbol{o}_t$ consisting of the object pose in robot body frame, hand-to-object transforms, finger contact forces, a precomputed basis point set (BPS) (Prokudin et al., 2019) shape encoding, and, critically, delta observations encoding the difference between the reference future object pose and the current simulated pose. It outputs a 64-dim latent residual $\Delta \boldsymbol{z}_t$ together with a 2-dim binary hand primitive $\boldsymbol{a}_t^{\text{hand}}$ that maps to 7 finger DoFs per hand:

$$(\Delta \boldsymbol{z}_t, \boldsymbol{a}_t^{\text{hand}}) = \pi_\phi(\boldsymbol{s}_t, \boldsymbol{o}_t), \quad \boldsymbol{a}_t^{\text{body}} = \mathcal{G}(\boldsymbol{z}_t + \lambda \Delta \boldsymbol{z}_t), \tag{6}$$

with $\lambda = 0.1$ scaling the residual before FSQ quantization. Each hand primitive produces a binary open/close grasp signal, which is mapped to 7 finger joint positions per hand via predefined grasp configurations. The BPS encoding provides the adaptor with object-shape awareness, enabling a single π_ϕ to track motions across diverse object geometries. An auxiliary ℓ_2 penalty is applied on $\Delta \boldsymbol{z}_t$ to encourage the adapted latent to remain close to the pretrained controller's behavior.

Scene-Aware Tracking. For tasks involving scene-aware interactions, such as stepping over curbs, climbing up stairs, and sitting on chairs, the controller's flat-ground prior is not directly usable. We therefore fine-tune the controller on a mixture of reconstructed 4D HOI scene-interaction trajectories and its original flat-ground data, while augmenting the encoder input with a local height map h_t around the robot processed by a 2D-convolutional projector ϵ_h. This preserves the base locomotion distribution while teaching the controller to adapt whole-body motions to terrain and scene geometry. To stabilize the training, in addition to the action decoder $\mathcal{G}$, we train a parallel kinematic decoder $\mathcal{G}_{\text{rec}}$ that reconstructs the input motion targets to provide an auxiliary MSE loss that regularizes the latent to remain faithful to the trajectory.

Reward Design. Both trackers share a *motion-tracking reward* R_t^{motion} that encourages the simulated robot to follow the retargeted reference, together with regularization penalties R_t^{reg} (*e.g.*, action rate and joint limits) for smoothness and safety. The motion-tracking reward is a sum of exponential terms over reference–simulation discrepancies:

$$R_t^{\text{motion}} = \sum_i w_i \exp\left(-\frac{\|\tilde{\boldsymbol{x}}_{i,t} - \boldsymbol{x}_{i,t}\|^2}{\sigma_i^2}\right), \tag{7}$$

where $\tilde{\boldsymbol{x}}_{i,t}$ and $\boldsymbol{x}_{i,t}$ are reference and simulated quantities spanning root pose, per-body positions and orien-

tations, and linear and angular velocities; for object-aware tracking we additionally boost the weight on the wrist links to encourage accurate hand placement. For object-aware tracking, the total reward adds an object term and a contact-gated grasp term, $R_t = R_t^{\text{motion}} + R_t^{\text{reg}} + R_t^{\text{obj}} + \mathbb{1}\{C_t\}\, R_t^{\text{grasp}}$, where $\mathbb{1}\{C_t\}$ is a per-frame contact indicator carried by the reconstructed trajectory, so the grasp reward is active only during contact phases. The *object tracking reward* penalizes deviation from the reference object pose:

$$R_t^{\text{obj}} = w_p \exp\!\big(-\alpha_p \|\widetilde{\boldsymbol{p}}_t^{\mathcal{O}} - \boldsymbol{p}_t^{\mathcal{O}}\|\big) + w_r \exp\!\big(-\alpha_r \|\widetilde{\boldsymbol{r}}_t^{\mathcal{O}} \ominus \boldsymbol{r}_t^{\mathcal{O}}\|\big)\,, \tag{8}$$

with scaling coefficients $\alpha_p, \alpha_r > 0$. The *grasp reward* combines three terms per hand:

$$\begin{aligned} R_t^{\text{grasp}} = w_c \underbrace{\min\!\left(\frac{N_t^{\text{contact}}}{N_{\min}},\, 1\right)}_{\text{contact time}} + w_d \underbrace{\big[-\cos(\boldsymbol{d}_t^{\text{thumb}},\, \boldsymbol{d}_t^{\text{index}})\big]^+}_{\text{grasp pose}} \\ + \; w_f \underbrace{\exp\!\Big(-\gamma \tfrac{1}{N_f} \textstyle\sum_j \|\boldsymbol{f}_{j,t} - \boldsymbol{c}_t\|\Big)}_{\text{contact proximity}}. \end{aligned} \tag{9}$$

The first term rewards sustained finger contact with the object (saturating at $N_{\min}$ contacts), the second encourages the thumb and index finger to approach from opposing sides for a stable pinch grasp ($\boldsymbol{d}_t^{\text{thumb}}$, $\boldsymbol{d}_t^{\text{index}}$ are vectors from the object center to each fingertip), and the third draws all fingertips $\boldsymbol{f}_{j,t}$ toward the object contact centroid $\boldsymbol{c}_t$. Since scene-aware tasks involve no hand-object interaction, the scene-aware tracker uses only R_t^{motion} and R_t^{reg}.

Training. We train each stage with PPO (Schulman et al., 2017) in Isaac Lab on 64 NVIDIA L40 GPUs, running for 30,000 iterations with 1,024 environments per GPU. Object-aware tracking updates only π_ϕ while the controller's encoder, quantizer, and decoder remain frozen. Scene-aware tracking instead fine-tunes the controller together with the height-map encoder ϵ_h. Multiple reference motions are trained jointly within each task family, with environments sampling motions from a shared 4D HOI pool, and we apply reference state initialization at every episode reset.

Full reward definitions and implementation details are provided in Appendix B.

3.4. Sim-to-Real Deployment

For sim-to-real deployment, we distill the object-aware and scene-aware tracking policies into separate egocentric visual policies (Chi et al., 2023; He et al., 2025a; Luo et al., 2025) for object pick-up and stair-climbing, respectively. The deployed models consume head-camera RGB inputs and output the latent tokens of the SONIC controller, and are trained with domain randomization to facilitate sim-to-real transfer. To deploy on the real Unitree G1, we connect the robot to a desktop with an NVIDIA RTX 5090 GPU and stream visual and proprioceptive input to the desktop before streaming robot actions to the G1. We use a Luxonis OAK-D W camera on the G1 and run inference at 10 Hz.

4. Results

Our experiments cover the three stages of the GRAIL pipeline. We first evaluate whether the generated 4D HOI sequences are more physically executable than existing generation baselines. We then ask whether these 4D HOI sequences can be converted into task-general loco-manipulation policies at scale, rather than only replayed through per-sequence tracking. Finally, we demonstrate the practical value of the generated data by deploying egocentric visual policies on the real robot for autonomous loco-manipulation.

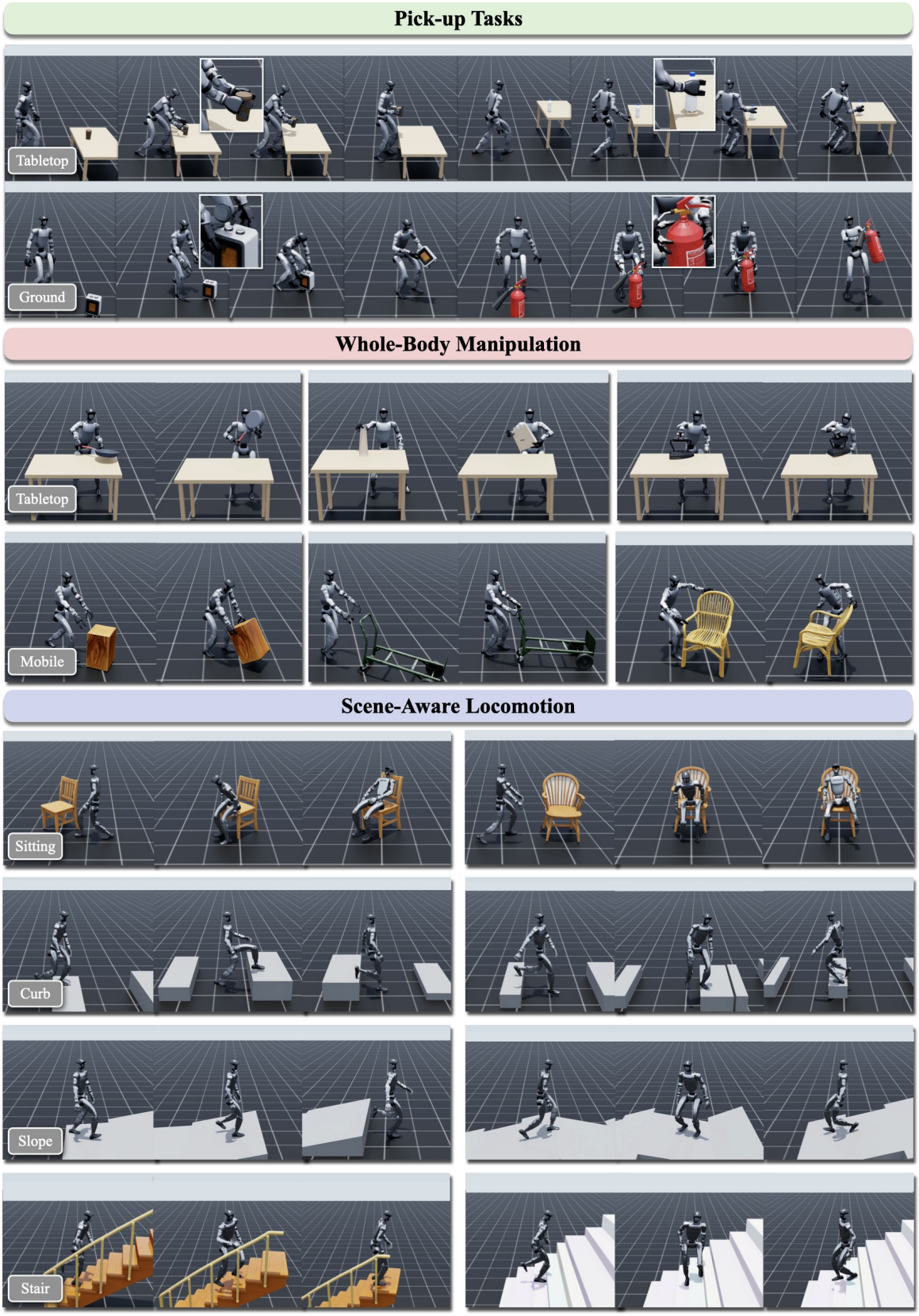

Figure 4: **Generated Loco-Manipulation Data.** Representative simulated Unitree G1 executions from the generated dataset span pick-up, whole-body manipulation, sitting, and terrain traversal across diverse objects and scene geometries.

Methods	Geometric		Perceptual	Smoothness		Physical Executability		
	Contact ↓	Pen. ↓	Inter. Score ↑	Human Smo. ↓	Obj Smo. ↓	SR ↑	Body Dev. ↓	Obj Dev. ↓
HOIDiff (Peng et al., 2025)	0.012	2.07%	1.79	0.0043	0.0118	15.8%	0.2120	0.3352
CHOIS (Li et al., 2024)	0.034	3.74%	2.47	0.0055	0.0062	10.5%	0.2564	0.3642
DAViD (Kim et al., 2025a)	0.246	1.46%	2.74	**0.0024**	0.0605	24.0%	0.4723	0.5826
Ours	**0.008**	**0.90%**	**3.58**	0.0033	**0.0022**	**88.9%**	**0.0913**	**0.0851**

Table 1: **Comparison of HOI Generation.** Geometric quality, perceptual realism (Interaction Score, 1-5 scale), motion smoothness, and physics-based tracking on the shared 20-object evaluation set.

Method	SR ↑	ObjPos ↓	MPJPE-L ↓
HDMI (Weng et al., 2025)	48.5%	0.283	122.3
ResMimic (Zhao et al., 2025)	49.2%	0.393	80.9
Ours w/o SONIC	45.0%	0.395	243.5
Ours w/o π_ϕ	39.7%	0.303	**37.1**
Ours w/o Rel. Obs.	57.9%	0.257	43.0
Ours (Full)	**81.4%**	**0.135**	41.8

Table 2: **Task-General Loco-Manipulation Tracking.** Comparison against loco-manipulation baselines (top) and ablations of GRAIL's object-aware adaptor (bottom). Metrics: success rate (SR), object position error (ObjPos), and local per-joint error (MPJPE-L).

4.1. Human-Object Interaction Generation

Setup. We compare the HOI generation component of GRAIL against training-based (CHOIS (Li et al., 2024), HOIDiff (Peng et al., 2025)) and training-free (DAViD (Kim et al., 2025a)) 4D HOI generation approaches on a shared evaluation set of 20 everyday objects from ComAsset (Kim et al., 2024). Detailed baseline configurations are provided in Appendix C.1.

Metrics. We evaluate the generated 4D HOI sequences along three axes. (i) *Geometric quality: contact distance* (Contact) is the average top-k vertex-to-vertex distance between the SMPL-X human surface and the object surface; *penetration ratio* (Pen.) is the percentage of SMPL-X vertices that interpenetrate the object mesh. (ii) *Perceptual realism: Interaction Score* (Inter. Score) is a VLM rating (OpenAI, 2024) of sampled keyframes on a 1-5 scale based on physical plausibility and affordance correctness; *motion smoothness* for the human (Human Smo.) and the object (Obj Smo.) is the second-order temporal derivative of their vertex trajectories. (iii) *Physical executability:* we apply InterMimic (Xu et al., 2025), an SMPL-X humanoid tracking framework, to reproduce each method's 4D HOI sequences in physics simulation using humanoids built from capsule primitives that conform to the input body shape (no motion retargeting required). We report the full-body mean per-joint position deviation (Body Dev.) and the mean object surface deviation (Obj Dev.), and define the tracking success rate (SR) as the fraction of frames where the normalized full-body and object deviations (divided by the object's maximum dimension) are both below 0.25.

Comparison. As shown in Table 1, GRAIL achieves the strongest performance across nearly all metrics: the lowest contact distance and penetration ratio, the highest interaction score, the smoothest object trajectories, and, by a large margin, the highest tracking success rate with the lowest body and object deviation. This confirms that the generated 4D HOI sequences are both perceptually realistic and physically executable, making them well-suited for downstream robot learning. We further conduct a user study and present additional qualitative comparisons in Appendix C.1.

	Seen Objects						Unseen Objects					
	Cube	Apple	Tea Box	Carrot	Wet Wipes	**Avg.**	Spray Can	Lint Roller	Peach	Flashlight	Medicine Bottle	**Avg.**
SR ↑	100%	60%	100%	70%	90%	**84%**	100%	50%	90%	80%	80%	**80%**

Table 3: **Pick-up results for seen and unseen objects.** Success rates are computed over 10 trials.

4.2. Task-General Loco-Manipulation Tracking

Scaling Loco-Manipulation Data. Using the asset-conditioned generation pipeline, we generate a large-scale loco-manipulation dataset for the Unitree G1 with 1,000 object assets sourced from Robocasa (Nasiriany et al., 2024), ComAsset (Kim et al., 2024), OMOMO (Li et al., 2023a), and Hunyuan3D (Team, 2025), paired with 1,000 procedurally generated terrain configurations. The resulting dataset contains over 20,000 sequences (Fig. 4) spanning four categories. *Pick-up* covers tabletops and the ground, exercising diverse grasp strategies across varying object shapes and placement heights. *Whole-body manipulation* covers tabletop manipulation motions and mobile interactions in which the robot carries, pushes, or repositions larger items such as boxes and carts while walking. *Sitting* spans diverse chair styles, requiring approach, lower-body adjustment, and settling into a seated posture. *Terrain traversal* covers procedurally generated curbs, slopes, and stairs, a setting essential for real-world deployment but underrepresented in existing datasets.

Baseline Comparison. We compare our method against two recent humanoid loco-manipulation baselines, HDMI (Weng et al., 2025) and ResMimic (Zhao et al., 2025), using their official implementations on a benchmark of 124 motions across 43 objects. We report manipulation success rate (SR), object position error (ObjPos), and local mean per-joint position error (MPJPE-L); SR is the fraction of episodes where the average object position error falls below 20 cm. Both train whole-body tracking policies from human references but differ from our approach in two key ways. First, neither method actuates per-finger DoFs, so their evaluated interactions rely on whole-arm or whole-body contact such as carrying, lifting, and pushing. Second, both train a separate policy per task: ResMimic trains a per-task residual on top of a general motion-tracking base, while HDMI trains one specialist policy per task. In contrast, GRAIL trains task-general policies across large in-family pools of 4D HOI trajectories. As shown in the top block of Table 2, GRAIL outperforms the baselines by a large margin across success rate, object position error, and body tracking accuracy.

Ablation Study. We ablate the object-aware latent adaptor for manipulation cases on the same benchmark and report results in Table 2 (bottom block). Removing SONIC and training from scratch substantially degrades body tracking and reduces success rate. Disabling the latent adaptor π_ϕ (i.e., vanilla SONIC) yields the lowest manipulation success rate despite the best body tracking, indicating that accurate body imitation alone is insufficient for object interaction. Replacing relative object observations with absolute ones also decreases success rate.

4.3. Sim-to-Real Deployment

To demonstrate GRAIL's real-world applicability, we deploy trained egocentric visual policies for stair-climbing and diverse object pick-up. For stair-climbing, a policy trained on diverse terrain-traversal sequences from GRAIL achieves a 90% real-world success rate, as shown in Fig. 5. For object pick-up, we train on 200 approach-and-pick-up sequences per object across cubes, apples, tea boxes, carrots, and wet wipes. The resulting policy achieves an 84% real-world success rate, as shown in Table 3, and transfers effectively to unseen objects, attaining an 80% success rate.

5. Conclusion

We presented GRAIL, a fully digital pipeline for generating humanoid loco-manipulation data from 3D assets and video priors, requiring the physical robot and environment only at deployment. Instead of reconstructing

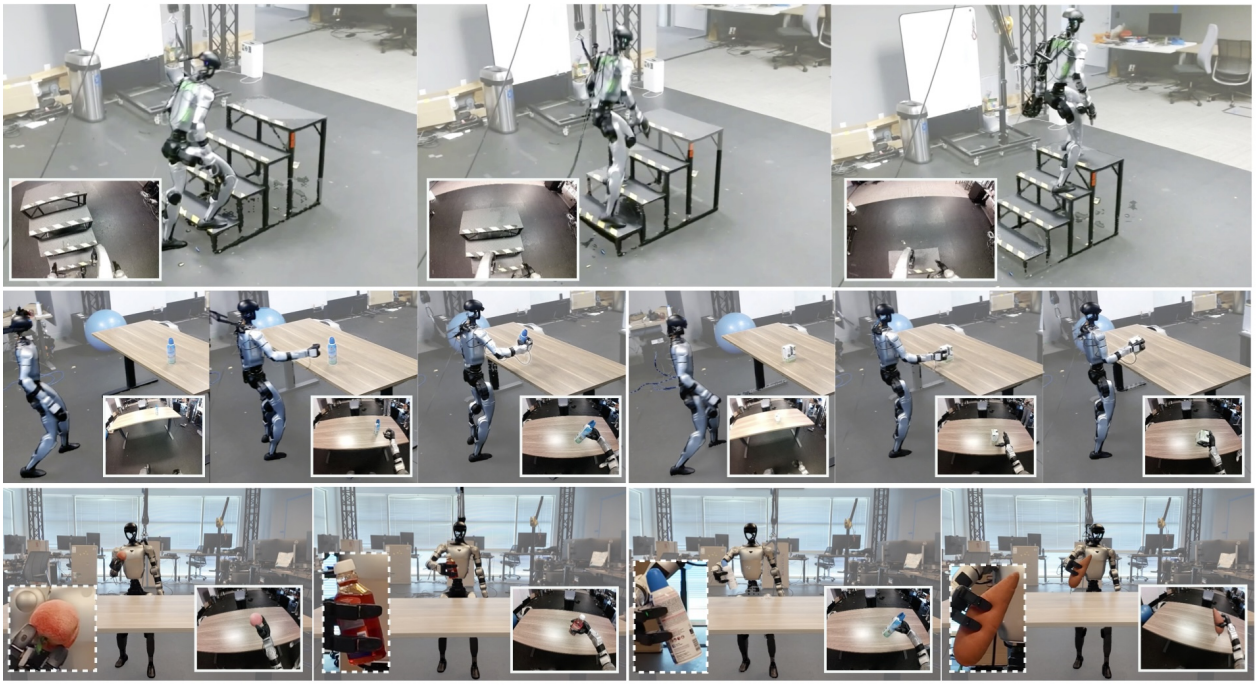

Figure 5: **Sim-to-Real Deployment.** Egocentric visual policies trained only on GRAIL-generated data transfer to a Unitree G1 for object pick-up and stair-climbing.

uncontrolled videos, GRAIL starts from fully specified 3D scenes where object geometry and texture, camera parameters, metric scale, environment depth, and robot-proportioned morphology are available by construction. This privileged setup turns key ambiguities in 4D HOI reconstruction into controlled inputs, enabling model-based object tracking, metric depth alignment, and interaction-aware optimization to recover robot-compatible trajectories. The recovered motions are retargeted to the Unitree G1 and converted into complementary task-general trackers: object-aware latent adaptation for manipulation and scene-aware height-map conditioning for terrain traversal and sitting. From over 20,000 generated sequences, we train egocentric visual policies using only GRAIL-generated data and deploy them on a real G1, achieving 84% pick-up success across diverse objects and 90% stair-climbing success. These results suggest that asset-conditioned generative data can complement teleoperation and motion capture as a scalable route toward humanoid loco-manipulation.

6. Limitations

Our pipeline assumes 3D object assets, simulator-ready scene setup, and a video foundation model that follows the requested interaction. Reconstruction quality degrades under severe occlusion, fast motion, or inconsistent object appearance from the VFM, and the failure-filtering step discards a non-trivial fraction of sequences. The task-general tracking policies amortize learning over related 4D HOI pools, but still require training or fine-tuning when the motion family changes substantially.

References

Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sanky Chan, Kenneth Draper, Debidatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, et al. ALOHA 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024. 2

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025. 3

João Pedro Araújo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking, 2025. URL https://arxiv.org/abs/2510.02252. ICRA 2026. 6

Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025. 2, 3

Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 3

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 8

Lingwei Dang, Ruizhi Shao, Hongwen Zhang, Wei Min, Yebin Liu, and Qingyao Wu. SViMo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios. In *NeurIPS*, 2025. 3

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2024. 3

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J Black, and Dimitrios Tzionas. Interactvlm: 3d interaction reasoning from 2d foundational models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22605–22615, 2025. 3

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 3

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024. 3

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning (CoRL)*, 2024. 3

Tairan He, Zi Wang, Haoru Xue, Qingwei Ben, Zhengyi Luo, Wenli Xiao, Ye Yuan, Xingye Da, Fernando Castañeda, Shankar Sastry, et al. Viral: Visual sim-to-real at scale for humanoid loco-manipulation. *arXiv preprint arXiv:2511.15200*, 2025a. 3, 8

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. HOVER: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025b. 3

Xialin He, Sirui Xu, Xinyao Li, Runpei Dong, Liuyu Bian, Yu-Xiong Wang, and Liang-Yan Gui. Ultra: Unified multimodal control for autonomous humanoid whole-body loco-manipulation. *arXiv preprint arXiv:2603.03279*, 2026. 3

Zhi Hou, Baosheng Yu, and Dacheng Tao. Compositional 3d human-object neural animation. *arXiv preprint arXiv:2304.14070*, 2023. 2, 3

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *GCPR*, 2022. 3

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. In *ICLR*, 2026. 3

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 3

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 3

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 2, 3

Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3D objects from pre-trained 2D diffusion models. In *European Conference on Computer Vision*, pages 400–419. Springer, 2024. 10, 11

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10330–10341, 2025a. 3, 10, 22, 23

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *CVPR*, 2025b. 3

Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. NCHO: Unsupervised learning for neural 3d composition of humans and objects. In *ICCV*, 2023. 2, 3

Kuaishou. Kling ai video generator (image-to-video model), 2025. URL https://klingai.com/. Version 2.1, image-to-video generative model. 4, 19

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 947–957, June 2024. 3

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742. 22

Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. In *3DV*, 2026. 3

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023a. 3, 11

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 3, 10, 22, 23

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A GENeralist model for human MOtion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 5, 6

Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20465–20474, 2024. 3

Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *arXiv preprint arXiv:2303.13129*, 2023b. 3

Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors. *arXiv preprint arXiv:2503.20118*, 2025. 3

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. *arXiv preprint arXiv:2504.10414*, 2025. 2, 3

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 3

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025. 2, 3, 6, 8, 19

Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pages 49–54, 2016. 4

Jiageng Mao, Siheng Zhao, Siqi Song, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, and Yue Wang. Learning from massive human videos for universal humanoid pose control, 2024. URL https://arxiv.org/abs/2412.14172. 3

Ruiqian Nai, Boyuan Zheng, Junming Zhao, Haodong Zhu, Sicong Dai, Zunhao Chen, Yihang Hu, Yingdong Hu, Tong Zhang, Chuan Wen, and Yang Gao. HuMI: Humanoid whole-body manipulation from robot-free demonstrations, 2026. URL https://arxiv.org/abs/2602.06643. 3

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024. 11

OpenAI. Chatgpt. https://chat.openai.com/, 2024. Large language model used for text generation and editing. 4, 6, 10, 19

Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations, 2025. URL https://arxiv.org/abs/2507.00990. 3

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 5

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 *CVPR 2025 Workshop of HuMoGen*, 2025. 3, 10, 22, 23

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 3

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 3

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *CVPR*, 2023. 2, 3

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 5

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 7

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12630–12641, 2023. 4, 19

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 19

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 5

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964. 5

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 8, 20

Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Yinghui Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration, 2026. URL https://arxiv.org/abs/2602.10106. 3

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 2, 3

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025. 11

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 5, 19

Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *3DV*, 2022. 2, 3

Yinhuai Wang, Qihan Zhao, Yuen Fui Lau, Runyi Yu, Hok Wai Tsui, Qifeng Chen, Jingbo Wang, Jiangmiao Pang, and Ping Tan. HumanX: Toward agile and generalizable humanoid interaction skills from human videos, 2026. URL https://arxiv.org/abs/2602.02473. 3

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 5

Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. HDMI: Learning interactive humanoid whole-body control from human videos, 2025. URL https://arxiv.org/abs/2509.16757. 3, 10, 11

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. 3

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *ECCV*, 2022. 2, 3

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. CARI4D: Category agnostic 4D reconstruction of human-object interaction. In *CVPR*, 2026. 2, 3

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 (ICCV)*, pages 14928–14940, October 2023. 3

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In *NeurIPS*, 2024. 3

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. 10

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 5

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. X-Humanoid: Robotize human videos to generate humanoid videos at scale, 2025. URL https://arxiv.org/abs/2512.04537. 3

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 3

Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C. Karen Liu, and Jiajun Wu. VisualMimic: Visual humanoid loco-manipulation via motion tracking and generation, 2025. URL https://arxiv.org/abs/2509.20322. 3

Justin Yu, Letian Fu, Huang Huang, Karim El-Refai, Rares Andrei Ambrus, Richard Cheng, Muhammad Zubair Irshad, and Ken Goldberg. Real2render2real: Scaling robot data without dynamics simulation or robot hardware. In *Conference on Robot Learning (CoRL)*, 2025. 3

Yanjie Ze, Zixuan Chen, Joao Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. TWIST: Teleoperated whole-body imitation system. In *Conference on Robot Learning (CoRL)*, 2025a. 2, 3

Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. TWIST2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025b. 2, 3

Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. *arXiv preprint arXiv:2309.03891*, 2023a. 3

Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. Interactanything: Zero-shot human object interaction synthesis via llm feedback and object affordance parsing. In *CVPR*, 2025. 3

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023b. 3

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3d human-object interactions. In *CVPR*, 2024. 3

Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning, 2025. URL https://arxiv.org/abs/2510.05070. 3, 10, 11

Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *CVPR*, 2023. 3

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object-to-hand correspondence for motion refinement. In *ECCV*, 2022. 3

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. 5

A. Human-Object Interaction Generation Details

A.1. Video Generation

We construct two candidate scene configurations using Infinigen (Raistrick et al., 2023): an indoor floor-only environment and a furnished room with a table (Fig. 6). For each input object, a VLM (OpenAI, 2024) determines whether it should be placed on the floor (*e.g.*, a sofa) or on the table (*e.g.*, a frypan) based on its typical affordances. We pair the object with a human asset prefitted to the Unitree G1's morphology, render the initial frame, and use a VLM (OpenAI, 2024) to generate a text prompt describing the intended interaction. The rendered frame and prompt are then passed to Kling 2.5 Turbo Pro (Kuaishou, 2025), which supports generating 5 or 10 second videos at 24 fps with resolution of 1920×1080. We can optionally render an ending frame to control the final positions of the human and object, and produce longer sequences auto-regressively by feeding the last frame of each segment as the initial frame of the next.

A.2. Generation Runtime

Table 4 reports the wall-clock time for each stage of the 4D HOI generation pipeline per sequence (a 5-second video at 24 fps, 121 frames), measured on a single NVIDIA A100 GPU. The full pipeline takes approximately 14 minutes per sequence. Video generation and initial motion estimation (human and object) together account for about 4 minutes. The optimization preprocessing stage, which runs MoGe-2 (Wang et al., 2025) for metric depth estimation and SAM2 (Ravi et al., 2024) for human and object segmentation to produce per-frame point clouds as optimization targets, takes roughly 2 minutes. The joint optimization stage dominates at approximately 8 minutes, as it jointly optimizes human and object trajectories across all frames.

Stage	**Time**
Video Generation (Kling API)	∼1 min
Human Motion Estimation	∼2 min
Object Pose Tracking	∼1 min
Optimization Preprocessing	∼2 min
Joint Optimization	∼8 min
Total	∼14 min

Table 4: **Runtime Breakdown.** Wall-clock time per stage of the GRAIL 4D HOI generation pipeline, measured on a single NVIDIA A100 GPU for a 5-second, 121-frame sequence.

A.3. Failure Case Filtering

While image-to-video models produce realistic HOI sequences, they may introduce artifacts such as texture inconsistencies or geometry mismatches across frames, causing FoundationPose to lose tracking. To automatically filter such failures, we compare SAM2 (Ravi et al., 2024) object masks $\{\widehat{\mathcal{M}}_t\}_{t=1}^{T}$ against rendered silhouettes $\{\mathcal{M}_t\}_{t=1}^{T}$ from the predicted poses, and compute the mask tracking error:

$$e^{\mathcal{M}} = \sum_{t=1}^{T} \frac{\text{Sum}\left((1-\mathcal{M}_t)\cdot\widehat{\mathcal{M}}_t\right)}{\text{Sum}\left(\mathcal{M}_t\right)}, \tag{10}$$

where $\text{Sum}(\cdot)$ counts non-zero pixels. This measures the fraction of SAM2-tracked mask pixels not covered by the predicted mask. We discard sequences where $e^{\mathcal{M}}$ exceeds $\tau = 0.2$, effectively removing cases caused by fast motion, blurry frames, or inconsistent object appearance.

B. Task-General Loco-Manipulation Tracking Details

We train two physics-based tracking policies on top of SONIC (Luo et al., 2025), a pretrained whole-body controller: an *object-aware adaptor* for manipulation 4D HOI trajectories and a *scene-aware tracker* for terrain- and chair-conditioned 4D HOI trajectories. Per-tracker policy observations are summarized in Table 5 and reward terms in Table 6.

Figure 6: **Candidate Scene Configurations.** Two pre-built 3D scene templates used to place objects: an indoor floor-only environment for ground-level objects (*e.g.*, a sofa), and a furnished room with a table for tabletop objects (*e.g.*, a frypan). A VLM selects between the two based on each object's typical affordances.

B.1. Object-Aware Adaptor

Architecture. The adaptor policy π_ϕ is a 3-layer MLP with hidden dimensions $[512, 256, 128]$ and SiLU activations. It outputs a 66-dim meta-action: 64 dims for the latent residual (matching the controller's token dim of 2×32) and 2 dims for left/right hand primitives. The residual is scaled by $\lambda = 0.1$ and added to the controller's encoder output *before* finite scalar quantization, then decoded to 29 body joint position targets. Each hand primitive is passed through a sigmoid and thresholded to a binary open/close signal, mapped to 7 finger joint positions per hand via predefined grasp configurations. The critic is a separate 3-layer MLP $[512, 256, 128]$ that receives privileged observations including full body state and ground-truth contact flags.

Training. We train π_ϕ with PPO (Schulman et al., 2017) on 64 NVIDIA L40 GPUs with 1,024 parallel environments per GPU in Isaac Lab for 30,000 iterations; the pretrained encoder, FSQ quantizer, and decoder remain frozen and only π_ϕ is updated. Environments sample reference motions from a shared task-family 4D HOI pool, and we apply reference state initialization at every episode reset: the start frame is sampled uniformly from the first 30 frames and clipped to precede the labeled hand-object contact. PPO hyperparameters: actor learning rate 2×10^{-5} with an adaptive schedule targeting KL $= 0.01$, critic learning rate 10^{-3}, discount factor $\gamma = 0.99$, GAE parameter $\lambda_{\text{GAE}} = 0.95$, clipping parameter $\epsilon = 0.2$, entropy coefficient 0.01, 24 steps per environment, 5 learning epochs with 4 mini-batches, and maximum gradient norm 0.1. Episodes terminate when the object's z-position deviates by more than $0.4\,\text{m}$ from the reference, the root height deviates by more than $0.25\,\text{m}$, or the root orientation error exceeds $1.0\,\text{rad}$.

B.2. Scene-Aware Tracker

For scene-level interactions such as stepping over curbs, traversing slopes and stairs, or sitting on chairs, the controller's flat-ground prior is not directly applicable. Instead of attaching a latent adaptor, we fine-tune the controller end-to-end together with a height-map encoder ϵ_h and an auxiliary kinematic decoder $\mathcal{G}_{\text{rec}}$ on the reconstructed scene-aware data.

Architecture. We construct an 11×11 height map grid centered on the robot with a total extent of $1.5\,\text{m}$ and a resolution of $0.15\,\text{m}$. At each grid point, a downward ray is cast against the scene mesh to obtain the terrain hit position; positions are then transformed into the robot's yaw-aligned local frame, yielding a $(11, 11, 3)$ tensor. The height map is processed by a 3-layer CNN with channels $[64, 128, 256]$, kernel size 3×3, stride 2, and LeakyReLU activations; spatial dimensions reduce as $11 \to 6 \to 3 \to 2$, and the output is flattened to a 1,024-dim feature vector. This vector is concatenated with the proprioceptive observation and the controller's tokenizer features, then passed through a fusion MLP ($[256]$, SiLU) that produces the latent input to the controller's motion decoder.

Training. We train with PPO (Schulman et al., 2017) on 64 NVIDIA L40 GPUs with 1,024 parallel environments per GPU in Isaac Lab for 30,000 iterations. Unlike object-aware tracking, we fine-tune the controller (encoder, FSQ quantizer, and action decoder $\mathcal{G}$) end-to-end together with the height-map encoder ϵ_h and the parallel kinematic decoder $\mathcal{G}_{\text{rec}}$. $\mathcal{G}_{\text{rec}}$ reconstructs the input motion targets to provide an auxiliary MSE loss (weight 0.01) that regularizes the latent to remain faithful to the reference. Reference state initialization is sampled

Term	Description	Dim	Used
Proprioception			
Joint position	robot joint angles	43 / 29	Both
Joint velocity	robot joint angular velocities	43 / 29	Both
Base angular velocity	in body frame	3	Both
Previous action	last meta-action	66 / 29	Both
Base linear velocity	in body frame	3	O
Gravity direction	in body frame	3	S
Reference motion targets			
Motion anchor position	target root position in body frame	3	O
Motion anchor orientation	target root orientation 6D	6	O
Current command	ref. joint position + velocity, current frame	58	O
Multi-future command	ref. joint position + velocity, 10 future frames	580	O
Multi-future anchor ori	10 future root orientations 6D	60	O
Object state			
Object position	current object pos in body frame	3	Both
Object orientation 6D	current object ori in body frame	6	Both
Object pos delta (10 fut.)	$\mathrm{ref}_{\mathrm{fut}} - \mathrm{sim}_{\mathrm{cur}}$, position	30	Both
Object ori delta (10 fut.)	relative rotation, 6D	60	Both
Target object position	reference object pos in base frame	3	O
Object BPS encoding	per-object shape descriptor	10	O
Table position	in body frame	3	O
Table orientation 6D	in body frame	6	O
Hand-object contact			
Hand-to-object transform	per right hand (3 pos + 6 ori)	9	O
Fingertip contact forces	3D force per fingertip	12	O
Scene			
Local height map	11×11 terrain z in body frame	121	S

Table 5: **Policy Observations.** Observations for the object-aware adaptor (O) and the scene-aware tracker (S), grouped by category. When two dims are listed (*e.g.*, 43 / 29), the first applies to the object-aware policy (43-DoF G1 including fingers) and the second to the scene-aware policy (29-DoF G1, no fingers). Proprioceptive and previous-action observations are stored as 10-frame histories; instantaneous values are listed above.

uniformly across the full motion at every episode reset. Since hand-object interaction is not involved in these tasks, manipulation-specific reward terms (grasp and object-pose tracking) are disabled. Episodes terminate under cumulative tracking-error thresholds with adaptive strict orientation and foot xyz constraints.

B.3. Training Cost

Because each tracking policy is trained jointly over a shared task-family pool rather than fit per sequence, we report the amortized training cost per motion, defined as the total training wall-clock of a run divided by the number of motions in its pool. A full policy is trained for $30{,}000$ PPO iterations on 64 NVIDIA L40 GPUs with $1{,}024$ environments per GPU, which takes roughly 30 hours, and each run trains $2{,}000$–$4{,}000$ motions jointly. The amortized cost is therefore only about 0.5–0.9 minutes per motion, far below the per-sequence cost of fitting a controller to each trajectory in isolation. As the pool grows with additional in-family motions, we do not retrain from scratch: we warm-start from the current policy and fine-tune, which typically converges within $6{,}000$ iterations (about one fifth of a full run, $\sim$6 hours), reducing the amortized cost of incorporating new motions to roughly one fifth of the full-training figure.

Term	Formula	Weight	Used
Motion tracking reward			
Anchor position	$\exp\big(-\|\widetilde{\boldsymbol{p}}_t^{\text{root}} - \boldsymbol{p}_t^{\text{root}}\|^2/\sigma^2\big)$	0.5	Both
Anchor orientation	$\exp\big(-\|\widetilde{\boldsymbol{r}}_t^{\text{root}} \ominus \boldsymbol{r}_t^{\text{root}}\|^2/\sigma^2\big)$	2.5	Both
Relative body position	$\exp\big(-\frac{1}{N}\sum_i \|\widetilde{\boldsymbol{p}}_{i,t} - \boldsymbol{p}_{i,t}\|^2/\sigma^2\big)$	1.0	Both
Relative body orientation	$\exp\big(-\frac{1}{N}\sum_i \|\widetilde{\boldsymbol{r}}_{i,t} \ominus \boldsymbol{r}_{i,t}\|^2/\sigma^2\big)$	5.0	Both
Body linear velocity	$\exp\big(-\frac{1}{N}\sum_i \|\widetilde{\boldsymbol{v}}_{i,t} - \boldsymbol{v}_{i,t}\|^2/\sigma^2\big)$	1.0	Both
Body angular velocity	$\exp\big(-\frac{1}{N}\sum_i \|\widetilde{\boldsymbol{\omega}}_{i,t} - \boldsymbol{\omega}_{i,t}\|^2/\sigma^2\big)$	1.0	Both
5-point local body	$\exp\big(-\frac{1}{\lvert\mathcal{S}_5\rvert}\sum_{i\in\mathcal{S}_5} \|\widetilde{\boldsymbol{p}}_{i,t}^{\text{loc}} - \boldsymbol{p}_{i,t}^{\text{loc}}\|^2/\sigma^2\big)$	2.0	S
Object reward			
Object pose tracking	$w_p \exp\big(-\alpha_p\|\widetilde{\boldsymbol{p}}_t^{\mathcal{O}} - \boldsymbol{p}_t^{\mathcal{O}}\|\big) + w_r \exp\big(-\alpha_r\|\widetilde{\boldsymbol{r}}_t^{\mathcal{O}} \ominus \boldsymbol{r}_t^{\mathcal{O}}\|\big)$	20.0	O
Grasp reward			
Grasp contact count	$\min\big(N_t^{\text{contact}}/N_{\min},\ 1\big)$	40.0	O
Grasp finger direction	$-\cos(\boldsymbol{d}_t^{\text{thumb}}, \boldsymbol{d}_t^{\text{index}})\,\mathbb{1}\{C_t\}$	10.0	O
Grasp contact center	$\exp\big(-\gamma\,\frac{1}{N_f}\sum_j \|\boldsymbol{f}_{j,t} - \boldsymbol{c}_t\|\big)\,\mathbb{1}\{C_t\}$	0.1	O
Regularization			
Latent residual ℓ_2	$\|\Delta\boldsymbol{z}_t\|^2$	0.1	O
Finger-primitive limit	$\mathbb{1}\big[\lvert a_{j,t}^{\text{fp}}\rvert > 0.5\big]$	-10.0	O
Action rate ℓ_2	$\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|^2$	-0.1	S
Anti-shake angular velocity	$\frac{1}{\lvert\mathcal{B}\rvert}\sum_{i\in\mathcal{B}}\big[\max(0, \|\boldsymbol{\omega}_{i,t}\| - \tau)\big]^2$	-5×10^{-3}	S
Ankle joint acceleration	$\sum_{j\in\mathcal{J}_{\text{ankle}}} \ddot{q}_{j,t}^2$	-2.5×10^{-7}	S
Kinematic reconstruction	$\|\widetilde{\boldsymbol{q}}_t - \mathcal{G}_{\text{rec}}(\boldsymbol{z}_t)\|^2$	0.01	S
Joint limit	$\sum_j \max\big(0,\ \lvert q_{j,t}\rvert - q_j^{\lim}\big)$	-10.0	S
Undesired body contact	$N_t^{\text{undesired}}$	-0.1	Both

Table 6: **Reward Terms.** Rewards for the object-aware adaptor (O) and the scene-aware tracker (S), organized by the four categories from Sec. 3.3. The rightmost column indicates which tracker uses each term. For grasp terms, the listed weight is the right-hand value; in object-aware training, the left hand uses half the listed weight. $\mathbb{1}\{C_t\}$ is the per-frame motion-label contact indicator that gates the grasp finger-direction and contact-center terms. The object pose-tracking reward is additionally gated by a simulated finger–object contact indicator, so it is active only while the hand is in contact with the object. For anti-shake, $\mathcal{B} = \{$left wrist, right wrist, head$\}$ with deadzone $\tau = 1.5$ rad/s. Per-term Gaussian-kernel bandwidth σ varies by term and is given in the released config. Negative weights indicate penalties.

C. Experiment Details

C.1. Human-Object Interaction Generation

Baselines. We compare against training-based and training-free 4D HOI generation approaches. Training-based baselines include CHOIS (Li et al., 2024), a controllable motion generation framework guided by language and sparse waypoints, and HOIDiff (Peng et al., 2025), a diffusion-based model for affordance-conditioned HOI synthesis. For training-free comparison, we evaluate DAViD (Kim et al., 2025a), which generates the first frame using an image generative model (Labs et al., 2025) and produces the video from that frame. Since image generation often fails under partial control signals (*e.g.*, Canny edge maps), we generate 24 images per object and manually select a successful result as the starting frame. To ensure fairness, DAViD uses Kling's image-to-video model under the same setting as our approach.

User Study. Beyond the quantitative metrics reported in the main paper, we conduct a user study with 30 participants to assess perceptual quality. In each trial, participants view sequences from three of the four methods drawn at random and select the result with the most appropriate object affordances (Aff. Real.) and the most physically plausible motion (Phys. Real.). As shown in Table 7, GRAIL is preferred by a wide margin on both criteria. This perceptual study complements, but does not replace, the physics-based tracking and robot-execution metrics reported in the main paper.

Methods	Aff. Real. ↑	Phys. Real. ↑
HOIDiff (Peng et al., 2025)	2.0%	1.9%
CHOIS (Li et al., 2024)	12.2%	16.8%
DAViD (Kim et al., 2025a)	11.2%	10.4%
Ours	**74.7%**	**70.9%**

Table 7: **User Study.** 30-participant pick rates for the most appropriate object affordances (Aff. Real.) and the most physically plausible motion (Phys. Real.) on the shared 20-object evaluation set. Pick rates have a theoretical upper bound of 75% under 3-of-4 random sampling.

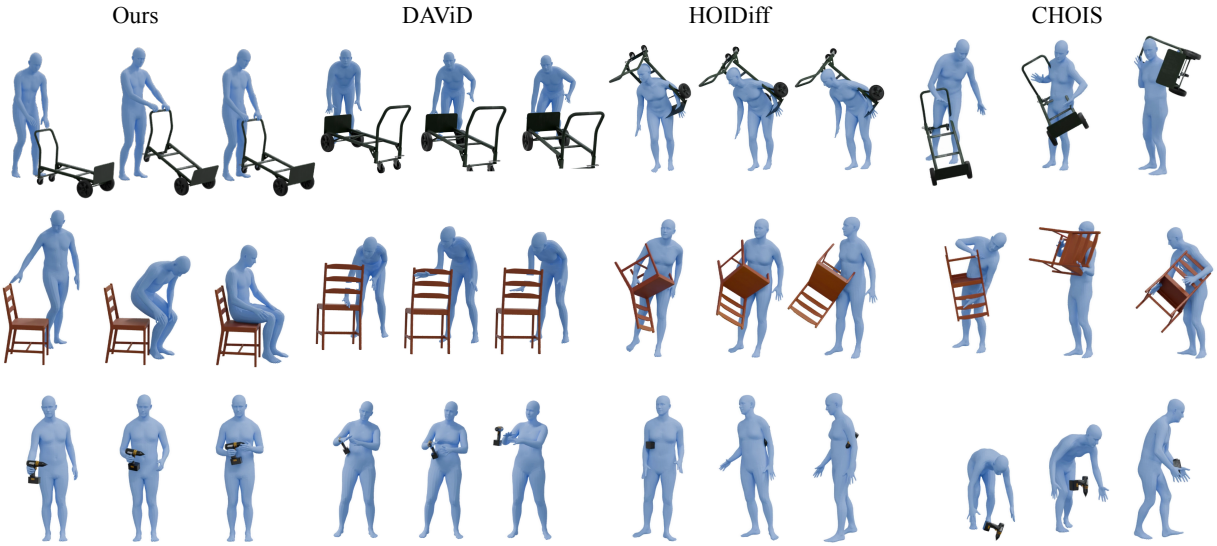

Figure 7: **Qualitative Comparison.** Representative 4D HOI sequences from each method on the shared 20-object evaluation set. GRAIL produces more coherent motions with accurate contact and natural hand poses, while baseline methods often yield unrealistically flat or static hand configurations.

	Reconstruction Quality					**Tracking Quality**		
Methods	Contact ↓	Pen. ↓	MPJPE ↓	Human Smo. ↓	Obj Smo. ↓	SR ↑	ObjPos ↓	MPJPE-L ↓
w/o L_{proj}	0.016	1.93%	16.70	0.0023	0.0011	41.6%	0.374	47.1
w/o L_{depth}	0.017	1.97%	4.35	0.0020	0.0011	42.6%	0.372	49.3
w/o L_{cont}	0.024	1.52%	4.81	0.0022	0.0009	53.3%	0.332	52.4
Ours (Full)	0.015	1.81%	4.89	0.0020	0.0009	81.4%	0.135	41.8

Table 8: **Ablation Study on Reconstruction Losses.** Reconstruction quality (Contact, Pen., MPJPE in pixel space, motion smoothness) and downstream tracking quality (SR, ObjPos, MPJPE-L) for each loss ablation. Each loss targets a different failure mode in reconstruction; the full model achieves the best downstream tracking despite not minimizing every reconstruction proxy in isolation.

Qualitative Comparison. Fig. 7 shows that GRAIL produces more coherent motions with accurate contact and natural hand poses, while baseline methods often yield unrealistically flat or static hand configurations unsuitable for downstream humanoid skill learning.

Reconstruction Ablation. We ablate the interaction-aware reconstruction losses (L_{proj}, L_{depth}, L_{cont}) on the 124-motion benchmark. As shown in Table 8, removing L_{proj} degrades image-space consistency, while removing L_{depth} or L_{cont} weakens metric interaction quality. These reconstruction-level errors propagate to downstream tracking: each ablation substantially reduces tracking success rate and increases trajectory deviation, while the

full model achieves the best overall downstream tracking quality.