SITUHOI: SCALABLE CAPTURE OF SCENE-SITUATED FULL-BODY HUMAN-OBJECT INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modeling human-object interaction (HOI) in real environments requires jointly representing whole-body motion, manipulated objects, and scene geometry, yet existing datasets trade off scalable capture against structured, scene-aligned supervision. We introduce SituHOI, a large-scale RGB-D dataset containing 18,429 annotated clips across 16 indoor rooms and approximately 86 hours of quality-screened activity. It pairs recordings from a single moving RGB-D camera with static scene scans, object geometry, language descriptions, and estimated human and object trajectories. Using the scans as persistent references, our pipeline combines camera localization and depth-based metric alignment with human mesh recovery, object tracking, and temporal refinement to register all geometric annotations in a shared floor-aligned frame. These data support FlexHOI-Bench, a goal-conditioned joint motion prediction benchmark with variable observation and prediction durations. Given observed histories, scene and object geometry, and a terminal object pose, models predict future human and object motion. We provide SharedHOI as a reference predictor and evaluate world-space accuracy and human-object coordination. On a MoCap subset, our pipeline reduces human G-MPJPE from 195.4 to 68.3 mm and object translation error from 43.0 to 28.7 mm. SharedHOI leads four adapted baselines on all three aggregate FlexHOI-Bench metrics. Data will be released upon acceptance. An anonymized code repository is available at https://anonymous.4open.science/r/iclr2027-anonymous-code-1DFB.

1 INTRODUCTION

Recent advances in human motion generation have enabled increasingly realistic and controllable whole-body motion (Tevet et al., 2023; Karunratanakul et al., 2023). However, the growing interest in humanoid robots and embodied agents is driving a shift from modeling human motion in isolation to scene-grounded human-object interaction (He et al., 2025; Xu et al., 2025b). In this setting, human motion must respect scene geometry and contact constraints while remaining temporally coordinated with the motion of manipulated objects (Diller & Dai, 2024; Li et al., 2024), making the problem substantially more challenging than conventional motion generation.

Learning such interactions requires data spanning diverse scenes, objects, and motions (Bhatnagar et al., 2022; Kim et al., 2025). Egocentric datasets collected with head-mounted cameras provide a convenient means of recording hand-object interactions (Liu et al., 2022; Banerjee et al., 2025; Hoque et al., 2026), but their restricted field of view and frequent occlusions make accurate full-body motion recovery difficult without additional sensing. Other datasets rely on marker-based motion-capture (MoCap) systems, which provide accurate motion measurements but require costly, instrumented environments (Taheri et al., 2020), or calibrated third-person multiview RGB systems, which demand substantial camera setup and provide no direct depth measurements (Zhang et al., 2023). Existing capture paradigms therefore trade off scalability and scene diversity against full-body coverage and geometric accuracy.

RGB-D cameras offer a promising balance between flexible visual capture and direct metric depth. Existing full-body RGB-D interaction datasets typically employ multiple fixed cameras to improve coverage and reduce occlusion (Bhatnagar et al., 2022; Huang et al., 2024; Kogashi et al., 2026), but these systems require installation, synchronization, and calibration in each new environment , limiting long-form collection across diverse indoor spaces. A single moving RGB-D camera provides a more deployable alternative by following interactions beyond a fixed volume while retaining metric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

depth. However, its moving viewpoint and lack of multiview constraints make camera, human, and object motion difficult to disentangle and register consistently in a common scene frame.

To address this challenge, we develop a capture and reconstruction pipeline that separates static-scene reconstruction from dynamic-activity recording. Each room is scanned once to construct a metric, gravity-aligned scene model, after which activities are recorded with a single handheld RGB-D camera. We combine scene-referenced camera localization with depth-based human fitting, object-tracking recovery, and temporal refinement to obtain aligned motion estimates from the moving viewpoint. The protocol replaces per-scene multiview setup with a one-time scene scan and object-asset preparation, and requires no body- or object-mounted trackers.

Using this pipeline, we introduce ***SituHOI***, a large-scale RGB-D dataset of scene-grounded full-body human-object interaction. Curated from roughly 90 hours of raw recordings, SituHOI comprises approximately 86 hours of activity across 16 indoor rooms, segmented into 18,429 language-annotated clips, including 10,335 dynamic-object clips covering 48.64 hours. Each recording includes RGB-D observations, reconstructed scene geometry, camera poses, and object assets, while each clip includes a natural-language description, estimated full-body human motion, and, where applicable, a rigid-object trajectory. All geometric annotations share a metric scene frame. To our knowledge, SituHOI is the first full-body HOI dataset to provide this combination at scale using a single moving RGB-D camera without body- or object-mounted trackers (Tab. 1).

Building on SituHOI, we introduce **FlexHOI-Bench** for *goal-conditioned joint human-object motion prediction* across variable observation and forecast durations. Given motion histories, aligned scene and object geometry, and a terminal object pose, a single model predicts future human and object motion for the requested time horizons. We further propose **SharedHOI**, a unified predictor of both motions and their relative coordination, and evaluate it using complementary world-space and body-relative metrics that capture global accuracy and interaction coherence.

In summary, our contributions are:

- *A large-scale, scene-grounded full-body HOI dataset.* We introduce SituHOI, comprising approximately 86 hours of RGB-D activity across 16 indoor rooms and 18,429 language-annotated intervals, including 10,335 intervals involving dynamic objects. The dataset provides reconstructed room-scale scenes and estimated camera, human, and rigid-object trajectories in a shared metric scene frame.
- *A scalable single-camera capture and reconstruction pipeline.* By separating one-time scene reconstruction from activity recording, our pipeline recovers temporally consistent camera, full-body human, and rigid-object motion from a single moving RGB-D stream. It requires neither a calibrated multi-camera rig nor body- or object-mounted trackers, facilitating deployment across diverse indoor environments.
- *A unified benchmark and model for goal-conditioned joint motion prediction.* We introduce FlexHOI-Bench, which maps observed human-object motion, scene and object geometry, and a target object pose to the future motions of both the human and the object. We further propose **SharedHOI** and evaluate it using complementary world-space and body-relative metrics that assess both global motion accuracy and human-object coordination.

2 Related Work

Scene-grounded human-object interaction datasets. Existing datasets cover scene-aware full-body motion and egocentric body capture in reconstructed environments (Hassan et al., 2019; Huang et al., 2022a; Zhang et al., 2022). Other resources capture full-body human-object interaction using motion capture or multiview RGB-D (Taheri et al., 2020; Lu et al., 2025; Bhatnagar et al., 2022; Huang et al., 2022b), or focus on articulated and egocentric hand-object manipulation (Fan et al., 2023; Liu et al., 2022; Banerjee et al., 2025; Hoque et al., 2026). Recent datasets further extend interaction capture to contextual environments, collaborative activities, and multiple humans or objects (Zhang et al., 2024; Kim et al., 2025; Liu et al., 2025; Kogashi et al., 2026). Across these efforts, complete human and object motion, explicit scene geometry, semantic supervision, and portable capture remain difficult to obtain jointly. SituHOI combines these properties through single-camera RGB-D observations, scene-aligned motion, room geometry, and language.

Scene-aligned interaction reconstruction. Constructing scene-aligned interaction data requires camera and scene reconstruction (Schönberger & Frahm, 2016; Sarlin et al., 2019; Fang et al., 2022),

Table 1: Comparison of capture setups, hours, sequence/segment counts, environments (Envs.), and annotations. Full-body 3D includes arms and legs; Object pose denotes dynamic-object position and orientation. Scene geom. includes point clouds and layouts; Scene mesh requires environment or aligned static-furniture meshes, not isolated manipulated-object meshes. Semantics covers action labels and language descriptions. Uninstr. excludes body/object tracking attachments, but permits capture cameras and calibration targets. ✓: documented support; ✗: criterion unmet or support unestablished; NR: not reported. Definitions and sources: Appendix A.3.

Dataset	Capture	Hours	Seq./seg.	Envs.	Full-body 3D	Object pose	Scene geom.	Scene mesh	Semantics	Uninstr.
Full-body human-scene capture										
PROX	1 RGB-D	NR	60 seq.	12	✓	✗	✓	✓	✗	✓
RICH	6–8 RGB	NR	142 seq.	5	✓	✗	✓	✓	✓	✓
Full-body interaction capture										
GRAB	54-cam MoCap	NR	1,334 seq.	NR	✓	✓	✗	✗	✓	✗
BEHAVE	4 RGB-D	NR	321 seq.	5	✓	✓	✗	✗	✗	✓
InterCap	6 RGB-D	NR	223 seq.	1	✓	✓	✗	✗	✗	✓
CHAIRS	4 RGB-D + hybrid MoCap	17.3	1,390 seq.	NR	✓	✓	✗	✗	✓	✗
OMOMO	12-cam MoCap	∼10	NR	NR	✓	✓	✗	✗	✓	✗
IMHD^2	32 RGB + IMUs	NR	295 seq.	NR	✓	✓	✗	✗	✗	✗
HOI-M^3	42 RGB + IMUs	>20	199 seq.	5	✓	✓	✓	✓	✗	✗
HIMO	20-cam MoCap + gloves	9.44	3,376 seq.	NR	✓	✓	✗	✗	✓	✗
ParaHome	70 RGB + suit/gloves	8.10	207 seq.	1	✓	✓	✓	✓	✓	✗
HUMOTO	2 RGB-D + suit/gloves	2.19	735 seq.	NR	✓	✓	✓	✓	✓	✗
MMHOI	4 RGB-D	NR	NR	NR	✓	✓	✗	✗	✓	✓
ACE-Data-0	8 exo + 4 ego + MoCap	150	75,000 ep.	2	✓	✓	✓	✗	✓	✗
HiPHI	Optical MoCap	245.7	NR	NR	✓	✓	✗	✗	✓	✗
Consolidated / augmented benchmark										
InterAct / -X	7 datasets + optimization	21.81 base 30.70 aug.	11,350 base 16,201 aug.	NR	✓	✓	✗	✗	✓	✗
Egocentric hand-object interaction										
HOI4D	Egocentric RGB-D	NR	4,000 seq.	610	✗	✓	✓	✗	✓	✓
HOT3D	Aria / Quest 3 + MoCap	13.88	NR	1	✗	✓	✓	✗	✗	✗
EgoDex	Vision Pro	829	338,000 ep.	NR	✗	✗	✗	✗	✓	✓
SituHOI	**1 RGB-D + prior scans**	**∼86 total 48.64 dyn.**	**10,335 seq. (dynamic)**	**16**	✓	✓	✓	✓	✓	✓

expressive body recovery (Pavlakos et al., 2019; Lin et al., 2023), and object segmentation and tracking (Ravi et al., 2025; Wen et al., 2024). These components provide complementary estimates but do not automatically share consistent scales, coordinate frames, or temporal conventions. Our pipeline integrates them through depth-based alignment, floor referencing, temporal refinement, and quality filtering to produce a unified metric representation.

Human-motion interaction modeling. Human-object models forecast interactions from histories, generate human motion from prescribed object trajectories, or jointly synthesize human and object motion conditioned on language, contacts, or sparse waypoints (Xu et al., 2023; Li et al., 2023; Diller & Dai, 2024; Li et al., 2024). These formulations provide different amounts of future object information. FlexHOI-Bench instead conditions on interaction history, aligned scene and object geometry, and only a terminal object pose, while requiring both future human and object motion. Its world-space and body-relative evaluation measures global accuracy and interaction coordination.

3 SituHOI Dataset

To capture scene-grounded human-object interactions with a single handheld RGB-D camera, we collect three complementary types of data: static room scans for scene reconstruction and camera localization, dedicated object scans for object mesh reconstruction and pose tracking, and interaction recordings for human and object tracking (Fig. 1). With this pipeline, we build a novel dataset **SituHOI** with RGB-D observations, human and object trajectories, and geometry in a shared metric frame for each of 16 indoor rooms. After screening, the dataset contains 18,429 language-annotated

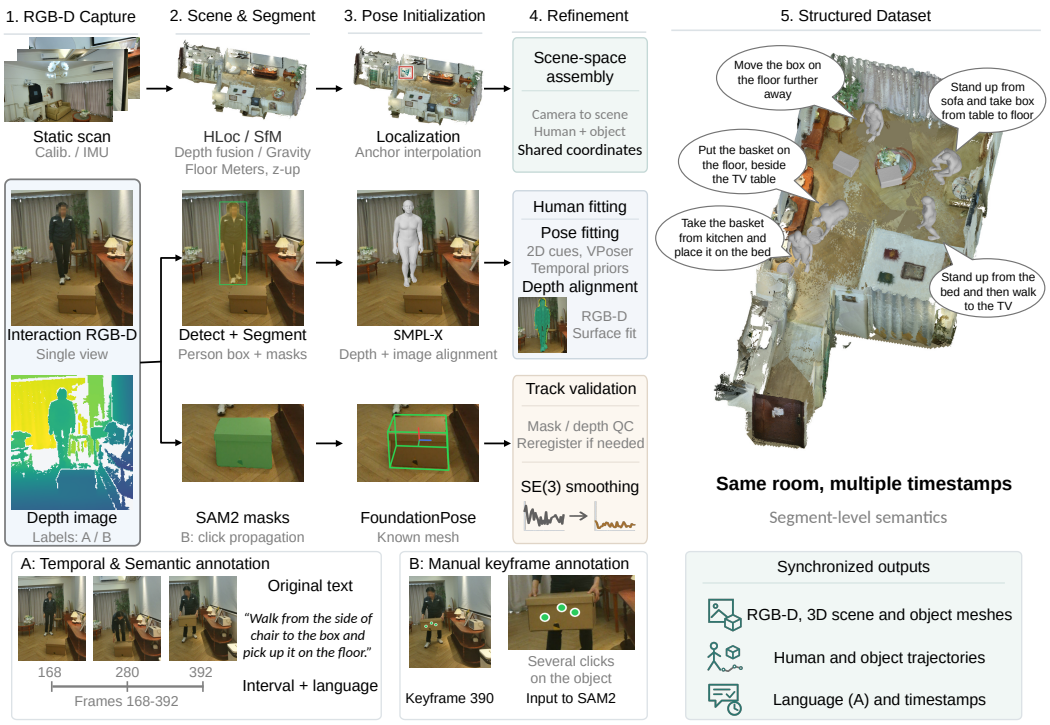


Figure 1: SituHOI construction pipeline. We collect room scans, RGB-D activity, object meshes, and manual annotations, which support camera, human, and object estimation in a shared scene space. Annotations specify activity intervals, language descriptions, and target-object clicks.

clips spanning approximately 86 hours, of which 10,335 clips covering 48.64 hours involve 13 dynamic objects, and the remaining clips capture interactions with static furnitures.

3.1 Static room scans and scene reconstruction

For each room, we capture a static RGB-D scan and retain the sensor calibration and available IMU measurements. We use HLoc (Sarlin et al., 2019) for image retrieval and feature matching across the scan, and COLMAP (Schönberger & Frahm, 2016) to recover the scan-camera poses and sparse scene structure from these correspondences. Since COLMAP cannot recover the correct metric scale, we estimate the metric scale from the median of the ratio between captured depth and COLMAP-recovered depth, then fuse the backprojected observations into a dense mesh. We estimate the vertical direction from camera orientations and reliable IMU data, fit the floor plane to 3D scene points, and align it with $z = 0$, with the vertical direction pointing along $+z$. The resulting metric, gravity-aligned reconstruction provides both room geometry and a reference map for localizing the interaction camera. Further details are given in Appendix B.1.

3.2 Dedicated object scans and mesh reconstruction

For each dynamic object, we acquire a dedicated RGB-D scan and reconstruct a textured mesh using BundleSDF (Wen et al., 2023). This scan is collected separately from the interaction recordings so that the object's geometry can be reconstructed before estimating its motion. The resulting mesh serves as known geometry for rigid-pose registration and tracking in the interaction recordings.

3.3 Interaction recordings and trajectory reconstruction

Recording and annotation. We record dynamic activities across 16 rooms using a single hand-held RGB-D camera at 30 FPS, with RGB resolution of 1280×960 and depth resolution of 640×480. We collect approximately 90 hours of raw recordings. After excluding segments with severe person-object occlusion, motion blur, or missing or corrupted depth, and omitting gaps between annotated

activities, the retained clips cover approximately 86 hours. This duration measures annotated clip coverage rather than continuous raw recording time.

Annotators mark clip boundaries, write free-form English descriptions of actions and spatial relations, and identify the involved objects and their types. For each dynamic object interaction clip, they choose a keyframe and provide one or more point clicks to initialize mask propagation. Fig. 1 shows an example in which a language description covers frames 168-392 and three clicks on the box at frame 390 initialize its mask.

Camera localization. We localize the interaction camera against the metric scene map constructed from the static room scan. By default, for every 30th recording frame, HLoc retrieves relevant scan images, matches visual features, and associates 2D image pixels with 3D map points. Its localization pipeline then estimates the camera pose using RANSAC-based PnP followed by refinement. The successfully localized frames serve as camera-pose anchors. Camera poses between retained anchors are obtained by linear interpolation of translation and spherical linear interpolation (SLERP) of rotation; poses outside the anchor span use the nearest retained endpoint. The resulting camera trajectory places the human and object motion reconstructed below in the same coordinate system.

Human motion. OSX (Lin et al., 2023) provides initial SMPL-X pose and shape from RGB. We get root translation by back-projecting person-region depth and transform the body to the scene space using the camera pose, then apply two-stage fitting. The first stage refines this estimate using confidence-weighted 2D keypoints, temporal regularization, and the VPoser body-pose prior (Pavlakos et al., 2019). The second stage further aligns the body surface with 3D points back-projected from depth pixels within SAM 2 person masks (Ravi et al., 2025).

Let J_t and V_t denote the SMPL-X joints and vertices at frame t, and let R_t, q_t, and z_t denote the optimized joint rotations, root translation, and VPoser pose code. The two stages minimize

$$E_h^{(1)} = \lambda_{2D} E_{2D} + E_{\text{prior}} + E_{\text{preserve}} + \sum_{x \in \{J,V,R,q,z\}} \lambda_x E_{\Delta^2 x_t}, \qquad E_h^{(2)} = E_h^{(1)} + \lambda_D E_D. \quad (1)$$

Here, $\Delta^2 x_t = x_{t+1} - 2x_t + x_{t-1}$ is the second temporal difference. E_{2D} is a robust reprojection loss between projected SMPL-X joints and confidence-weighted RTMW detections (Jiang et al., 2024), using 63 matched body, foot, and hand landmarks. E_D is a robust bidirectional Chamfer distance between masked RGB-D person points and sampled body vertices. E_{prior} regularizes body and hand pose, while E_{preserve} discourages deviations from the OSX initialization. For rotational variables, temporal differences are computed on rotation-matrix entries. Appendix B.2 gives the complete loss definitions and optimization settings.

Object motion. We propagate object masks from the annotated keyframe clicks using SAM 2 (Ravi et al., 2025) and estimate the pose of each object using its dedicated-scan mesh and FoundationPose (Wen et al., 2024). Our tracking safeguards include registering candidate poses on visible frames around the annotated keyframe and selecting an initialization based on mask overlap, depth agreement, and whether the object is truncated by the image boundary. FoundationPose then tracks the object forward and backward from the selected frame.

During tracking, we run a safeguard process to improve tracking robustness. Low overlap between the projected mesh and propagated mask triggers a new registration proposal. We accept the proposal only if it improves mask agreement and passes the depth, visibility, and trajectory-continuity checks in Appendix B.3. The same recovery procedure applies when an object reappears after an occlusion. Rejected poses within visible spans are interpolated when bounded by reliable estimates. During invisible spans, we retain the last pose as a placeholder and mark those frames as unobserved.

We apply object refinement to visible trajectories to reduce temporal jitter while keeping them close to the tracked poses. Let p_t^0 and Q_t^0 denote the tracked scene-frame translation and rotation. We optimize bounded corrections δp_t and δr_t such that

$$p_t = p_t^0 + \delta p_t, \qquad Q_t = \text{Exp}(\delta r_t) Q_t^0, \quad (2)$$

where Exp maps an axis-angle correction to a rotation matrix. Defining

$$\omega_t = \text{Log}(Q_{t+1} Q_t^\top), \qquad \Delta\omega_t = \omega_{t+1} - \omega_t,$$

we minimize

$$E_o = \lambda_p \,\text{MSE}(\Delta^2 p) + \lambda_\omega \,\text{MSE}(\Delta\omega) + \lambda_{p0} \,\text{MSE}(\delta p / s_p) + \lambda_{r0} \,\text{MSE}(\delta r / s_r). \quad (3)$$

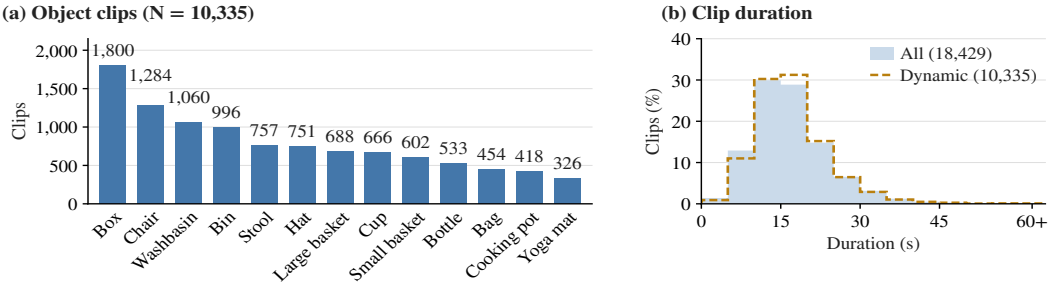


Figure 2: Dataset distributions: (a) object counts; (b) clip durations, normalized per group.

The first two terms reduce translational and rotational jitter; the last two penalize deviations from the tracked trajectory. We bound corrections by $s_p = 0.05$ m and $s_r = 0.25$ rad, which also normalize the fidelity terms. Corrections are fixed to zero on invisible frames, leaving their held placeholder poses unchanged. Appendix B.3 provides the remaining optimization details.

3.4 Quality Control

We first screen the recordings for severe occlusion, motion blur, and missing or corrupted depth, as described above. After reconstruction, we combine manual inspection with automated checks of source-data integrity, camera-pose coverage, trajectory continuity, and the agreement of human and object geometry with RGB-D masks and depth. Appendix B.3 gives the checks and acceptance thresholds. We further assess trajectory quality through component ablations and independent synchronized MoCap validation in Section 5.1.

3.5 Corpus Coverage and Comparison

Comparison with existing resources. Tab. 1 highlights SituHOI's combination of single-camera RGB-D capture, full-body and dynamic-object trajectories, room meshes, and language without body- or object-mounted trackers. Their common scene frame links explicit surfaces and motion for geometric analysis and scene-conditioned modeling (criteria and sources: Appendix A.3).

Object and clip distributions. Fig. 2 shows object frequencies and clip durations. The 13 types span furniture, containers, and smaller household objects; boxes and chairs dominate across varied appearances and manipulations. Longer clips support history/forecast resampling beyond current benchmark windows; Appendix A.1 details the inventory, durations, and language statistics.

4 FlexHOI-Bench

Existing HOI prediction and synthesis protocols provide different amounts and types of future information, making their results difficult to compare under a common task (Xu et al., 2023; Li et al., 2023). To provide a consistent setting, we introduce **FlexHOI-Bench** on SituHOI. Given observed human-object motion, scene and object geometry, and a target object pose, the benchmark tests whether a model can predict their coordinated motion toward that target across varying observation and forecast durations. We evaluate both world-space trajectory accuracy and human-object coordination, and propose **SharedHOI** as a reference method.

4.1 Benchmark Definition and Protocol

We formalize each interaction as a sequence of human and object states. At time t, the state is $X_t = (q_t, \Theta_t, p_t, R_t)$, where q_t is the human-root translation, Θ_t contains the scene-frame root orientation and 21 SMPL-X body rotations, and (p_t, R_t) is the object translation and rotation.

Let $t = 0$ denote the last observed time. For a history duration H and forecast duration F, let $\mathcal{Q}_H \subset [-H, 0]$ and $\mathcal{Q}_F \subset (0, F]$ denote the sampled history and future timestamps. The task is

$$\widehat{X}_{\mathcal{Q}_F} = f_\theta(X_{\mathcal{Q}_H}, \mathcal{S}, \mathcal{O}, g, \mathcal{Q}_H, \mathcal{Q}_F), \qquad g = (p_F, R_F), \tag{4}$$

where $\mathcal{S}$ and $\mathcal{O}$ denote the static scene and object geometry, and g is the terminal object pose obtained from the reconstructed trajectory. At inference time, g is provided as a condition. The

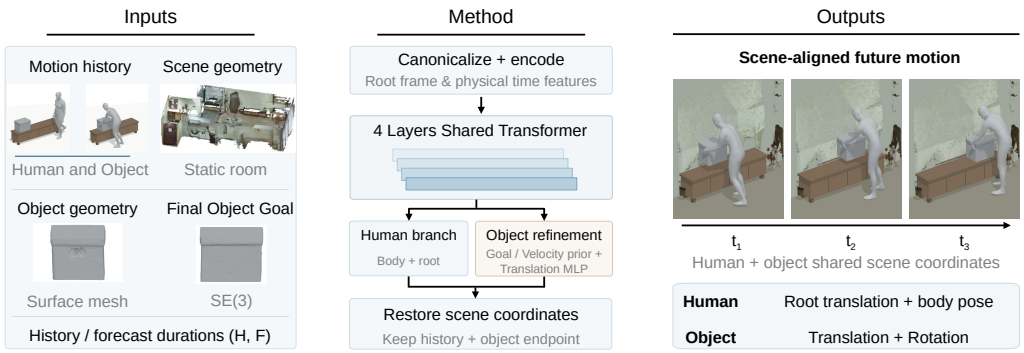


Figure 3: SharedHOI predicts joint motion from history, scene/object geometry, the object goal, and query times. An endpoint-preserving head refines object translation before both trajectories are restored to the metric scene frame.

model predicts the complete future human motion and the intermediate object trajectory, with the object pose at $t = F$ constrained to match g.

Data and partition. We extract 135-frame source windows from quality-screened dynamic-object clips with paired human and object trajectories and known object meshes. Each window spans approximately 4.5 s at 30 Hz. This yields 26,404 training, 5,527 validation, and 6,948 testing windows. We partition the data by recording so that overlapping windows from the same recording never cross splits, and compute all normalization statistics using training data only. Appendix D details the window-extraction and trajectory-validity requirements.

Flexible-time protocol. We train a single model across variable history and forecast durations. For each source window, we sample the last-observation boundary $t = 0$ together with continuous durations (H, F) satisfying $H + F \leq 4$ s; the remaining portion of the 4.5-second window allows the boundary to shift across training examples. Motion states are sampled at the requested timestamps using linear interpolation for translations and spherical linear interpolation for rotations. At evaluation, the same trained model is applied without retraining to 11 fixed duration pairs at 16 Hz. Training excludes small neighborhoods around $(H, F) = (0.85, 1.6)$ s and $(1.5, 0.8)$ s, so these pairs test generalization to held-out duration combinations. Checkpoint selection uses six validation pairs outside the excluded neighborhoods. Appendix D specifies the query counts, excluded ranges, and validation grid; Section 5.2 reports the evaluation protocol and results.

4.2 SharedHOI: shared prediction with object refinement

We further propose SharedHOI as reference model for FlexHOI-Bench. It couples human and object predictions through shared temporal features, then refines object translation while preserving the supplied terminal pose (Fig. 3).

To represent motion and geometry consistently, we express the observed states, target pose, and scene geometry in a local frame centered at the last observed human root and aligned with its yaw. This removes arbitrary global translation and heading while preserving their relative geometry. Separate point encoders aggregate 256 samples from the object mesh in its local frame and the 512 scene points nearest the last observed root.

A four-layer, width-192 Transformer processes observed-state and future-query tokens with continuous-time and history/future/goal role encodings. It also receives the last observed state, terminal object pose, history and forecast durations, object geometry, and local scene geometry. Its shared temporal features predict human and object residuals relative to simple reference motions: the human state is held at its last observed value, while the object pose is interpolated toward the terminal goal. We represent body rotations in continuous 6D form (Zhou et al., 2019).

We train the shared predictor for 120 epochs with weighted state-reconstruction and physical-time velocity losses. To supervise human-object coordination, we also predict the distances from both wrists to the object surface, computing wrist positions through SMPL-X forward kinematics with a neutral body shape. We then freeze the shared predictor and train only the object-translation

Table 2: Accuracy on the 30-minute MoCap subset. OSX (Lin et al., 2023) and FoundationPose (Wen et al., 2024) are the initial human and object estimates. Each "w/o" row removes a component from the full pipeline: E_D is the depth-based human surface loss, and $E_{\Delta^2 x_t}$ denotes all human temporal terms in Eq. 1; object refinement is defined in Eq. 3, and tracking safeguards are described under Object motion in Sec. 3.3. All configurations use the same calibration and valid samples for each applicable metric. Dashes denote inapplicable or unreported metrics.

Pipeline configuration	**G-MPJPE**/mm↓	**RA-MPJPE**/mm↓	**Obj. trans.**/mm↓	**Obj. rot.**/° ↓	**HOI-relative**/mm↓
OSX	195.35	168.84	–	–	–
FoundationPose	–	–	43.0	11.6	–
w/o E_D	73.6	50.0	28.7	7.5	55.2
w/o $E_{\Delta^2 x_t}$	79.3	55.4	28.7	7.5	61.5
w/o object refinement	68.3	47.6	32.2	8.9	54.8
w/o tracking safeguards	68.3	47.6	48.7	13.2	69.6
Full pipeline	**68.3**	**47.6**	**28.7**	**7.5**	**51.6**

refinement head for up to 40 additional epochs, preserving the predicted human motion. Architecture and loss details are provided in Appendix D.2.

5 Experiments

We evaluate three questions: (1) how accurately the pipeline reconstructs scene-aligned human-object motion and which components contribute to its performance; (2) whether a single model can jointly predict human and object motion across different observation and forecast durations; and (3) whether SituHOI provides useful supervision beyond FlexHOI-Bench through cross-dataset transfer and aligned language and scene context.

5.1 Reconstruction Quality and Component Ablations

We assess 3D accuracy on approximately 30 minutes of synchronized RGB-D and motion-capture (MoCap) interactions from two participants. Fixed space-time calibration aligns matched joints and object-mesh origins. Global mean per-joint position error (G-MPJPE) uses scene coordinates without per-frame alignment, whereas root-aligned MPJPE (RA-MPJPE) removes only root translation. We also report object translation and rotation errors and object-local wrist error (HOI-relative). All configurations use the same valid samples for each applicable metric (Appendix C.1).

OSX (Lin et al., 2023) and FoundationPose (Wen et al., 2024) provide the initial human and object estimates in our pipeline (Sec. 3). These estimates have substantial 3D errors (Tab. 2): OSX yields 195.35 mm G-MPJPE and 168.84 mm RA-MPJPE, while FoundationPose yields 43.0 mm object translation error and 11.6° rotation error. The full pipeline reduces the corresponding errors to 68.3 mm, 47.6 mm, 28.7 mm, and 7.5°.

Human pose fitting contributes substantially to this improvement. Removing the temporal terms $E_{\Delta^2 x_t}$ increases G-MPJPE and RA-MPJPE by 11.0 and 7.8 mm, respectively; removing the depth-based surface loss E_D increases them by 5.3 and 2.4 mm. The two ablations also increase HOI-relative error from 51.6 mm to 61.5 and 55.2 mm, respectively, indicating that human fitting improves both joint accuracy and human-object alignment. Object trajectory refinement reduces translation and rotation errors from 32.2 mm and 8.9° to 28.7 mm and 7.5°. Object tracking safeguards have a larger effect: they reduce these errors from 48.7 mm and 13.2° to 28.7 mm and 7.5°, and reduce HOI-relative error from 69.6 to 51.6 mm. Fig. 4 illustrates the resulting improvement in tracking continuity and mask alignment.

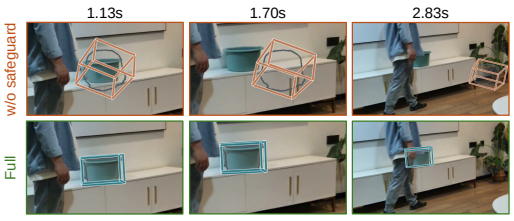


Figure 4: Tracking over 1–3 s without (top) and with (bottom) safeguards.

Since different components in our data pipeline affect different aspects of tracking quality, we provide more extensive ablations of these components and their specific effects in Appendix C.2.

5.2 Evaluation on FlexHOI-Bench

We compare SharedHOI with four task-adapted baselines, including CHOIS (Li et al., 2024), DecHOI (Jung et al., 2026), InterDiff (Xu et al., 2023), and OMOMO (Li et al., 2023), on FlexHOI-Bench. All methods share the same partition and state/geometry/goal/duration interface and are trained for 120 epochs. SharedHOI adds up to 40 refinement epochs (details in Appendix D.2).

Table 3: FlexHOI-Bench results: 6,948 testing windows, equal-weight averages over 11 pairs, and representative short, training-excluded, and long forecasts.

Method	All 11 pairs			$H = 3.0$ s, $F = 0.5$ s			$H = 0.85$ s, $F = 1.6$ s			$H = 0.25$ s, $F = 3.75$ s		
	Root/mm↓	Obj./mm↓	F1(%)↑	Root/mm↓	Obj./mm↓	F1(%)↑	Root/mm↓	Obj./mm↓	F1(%)↑	Root/mm↓	Obj./mm↓	F1(%)↑
CHOIS	244.7±5.4	65.8±0.04	68.2±0.6	126.4±1.9	21.8±0.06	80.9±0.9	248.0±3.4	69.9±0.05	65.2±0.2	457.7±9.6	143.5±0.02	48.6±0.6
DecHOI	190.8±0.5	65.8±0.02	72.1±0.2	98.5±0.2	21.8±0.00	83.1±0.5	188.6±0.8	69.7±0.02	70.0±0.2	362.2±1.3	143.6±0.01	54.4±0.6
InterDiff	265.8±0.6	83.5±0.33	66.7±0.3	137.2±4.2	42.8±2.21	79.8±0.6	263.3±1.8	87.8±0.30	63.8±0.9	507.6±2.4	158.8±0.56	45.7±0.1
OMOMO	322.3±4.8	65.4±0.00	60.9±0.6	149.1±1.4	21.4±0.00	78.0±0.9	334.4±6.1	69.3±0.00	55.6±1.1	628.5±25.1	143.1±0.00	37.7±0.7
SharedHOI	**187.2±3.3**	**61.0±0.22**	**74.9±0.4**	**96.3±3.0**	**19.7±0.02**	**85.3±0.4**	**186.9±3.2**	**64.8±0.22**	**73.1±0.4**	**355.0±4.1**	**132.9±0.66**	**58.1±0.7**

Table 4: Transfer to BEHAVE under matched 6k- and 12k-step training budgets. For each budget, the control trains entirely on BEHAVE, while the transfer variant trains for the first 3k steps on SituHOI and the remaining steps on BEHAVE. All models are evaluated on BEHAVE. InterDiff reports eight DDIM sampling seeds and HOIMotion three runs; values are means ± sample SD. Bold denotes the better mean within each comparison.

		6k total steps		12k total steps	
Method	Metric (mm↓)	BEHAVE only	+SituHOI pretrain	BEHAVE only	+SituHOI pretrain
InterDiff	Object ADE	157.39±4.76	**141.32±4.25**	155.24±4.32	**139.36±4.68**
HOIMotion	Body MPJPE	67.93±0.34	**56.59±0.47**	67.73±0.18	**55.43±0.49**

Evaluation metrics. We compare methods across history and forecast duration pairs using three metrics. Human-root and object translation errors measure trajectory accuracy against ground truth. Proximity F1 measures whether the predicted wrists are near the object at the correct interactoin frames: for each wrist, we classify its distance to the object surface as near when it is below 20 cm, then compare these states with the ground truth. More details can be found in Appendix D.3.

Benchmark comparison and duration generalization. Tab. 3 shows that SharedHOI leads on all three metrics in the aggregate and in every duration setting shown. Averaged equally over the 11 pairs, it reduces root error by 1.9% relative to DecHOI (187.2 vs. 190.8 mm) and object error by 6.7% relative to the best baseline (61.0 vs. 65.4 mm), while improving proximity F1 by 2.8 percentage points (74.9% vs. 72.1%). The advantage persists for the long forecast ($F = 3.75$ s), where SharedHOI achieves 355.0 mm root error, 132.9 mm object error, and 58.1% proximity F1. These results indicate that SharedHOI improves both trajectory accuracy and human-object coordination across the evaluated forecast horizons.

5.3 Cross-dataset transfer

We test whether SituHOI benefits existing HOI methods on BEHAVE (Bhatnagar et al., 2022). For InterDiff (Xu et al., 2023) and HOIMotion (Hu et al., 2024), we compare training entirely on BEHAVE with training for the first 3k steps on SituHOI and the remaining steps on BEHAVE, holding the total budget at either 6k or 12k steps. All models are evaluated on BEHAVE. As shown in Tab. 4, pretraining reduces InterDiff object ADE from 157.39 to 141.32 mm at 6k steps (10.21%) and from 155.24 to 139.36 mm at 12k steps (10.23%). It also reduces HOIMotion body MPJPE from 67.93 to 56.59 mm (16.69%) and from 67.73 to 55.43 mm (18.16%), respectively. These gains across both methods and training budgets show that SituHOI provides useful pretraining data for HOI modeling beyond FlexHOI-Bench.

6 Conclusion

We present SituHOI, a scene-grounded HOI dataset with 86 hours of recordings, 18,429 annotated clips, and 16 indoor rooms. Captured with a single RGB-D camera, it aligns full-body and rigid-object trajectories with scene geometry and segment-level language. FlexHOI-Bench evaluates goal-conditioned joint prediction across observation and forecast durations, with SharedHOI as a reference model. MoCap validation supports reconstruction quality, while benchmark comparisons and external-model evaluations demonstrate its value for human-object motion modeling.

Limitations. The data covers indoor interactions with rigid objects, while the benchmark provides the terminal object pose and limits $H + F$ to 4 s for manageable training. Future work could extend coverage to articulated and deformable objects and longer interactions, infer goals from observations, and evaluate contact, support, and collisions more explicitly.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking, LaTeX formatting, and format checks. AI-assisted compiled manuscript was inspected for formatting and layout issues. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

SituHOI was collected with the consent of all participants for recording, public data release, and the publication of identifiable images in the paper and dataset. Third-party assets remain subject to their respective licenses. The resource supports research in human–object interaction, scene-aware motion modeling, and multimodal learning. As RGB-D observations and reconstructed indoor scenes can retain identifying information, downstream users should respect the agreed scope of consent and avoid re-identification or non-consensual surveillance. Responsible reuse should also consider the coverage of the collected participants and environments when deploying models beyond the dataset.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe dataset construction, FlexHOI-Bench, and the experimental protocols. Appendices A and B document data formats, coordinate conventions, reconstruction objectives, and implementation settings. Appendix C specifies ablation cohorts, metrics, and aggregation. Appendix D provides window construction, recording-level partitions, temporal queries, model architectures, training settings, random seeds, and baseline adaptations. External-model protocols are detailed in Appendix E; synchronized MoCap calibration and accuracy metrics are described in Section 5.1. An anonymized code repository is available at `https://anonymous.4open.science/r/iclr2027-anonymous-code-1DFB`. The full dataset and code will be publicly released upon acceptance.

REFERENCES

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, 2016.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7071, 2025. URL `https://facebookresearch.github.io/hot3d/`.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Yukang Cao, Haozhe Xie, Beichen Wen, Runmao Yao, Yinghao Liu, Yue Huang, Zhichao Liao, Yunxiang Wang, Haiheng Liu, Xingshun Tian, Dawei Su, Long Zhuo, Dacheng Tao, Xiaogang Wang, Liang Pan, and Ziwei Liu. ACE-Data-0: Human-centric ambient capture as embodied data engine. *arXiv preprint arXiv:2607.28625*, 2026. URL `https://arxiv.org/abs/2607.28625`.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

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*, pp. 19888–19901, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arctic.is.tue.mpg.de/.

Qihang Fang, Yingda Yin, Qingnan Fan, Fei Xia, Siyan Dong, Sheng Wang, Jue Wang, Leonidas J. Guibas, and Baoquan Chen. Towards accurate active camera localization. In *Proceedings of the European Conference on Computer Vision*, pp. 122–139, 2022. doi: 10.1007/978-3-031-20080-9_8.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2282–2292, 2019.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1516–1540, 2025. URL https://proceedings.mlr.press/v270/he25b.html.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.11709.

Zhiming Hu, Zheming Yin, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. HOIMotion: Forecasting human motion during human-object interactions using egocentric 3D object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics*, 30(11):7375–7385, 2024. doi: 10.1109/TVCG.2024.3456161. URL https://zhiminghu.net/hu24_hoimotion.html.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

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

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *Pattern Recognition: 44th DAGM German Conference*, volume 13485 of *Lecture Notes in Computer Science*, pp. 281–299, 2022b. URL https://intercap.is.tue.mpg.de/.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision*, 2024. doi: 10.1007/s11263-024-01984-1. URL https://intercap.is.tue.mpg.de/.

Jiahao Ji, Ji Ma, Runhan Zhang, Runyi Yu, Wenjia Wang, Weiheng Chi, Qianqian Peng, Weichao Yan, Yongfei Gu, Ye Tian, Ting Wu, Longwei Li, Chun Yuan, Ruoli Dai, and Lei Han. HiPHI: A large-scale benchmark for high-precision human motion and object-interaction. *arXiv preprint arXiv:2608.16222*, 2026. URL https://arxiv.org/abs/2608.16222.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Tao Jiang, Xinchen Xie, and Yining Li. RTMW: Real-time multi-person 2D and 3D whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024.

Hwanhee Jung, Seunggwan Lee, Jeongyoon Yoon, SeungHyeon Kim, Giljoo Nam, Qixing Huang, and Sangpil Kim. Decoupled generative modeling for human-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2253–2263, 2026. URL https://kuai-lab.github.io/cvpr2026dechoi/.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing everyday home activities towards 3D generative modeling of human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://jlogkim.github.io/parahome/.

Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. MMHOI: Modeling complex 3D multi-human multi-object interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1512–1521, 2026. URL https://jk-vision.github.io/mmhoi/.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6), 2023. URL https://lijiaman.github.io/projects/omomo/.

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*, 2024. URL https://lijiaman.github.io/projects/chois/.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21159–21168, 2023.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. CORE4D: A 4D human-object-human interaction dataset for collaborative object REarrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1769–1782, 2025. URL https://core4d.github.io/.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D dataset of mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10886–10897, 2025.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://lvxintao.github.io/himo/.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. 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*, pp. 10975–10985, 2019. URL https://smpl-x.is.tue.mpg.de/.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9488–9497, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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*, 2020. URL https://grab.is.tue.mpg.de/.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 606–617, 2023.

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*, pp. 17868–17879, 2024.

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*, pp. 14928–14940, 2023.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. URL https://sirui-xu.github.io/InterAct/.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8834–8845, 2023.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M^3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://juzezhang.github.io/HOIM3_ProjectPage/.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, pp. 180–200, 2022. URL https://sanweiliti.github.io/egobody/egobody.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://afterjourney00.github.io/IM-HOI.github.io/.

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*, pp. 5745–5753, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html.

APPENDIX

A DATASET DOCUMENTATION

A.1 INVENTORY AND COUNTING CONVENTIONS

The corpus covers 16 distinct indoor rooms. A *recording* is a continuous RGB-D capture; an *annotated clip* (or *segment*) is an annotated activity interval, and a processed *sequence* stores its fitted trajectories. *Evaluation clips* are intervals selected for ablation; benchmark *windows* are resampled sequence excerpts. A *room group* (capture family) groups recordings by physical room. Excluding segments with recording blur, severe occlusion, or unusable visual evidence leaves approximately 86 hours of annotated activity. Acquisition screening assesses recording usability; geometric checks and independent MoCap validation assess trajectory consistency and 3D accuracy. Table 5 separates all semantic activity from the dynamic-object subset; the remainder includes human interactions with static furniture, not merely idle footage. Each room has its own metric coordinate frame. Repeated scans, dates, and alternate scene names do not increase the room count or establish participant identities. SituHOI was recorded with 15 distinct participants. Each participant contributed the same total recording duration.

Table 5: Corpus inventory. Hours count unions of annotation spans within recordings; the dynamic-object subset is contained in the full corpus.

Quantity	Full semantic corpus	Dynamic-object subset
Distinct indoor rooms	16	16
Semantic intervals	18,429	10,335
Annotation-covered time (h)	85.98	48.64
Dynamic object types	—	13
Median interval span (s)	15.80	16.03
Interval-span interquartile range (s)	12.07–20.43	12.47–20.35
Median description length (words)	18	19
Description-length interquartile range	14–22	15–23

A.2 DATA CONTRACT AND COORDINATES

Interaction recordings use a single handheld RGB-D camera at 30 FPS, with 1280×960 RGB and 640×480 depth. We retain sensor calibration and available inertial measurement unit (IMU) readings with the room scans. In Figure 1, a language description covers frames 168–392 and three clicks on the box at frame 390 initialize its mask. Clicks specify tracking targets, not 3D correspondences or object poses; clip and source-frame IDs link both annotation types to the observations.

Measured RGB/depth and calibration accompany human-annotated intervals, English descriptions, object identities, and keyframe clicks. Estimated fields include masks, 63 confidence-bearing keypoints, camera-to-scene transforms, SMPL-X motion (translation, 165 axis-angle pose entries, ten shape coefficients), rigid object-to-scene transforms, visibility, and QC/provenance. Lengths follow stored source frame IDs rather than a fixed temporal grid. The 165 pose entries include root, body, jaw/eyes, and hands; local joint rotations are not all world rotations.

For homogeneous points, $X^s = T^s_{c,t} X^c$ and $X^s = T^s_{o,t} X^o$ give camera/object-to-scene mappings. Object projection is $u = \pi(K(T^s_{c,t})^{-1} T^s_{o,t} X^o)$, where K denotes camera intrinsics and π perspective division; scene-aligned human vertices need only the inverse camera transform. Estimator outputs are restored to raw image orientation before projection. Static-scan and dynamic-camera calibration remain distinct, and resizing a depth display does not create new calibrated measurements. Body-model versions and parameter conventions accompany exports.

Object mesh acquisition. We reconstruct the meshes for the 13 object types in Figure 5 from dedicated RGB-D scans, separate from the interaction recordings, using BundleSDF (Wen et al., 2023). Calibrated color images, depth maps, and object foreground masks support multi-view pose tracking and optimization of a neural signed distance field (SDF), whose zero level set defines the object surface. Global refinement jointly optimizes this field and the selected observation poses. We extract a triangle mesh, map the captured RGB images onto its surface as texture, and export it in metric object coordinates. These instance-specific assets provide the known geometry for FoundationPose's six-degree-of-freedom (6-DoF) tracking in the interaction recordings.



Figure 5: SituHOI's 13 dynamic object types, independently scaled.

A.3 Comparison-table definitions and sources

Table 1 compares resources, not pose accuracy. ✓ means documented support; ✗ means not satisfied or not established, not uniformly verified absence. NR is unreported, not zero. Full-body 3D covers torso, arms, and legs, without requiring fingers or a specific mesh model. Object pose means dynamic-object trajectories, not static scene assets. Scene geom. includes layouts, scanned/artist-built assets, or partial point clouds. Scene mesh requires explicit surface meshes for the environment, including aligned static-furniture meshes; isolated manipulated-object meshes alone do not qualify. This distinguishes a provided surface representation from point-only scene geometry, without imposing a watertightness requirement. Semantics includes prompts, intents, action/sequence labels, and descriptions of unequal granularity. Sequence counts retain their source units; multi-view images are not multiplied into capture hours. Envs. counts physical capture spaces where established; HOI-M^3's entry of five denotes scene types, not a count of physical rooms. SituHOI has 16 distinct indoor rooms.

Acquisition and release scope. Uninstr. excludes body/object tracking markers, IMUs, and mocap suits/gloves, but not capture cameras, headset localization sensors, or calibration-only targets. MMHOI's AprilTags calibrate cameras rather than track the interaction objects (Kogashi et al., 2026). PROX denotes its 60-sequence, 12-room qualitative release, with scanned scene meshes, not the separate Vicon-referenced quantitative subset (Hassan et al., 2019). RICH has five mostly outdoor scanned scene meshes; its semantic check refers to action-bearing sequence names, not interval captions (Huang et al., 2022a). Neither provides paired dynamic-object trajectories.

We retain the reported capture modes of GRAB and BEHAVE (Taheri et al., 2020; Bhatnagar et al., 2022), InterCap and CHAIRS (Huang et al., 2024; Jiang et al., 2023), and OMOMO and IMHD^2 (Li et al., 2023; Zhao et al., 2024). InterCap uses the expanded 223-sequence journal version; CHAIRS uses the final ICCV 17.3-hour release rather than its earlier preprint. Sparse annotation instants are not converted to unsupported hours. HOI-M^3 uses original CVPR counts, HIMO distinguishes whole-sequence and segmented language, and ParaHome has a common room setup. HOI-M^3 and ParaHome provide scanned, placed furniture assets; HUMOTO provides artist-built scene assets. These qualify as scene meshes under the same surface-representation criterion, without relabeling artist-built assets as measured scans (Zhang et al., 2024; Lv et al., 2024; Kim et al., 2025; Lu et al., 2025).

ACE-Data-0 reports two capture environments and tracked object meshes; an explicit background-scene surface mesh is not established in its documented release. HiPHI's 245.7 hours are a mirrored HOI subset, not unaugmented capture time (Cao et al., 2026; Ji et al., 2026). InterAct/-X consolidates seven resources: 11,350/16,201 base/augmented clips and 21.81/30.70 hours are not independent new acquisitions; mixed instrumentation explains its cross (Xu et al., 2025a). HOI4D spans 610 rooms with hand-focused supervision and reconstructs static scene point clouds after masking moving content; HOT3D's scene clouds refer to Aria recordings in one laboratory. Both receive Scene geom. checks; their documented point clouds do not establish Scene mesh support. SituHOI supplies

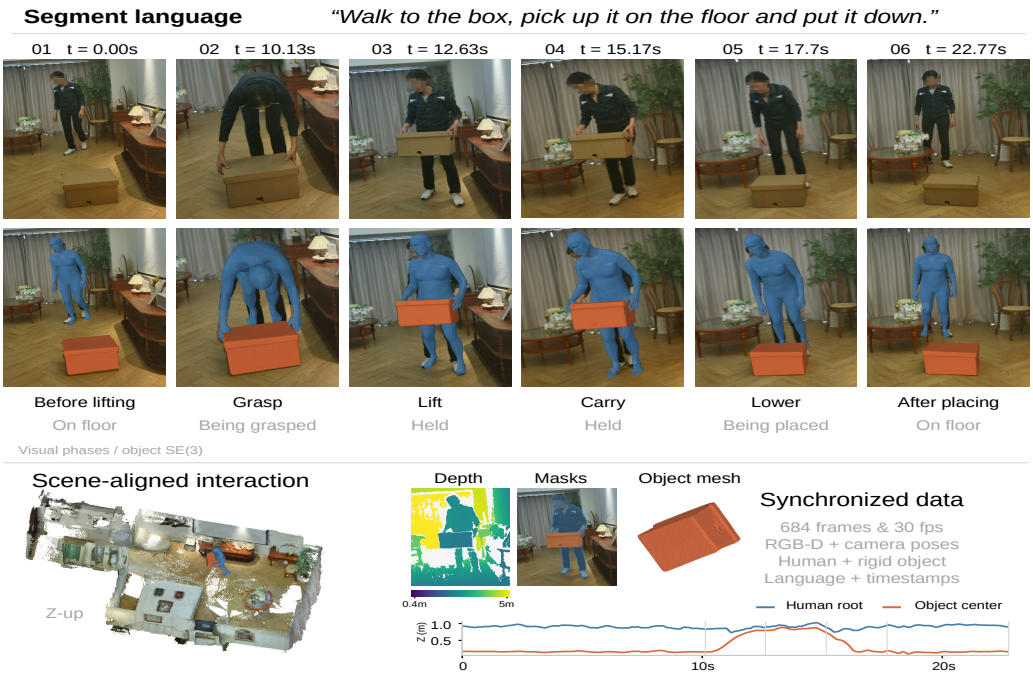


Figure 6: One continuous 684-frame interaction, with six real moments spanning 22.77 s. Saved SMPL-X/object estimates, original interval text, scene, depth, masks, mesh, and height traces share the source timeline. Phase/state labels are visual interpretations, not extra annotations. Insets use $t = 12.63$ s; depth spans 0.4–5 m, with invalid pixels gray.

reconstructed room meshes aligned with human/object trajectories across 16 rooms. EgoDex uses the March 2026 ICLR-associated author version, with upper-body/hand rather than full-body/object trajectories (Liu et al., 2022; Banerjee et al., 2025; Hoque et al., 2026).

A.4 Language and continuous multimodal example

Normalization lowercases word/number tokens and joins them with single spaces, preserving internal apostrophes and decimal points but ignoring other punctuation. There are 12,272 distinct normalized descriptions in the full corpus and 7,856 in the dynamic subset. At a matched sample budget of 1,000 descriptions without replacement, the expected distinct-string counts are 944.18 for SituHOI and 102.03 for the official OMOMO text archive (4,912 records, 103 unique strings).

Figure 6 uses annotation 2, frames 610–1293 at 30 FPS; displayed frames are 610, 914, 989, 1065, 1141, and 1293, with modality insets at 989 and manual clicks at keyframe 1293. All 684 estimated frames contribute to the traces.

B Pipeline implementation details

These are current ordinary-path defaults; archived experiments retain their recorded configurations and source versions.

B.1 Scene reconstruction and alignment

For static SfM, HLoc retrieves up to 40 neighbors plus ten-image sequential overlap, using SuperPoint, SuperGlue, NetVLAD, and COLMAP (DeTone et al., 2018; Sarlin et al., 2020; Arandjelović et al., 2016; Sarlin et al., 2019; Schönberger & Frahm, 2016). We reject outlying measured-to-SfM depth ratios using interquartile-range filtering and take the median retained ratio as the global metric scale. Fusion defaults use 0.35–7.6 m depth, 0.012 m voxels, and at least three observations; discon-

tinuity filtering suppresses mixed-depth boundaries. Camera orientations and reliable IMU readings estimate the vertical direction. Random sample consensus (RANSAC) fits the floor plane; rotation and translation align the upward direction to $+z$ and the floor to $z = 0$. For interaction recordings, the default pipeline samples one query every 30 frames. HLoc retrieves up to 30 static images per query and matches SuperPoint features with SuperGlue to obtain 2D-3D correspondences. Its PyCOLMAP backend performs RANSAC-based PnP with a 12-pixel reprojection threshold, followed by pose refinement. Retrieved images are grouped by shared landmarks, and the pose with the most inliers is retained. After metric scene alignment and anchor filtering, translation/SLERP interpolation fills intervening frames; poses outside the anchor span use the nearest endpoint. Translation smoothing uses a nine-frame window. Exported frame IDs and calibration are retained.

B.2 Human fitting

A 32D VPoser pose code $z_t = z_t^0 + \delta z_t$ parameterizes the body, with initial code z_t^0 and optimized correction δz_t (Pavlakos et al., 2019). Root translation/orientation and hands receive corrections; shape, jaw, and eyes stay fixed. The fit maps 63 RTMW body, foot, and hand landmarks to SMPL-X, discarding confidence below 0.70 and weighting hand landmarks by 0.65.

Observation losses. For $\rho_\sigma(r) = \sigma^2 r^2/(r^2+\sigma^2)$, image diagonal d, projected landmarks $\hat{u}_{tj}$, detections u_{tj}, and confidence/group/positive-depth weights w_{tj}, the reprojection term is

$$E_{2D} = \frac{\sum_{t,j} w_{tj} \sum_{a\in\{x,y\}} \rho_{35/d}((\hat{u}_{tj,a} - u_{tj,a})/d)}{\max(2\sum_{t,j} w_{tj}, 1)}. \tag{5}$$

The penalty is componentwise, not applied to the 2D residual norm. Fixed person-mask/depth samples P_t and sampled body vertices $\widetilde{V}_t$ give

$$E_D = \mathrm{mean}_{t,p\in P_t}\, \rho_{.08}\left(\min_{v\in\widetilde{V}_t} \|p - v\|_2\right) + .2\,\mathrm{mean}_{t,v\in\widetilde{V}_t}\, \rho_{.08}\left(\min_{p\in P_t} \|v - p\|_2\right). \tag{6}$$

Distances are in meters; the reverse term is weaker because the scan is partial. Defaults sample supported frames at stride eight, 64 scan points and 128 vertices per frame, depth 0.3–5 m, and initial target distance at most 0.20 m. This is nearest-vertex agreement, not exact triangle-surface distance.

Temporal terms, priors, and schedule. Let b_t, h_t, o_t denote body, hand, and root-orientation axis angles, $\mathcal{Y} = \{z, b, o, h, q\}$ the preserved parameters, and y^0 their initial values. With $\langle\cdot\rangle$ averaging over frames and scalar entries, the remaining terms are

$$\begin{aligned} &E_{\mathrm{prior}} = \alpha_z\langle z^2\rangle + \alpha_h\langle h^2\rangle + \alpha_a\langle \exp(s_k b_{t,i_k})\rangle_{t,k}, \qquad E_{\mathrm{preserve}} = \sum_{y\in\mathcal{Y}} \mu_y \langle (y - y^0)^2\rangle, \\ &E_{\Delta^2 J_t} = \frac{\sum_{t,j} w_j \|\Delta^2 J_{t,j}\|_2^2}{3(T-2)\sum_j w_j}, \qquad E_{\Delta^2 x_t} = \langle(\Delta^2 x_t)^2\rangle, \quad x \in \{V, R, q, z\}. \end{aligned} \tag{7}$$

Here T is the frame count and w_j the joint weights. The elbow/knee prior uses zero-based body-pose indices $i = (52, 55, 9, 12)$ with signs $s = (1, -1, -1, -1)$; α, μ are the prior/preservation weights below. Joint second differences use selected outputs among the first 76 SMPL-X joints, with body/hand weights 1/1.35; the vertex term samples 192 vertices. Translation and latent terms use elementwise second differences; rotation uses rotation-matrix entry differences, excluding fixed jaw/eye slots. These losses differ from the body-22 evaluation metric. Temporal weights for joints, vertices, rotations, translation, and latent are 180, 20, .25, 50, and .20. Latent/angle/hand prior weights are .005/.005/.01; initialization-preservation weights for latent, body, root, hands, and translation are .40, .80, 1, 1, and 1. Image and scan weights are 320 and 30. Adam runs 70 iterations: 15 at .006 without scan loss, then 55 at .0025 with scan loss, with gradient norm clipped to eight and bounded corrections. Removing scan or reprojection loss does not remove depth or image information from initialization.

B.3 Object smoothing and tracking safeguards

Object updates are $Q_t = \mathrm{Exp}(a_t \delta r_t) Q_t^0$ and $p_t = p_t^0 + a_t \delta p_t$, where Q_t^0, p_t^0 are input poses and the binary update mask a_t is one on optimized frames and zero on inactive/held frames. Rotations are

left-composed in scene coordinates; angular smoothness uses differences of $\omega_t = \mathrm{Log}(Q_{t+1}Q_t^\top)$ rather than global axis-angle second differences. Adam uses 100 iterations at .012, translation/angular temporal weights 80/6, and preservation weights 1/1. Correction norms are bounded by $s_p = .05$ m/$s_r = .25$ rad; preservation normalizes by these scales. No human, contact, silhouette, depth, or mesh-fitting term enters this smoother.

Actual keyframe clicks and known meshes initialize SAM 2 masks and FoundationPose poses (Ravi et al., 2025; Wen et al., 2024). We register candidates on visible frames near the annotated keyframe and select the initialization by mask overlap, depth agreement, and image-boundary truncation, then track forward and backward. Low mask overlap and reappearance after occlusion trigger re-registration. Replacement requires improved mask agreement and satisfactory depth, visibility, reliable-coverage, and trajectory-continuity checks. Rejected visible poses are interpolated only between reliable estimates; invisible spans hold the last pose and are marked unobserved. The smoother leaves these held poses unchanged.

Quality-control checks. Integrity checks align arrays by source-frame ID and verify finite entries, homogeneous transform rows, rotation orthogonality, and unit determinants. Camera coverage uses the retained-anchor count, the fractions before the first and after the last anchor, and the span between them relative to the segment length. At most one anchor, an uncovered prefix/suffix above 50%, or a covered span below 25% raises a failure flag. Anchor residuals measure the Euclidean displacement between assembled and localized camera positions at anchor frames; a 90th-percentile residual above 0.75 m is flagged. Motion checks measure consecutive translation distances and rotation angles. Camera/human-root/object translation steps above 0.2/0.25/0.3 m trigger warnings and above 0.5/2.0/0.8 m trigger failure flags; camera/object rotation steps use 30/45-degree warning and 60/90-degree failure thresholds. Geometric diagnostics use minimum mesh height relative to the floor and nearest-surface distances to the room geometry above the floor; body/object vertices below -0.03 m trigger floor warnings. Object-projection checks measure rendered-mask overlap, the rendered silhouette fraction inside the mask, normalized center separation, and depth disagreement; their support requirements and frame-to-segment decision rule appear in Appendix C.1.

C Validation protocols and extended results

C.1 Metrics and statistical unit

Independent MoCap capture setting. To measure absolute 3D reconstruction accuracy, we collect approximately 30 minutes of synchronized RGB-D and MoCap interactions from two participants (Table 2). Validation takes place in original SituHOI capture environments and reuses the same physical object instances from dataset collection. This ties the reference measurements directly to the scene geometry and object shapes and scales encountered during dataset construction. We process the RGB-D recordings with the standard pipeline; MoCap trajectories serve as an independent evaluation reference. Fixed spatial calibration and time synchronization align the two trajectory sets, using matched body joints and the same object-mesh origins. All configurations share this calibration and the valid evaluation samples, enabling matched comparisons of pipeline components under the dataset's capture conditions. For Table 2, each metric is first averaged over valid frames within a sequence, then averaged equally across sequences.

G-MPJPE averages Euclidean position errors over the 22 SMPL-X body joints (indices 0–21, including the root), in the calibrated scene frame. RA-MPJPE uses the same joints after subtracting each trajectory's own root position, without rotation or scale alignment. Object translation error compares the common mesh origin. Object rotation error is the geodesic angle $\arccos((\mathrm{tr}(\hat{R}_t^\top R_t^{\mathrm{ref}}) - 1)/2)$ in degrees, with the argument clipped to $[-1, 1]$. For wrist $w \in \{L, R\}$ (joint indices 20/21), define its object-local position as $r_{t,w} = R_t^\top (J_{t,w} - p_t)$, where (p_t, R_t) is the object pose. HOI-relative averages $\|\hat{r}_{t,w} - r_{t,w}^{\mathrm{ref}}\|_2$ over both wrists and valid frames, using the estimated and reference object poses for their respective coordinate transforms. Position errors are reported in mm.

Table 6 uses 390 short clips and 130 long clips for human fitting, 140 stress clips for tracking, and 900 candidate decisions for the acceptance guard. The object audit contains 156 clips: 104 exceed both second-order thresholds below, while 52 fall at or below at least one. Relative smoothing reductions use the 104-clip group to avoid division by near-zero initial scores; strict geometric checks cover all 156 clips. Short human clips are hash-selected within room and low/mid/high initial-jitter

strata; long clips use room-wise hash selection, excluding prior clips. Tracking stress clips are flagged for weak reliable-update coverage, poor box overlap, or large pose jumps. Each comparison fixes inputs, initialization, and evaluation frames.

For 22 body joints (index j, coordinate c), the temporal score is $\mathrm{p95}_t \sqrt{\sum_{j,c}(\Delta^2 J_{tjc})^2/(22 \cdot 3)}$. Objects use $\mathrm{p95}_t \, \|\Delta^2 p_t\|_2$ and $\mathrm{p95}_t \, \|\omega_{t+1} - \omega_t\|_2$. Differences are not divided by physical Δt^2: units are mm/frame^2 or rad/frame^2, requiring matched recording rates. Human scan distance takes nearest-vertex distances on fixed depth points, then frame and clip medians, unlike Equation 6. Evaluation indices exclude the every-eighth-frame scan-loss samples. Silhouette and bounding-box IoU remain separate; paired object depth uses common supported pixels.

Strict geometric rule. At silhouette render scale .25, weak overlap requires box IoU $< .08$ and silhouette IoU $< .06$. Spatial inconsistency means silhouette-inside-mask fraction $< .30$ or center distance divided by mask-box diagonal $> .75$ (diagonal clamped to one pixel). Depth inconsistency requires at least 40 points and median absolute residual $> .18$ m. Severe separation requires inside fraction $< .10$ and normalized center distance $> .75$. A bad frame has severe separation, or spatial inconsistency together with weak overlap or depth inconsistency. Fewer than two supported frames is *insufficient*; otherwise at least two bad frames and a bad fraction $\geq .35$ constitute failure.

Aggregation. Continuous metrics in Tables 6 and 7 are cohort medians of clip-level values. For these metrics, paired benefits in Table 6 are medians of within-clip changes. Relative reduction is $\mathrm{median}_i[100(c_i - f_i)/c_i]$ for valid positive denominators, not a ratio of cohort medians. Ordinary-motion clips have initial translation and rotation temporal p95 above 10^{-6} m/frame^2 and 10^{-6} rad/frame^2; near-zero clips fall at or below either threshold. These measure second-order variation, not velocity. Repeated configurations do not enlarge the independent cohort.

C.2 Component ablations and additional controls

Configuration definitions. Table 6 compares the full pipeline with matched ablations or controls using the cohorts and metrics in Appendix C.1. Human temporal ablations remove all five $E_{\Delta^2 x_t}$ terms in Equation 1, or only the joint term $E_{\Delta^2 J_t}$; the long-clip row repeats the all-term ablation. Depth-scan and keypoint ablations remove E_D and E_{2D}, respectively, while retaining the same initialization. Translation and rotation smoothing rows measure two outcomes of the same before/after comparison for Equation 3, rather than separate loss-term removals. The tracking-safeguard rows evaluate the same control/full comparison in camera coordinates and after scene assembly; the safeguards are described in Section 3.3 and Appendix B.3. The acceptance-guard row separately replays candidate decisions with and without replacement screening, counting silhouette-IoU regressions.

Table 6: Effects of human fitting, object smoothing, and tracking safeguards. Each row compares the full pipeline with a component removed or a control setting on the same samples. For continuous metrics, Ablated/control and Full report median clip scores; Paired benefit reports the median improvement calculated clip by clip. Strict pass counts clips passing geometric checks. N: clips (acceptance guard: decisions); pp: percentage points.

Component / setting	N	Metric	Ablated / control	Full	Paired benefit
Temporal terms	390	Body Δ^2 p95 (mm/f^2) ↓	91.61	**11.88**	86.41% ↓
Joint second order	390	Body Δ^2 p95 (mm/f^2) ↓	17.78	**11.88**	28.91% ↓
Temporal terms (long)	130	Body Δ^2 p95 (mm/f^2) ↓	85.44	**11.08**	87.03% ↓
Depth-scan loss	390	Scan distance (mm) ↓	61.82	**60.64**	1.94 mm ↓
Keypoint loss	390	Scan distance (mm) ↓	62.62	**60.64**	2.03 mm ↓
Keypoint loss	390	Person IoU ↑	0.6812	**0.6930**	+0.0113
Translation smoothing	104	Trans. Δ^2 p95 (mm/f^2) ↓	19.24	**12.39**	41.44% ↓
Rotation smoothing	104	Rot. $\Delta\omega$ p95 (rad/f^2) ↓	0.196	**0.075**	61.61% ↓
Tracking safeguards: camera	140	Strict pass ↑	82/140	**124/140**	+30.0 pp
Tracking safeguards: camera	140	Object IoU ↑	0.0746	**0.2606**	+0.0604
Tracking safeguards: scene	140	Strict pass ↑	89/140	**128/140**	+27.9 pp
Candidate acceptance guard	900	IoU regressions ↓	15/900	**0/900**	15 fewer

Human temporal and observation consistency. Temporal regularization reduces body-motion p95 by 86.41%, with an 87.03% reduction on long clips. Depth and keypoint supervision improve

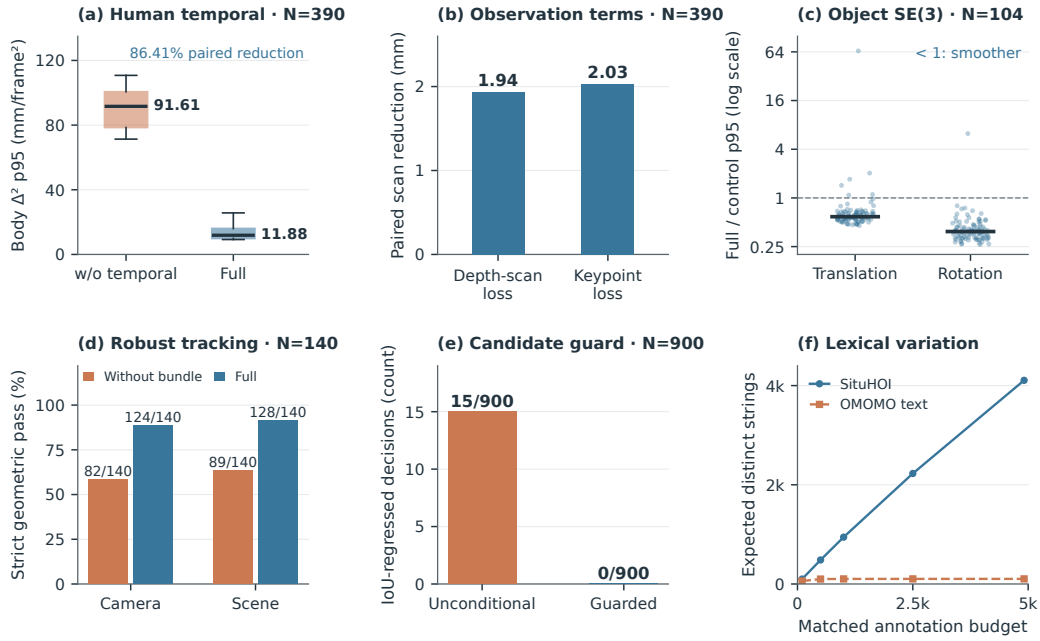


Figure 7: Pipeline ablations corresponding to Table 6: (a) human temporal p95 and (b) median paired scan-distance reductions on 390 clips; (c) temporal ratios on 104 object clips; (d) strict passes on 140 tracking clips; (e) IoU regressions over 900 candidate decisions; and (f) matched-budget lexical variation. Boxes in (a) span the 25th–75th percentiles, with 10th–90th percentile whiskers; dark bars in (a,c) mark medians.

agreement with RGB-D observations (Table 6). These results complement the independent MoCap comparisons in Section 5.1.

Object refinement and tracking safeguards. SE(3) smoothing reduces translation/rotation p95 by 41.44%/61.61% on the 104 moving clips. On difficult RGB-D clips, tracking safeguards raise automated pass counts from 82 to 124 of 140 before scene assembly and from 89 to 128 afterward. The acceptance guard reduces IoU regressions from 15 to zero over 900 archived candidate decisions. Figure 7 summarizes the component comparisons.

Human-fitting component and weight ablations. Table 7 reports component and temporal-weight ablations of human fitting, extending Table 6 on the same 390 short clips, with replication on 130 long clips. All configurations within each cohort use matched evaluation frames; scan distances are evaluated on held-out depth observations.

Increasing temporal weight from quarter to half to full reduces body p95 from 19.05 to 14.77 to 11.88 mm/frame^2, while person IoU rises from 0.6644 to 0.6804 to 0.6930. Restoring the scan loss reduces median scan distance from 61.82 to 60.64 mm; restoring keypoint supervision reduces the median 2D residual from 37.52 to 20.48 pixels. On 130 long clips, full fitting reduces body p95 from 85.44 to 11.08 mm/frame^2 and raises person IoU from 0.6993 to 0.7175, extending the temporal and silhouette improvements to longer sequences.

D FlexHOI-Bench and SharedHOI details

D.1 Temporal queries and reproducible partition

Eligible annotated segments have paired finite human/object trajectories and an object mesh. We extract full 135-frame windows at 30 Hz, with strides of 60 frames for TRAIN/HOLDOUT and 135 for VAL, appending a final tail-aligned window. A window is rejected if either root or object translation deviates by ≥ 5 m from its first position or moves ≥ 1 m between cached steps.

Table 7: Human-fitting component and weight ablations on 390 short clips and 130 long clips: cohort medians of per-clip metrics. Temporal p95 is in mm/frame^2, scan distance in mm, and keypoint residual in pixels.

Configuration	N	Body p95 ↓	Scan ↓	IoU ↑	2D ↓
Fixed short-clip cohort					
Full fitting	390	11.88	60.64	0.6930	20.48
No second-order terms	390	91.61	53.47	0.6774	14.18
No joint second-order term	390	17.78	54.25	0.6781	20.91
Quarter temporal weight	390	19.05	53.41	0.6644	21.41
Half temporal weight	390	14.77	61.95	0.6804	26.11
No scan loss	390	12.29	61.82	0.6831	29.12
No keypoint loss	390	13.78	62.62	0.6812	37.52
Long-clip replication					
Full fitting	130	11.08	46.75	0.7175	18.05
No second-order terms	130	85.44	46.58	0.6993	17.30

Each source window is stored as 72 precomputed motion states (the cache), which are interpolated at the requested query times. The cache uses source times $\mathrm{rint}(\mathrm{linspace}(0, 134, 72))/30$, not uniform 16-Hz indices. The observation boundary is a cached anchor; historical interpolation brackets never cross it. At query rate $\nu = 16$ Hz, $n_H = \max(2, \lceil \nu H \rceil + 1)$ points span $[-H, 0]$ and $n_F = \max(1, \lceil \nu F \rceil)$ points span $(0, F]$, including both requested endpoints. Thus nominal rate does not imply exact $1/16$ s spacing for every duration. The interface requires $H, F \geq 1/30$ s, $H + F \leq 4$ s, cache support, and at most 160 points. The four-second budget leaves anchor-selection margin within the source windows. Longer supports can be constructed from source recordings by resampling and retraining; they are not evaluated here.

Training samples legal anchors, then continuous durations, with 20% full-span boundary draws where feasible. It excludes $[.75, .95] \times [1.4, 1.8]$ and $[1.4, 1.65] \times [.65, .9]$ in (H, F) seconds. One duration request is shared within a batch; weights are shared across batches. The first six rows of Table 8 form the VAL selection grid; the marked pairs probe withheld duration combinations. The final three are additional checks.

The full-corpus protocol uses 12 physical rooms for TRAIN/HOLDOUT and the remaining four rooms for VAL and checkpoint selection. Within each of the 12 rooms, recordings are deterministically ranked by hash; 20%, rounded up, are assigned to HOLDOUT while retaining at least one TRAIN recording. All windows from a recording stay in one partition. Thus TRAIN and HOLDOUT share room identities but have disjoint recordings; VAL is room-disjoint from both. Room counts refer to physical rooms, not separate scans.

D.2 SHAREDHOI: MODEL AND TRAINING DETAILS

The 146D physical state contains root translation (3), body rotations (22×6), object translation/rotation (3/6), and wrist distances (2). The 6D encoding stores the first two rotation-matrix columns; each object's 256 points are fixed-seed samples of its mesh vertices in mesh coordinates. The shared network predicts 143 residual channels, replacing object 6D rotation by a three-dimensional rotation-vector correction, right-multiplied onto the goal-bridge rotation after applying Exp. Neutral SMPL-X forward kinematics recomputes wrists 20/21 after interpolation. Five memory tokens encode initial state, goal, durations, object, and scene; point MLPs pool max/mean geometry features. Attention has six heads, feed-forward width 768, and dropout .1. With t, H, F measured in seconds, the common 22D time vector contains $t/4$, phase t/H for history or t/F for future, $H/4$, $F/4$, two history/future indicators, and $\sin(2\pi f t), \cos(2\pi f t)$ at frequencies $f \in \{.25, .5, 1, 2, 4, 8, 12, 16\}$ Hz.

Reconstruction weights for root, body, object position, object rotation, and distances are 4, 1, 8, 1, and .2. Reconstruction uses mean squared error on normalized residual channels. Velocity error uses consecutive differences divided by query-time spacing, scaled to a $1/16$ s reference step, with weight .2. AdamW uses batch 128, learning rate $3 \cdot 10^{-4}$, weight decay 10^{-4}, cosine decay, gradient

clipping at one, and 120 epochs. Three seeds are 20260910/11/12. Checkpoints minimize mean normalized VAL reconstruction over the six selection pairs, checked every ten epochs.

For the object-translation refinement head, let p_0, R_0 be the last observed object pose, p_F, R_F the goal, v_0, ω_0 the observed linear and angular velocities, $u = t/F$, $\delta = \mathrm{Log}(R_0^\top R_F)$, and $b(t) = \frac{\tau}{\tau+F} F u(1-u)^2$, with $\tau = 1$ s. Then

$$p_{\mathrm{prior}}(t) = (1-u)p_0 + u p_F + b(t)\big(v_0 - (p_F - p_0)/F\big),$$
$$\hat{R}(t) = R_0 \,\mathrm{Exp}[u\delta + b(t)(\omega_0 - \delta/F)]\,. \tag{8}$$

Least-squares velocity uses the last .2 s (all history if fewer than two points), clipped to 2 m/s and 2π rad/s. The 225–128–128–3 MLP uses sigmoid-weighted linear unit (SiLU) activations and combines frozen 192D features, 22D time features, goal displacement, velocity, anchor position, and two observed wrist distances; its last layer starts at zero. Its dimensionless output $r_\phi(t)$ refines translation as

$$\hat{p}(t) = p_{\mathrm{prior}}(t) + 4u(1-u)c(F)r_\phi(t), \qquad c(F) = 0.1\,\mathrm{min}\left(\frac{F}{1\,\mathrm{s}}, 1\right)\mathrm{m}. \tag{9}$$

The envelope vanishes at $t = 0, F$, preserving both endpoints; $c(F)$ reduces the correction scale for short forecasts. Rotation follows Equation 8. Decoding restores scene coordinates, observed history, and the supplied object endpoint. Its loss is intermediate object ADE plus .01 times squared metric translation residual. The shared predictor stays frozen during 40 AdamW epochs (learning rate $3 \cdot 10^{-4}$, weight decay .01). Head selection checks six-profile VAL ADE every five epochs with proximity/rotation guards, selecting epochs 40/30/40. These extra epochs are not a compute-matched advantage claim.

Baseline implementations. The compared methods were developed for different conditioning and prediction settings, and their original interfaces do not directly match FlexHOI-Bench. We therefore adapt the authors' core implementations of CHOIS, DecHOI, InterDiff, and OMOMO to a common scene-grounded, goal-conditioned joint prediction interface with variable history and forecast durations, while retaining their core temporal architectures. A shared motion representation and geometry/goal interface align the available information and evaluation targets. The same resampled query times and 22D physical-time features explicitly encode elapsed seconds and the requested history/forecast durations, so each model can distinguish different temporal requests within one set of weights. For OMOMO, a bridge from the last observed object pose to the supplied goal provides the trajectory conditioning required by its hand-to-body stages using only the benchmark inputs. The adapted cores retain their respective reconstruction/denoising objectives on the common trajectory targets. To focus the comparison on this shared prediction task, method-specific auxiliary losses and contact/physics refinement remain outside its scope. These adaptations support consistent evaluation across model architectures, input durations, and forecast horizons.

D.3 Evaluation metrics and duration-wise results

Root and object average displacement errors (ADE) measure Euclidean translation error against the fitted dataset trajectories, in mm. Object translation and geodesic rotation errors exclude the supplied endpoint $t = F$; root, body-rotation, and proximity metrics include it. Errors are averaged over valid future queries within each window, then equally across windows.

For each wrist and query, proximity is positive when its minimum distance to the 256 object samples is below 0.20 m. Predicted and reference states are computed from their respective human/object geometry, rather than the model's auxiliary distance channels. Within each duration pair and seed, true positives (TP), false positives (FP), and false negatives (FN) are pooled over both wrists, future queries, and all evaluated windows; $\mathrm{F1} = 100 \cdot 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$. An empty denominator is undefined. The "All 11 pairs" columns in Table 3 equally average pair-level scores within each seed, then report the mean $\pm$ sample standard deviation across the three training seeds. Individual duration columns use the same seed-level summary.

The deterministic Hold reference retains the last observed human root and body pose, while linearly interpolating object translation and using SLERP for object rotation toward the supplied goal. Under the same eleven-pair aggregation as Table 3, rotation errors (degrees) are 12.51 for Hold,

12.53 ± 0.00 for CHOIS, 12.53 ± 0.00 for DecHOI, 14.26 ± 0.14 for InterDiff, 12.51 ± 0.00 for OMOMO, and 12.47 ± 0.00 for SharedHOI. Body-rotation error averages geodesic angles over all future queries and 22 joints: the root uses scene-frame orientation and the remaining joints use parent-relative rotations. Means over seeds and the eleven pairs are 7.36/6.84/6.94/6.90/7.88/5.68 degrees for Hold/CHOIS/DecHOI/InterDiff/OMOMO/SharedHOI. Across the six selection pairs on VAL, object ADE is 55.736 mm for the unrefined predictor, 49.529 for a goal bridge, 47.887 after velocity correction, 47.818 with a kinematics-only head, and 47.466 with frozen shared features.

Table 8: SharedHOI across all 11 duration pairs on all 6,948 HOLDOUT windows. Entries are three-seed means; object ADE also shows sample SD. Daggers mark pairs inside excluded training-duration regions.

H (s)	F (s)	Root (mm)	Object (mm)	Rotation (deg)	F1 (%)
0.25	0.50	102.09	21.43±0.01	5.54	85.75
0.50	1.00	145.45	45.13±0.06	9.39	79.55
0.50	3.50	331.21	123.83±0.70	23.36	58.90
1.00	2.00	217.63	76.68±0.26	15.93	69.29
2.00	1.00	153.42	43.07±0.05	9.19	78.25
3.00	0.50	96.28	19.68±0.02	5.23	85.30
$0.85^{\dagger}$	1.60	186.86	64.83±0.22	13.37	73.14
$1.50^{\dagger}$	0.80	131.88	33.41±0.05	7.64	81.03
0.25	3.75	355.04	132.92±0.66	24.44	58.08
3.75	0.25	64.87	9.39±0.00	2.77	90.93
0.37	2.81	274.27	100.41±0.53	20.29	63.51

E External-model evaluation details

Transfer protocol and adaptations. The source subset contains 512 TRAIN segments, selected round-robin from hash-ranked recording-object groups. Hash selection retains at most 16 non-overlapping 35-frame windows per segment within approved TRAIN ranges, yielding 6,562 windows. BEHAVE TRAIN/VAL/TEST contain 1,450/2,239/3,602 windows from 37/50/85 disjoint recordings; TEST uses Date03. Both domains use a neutral 22-joint skeleton and coordinates anchored to the first observed root pose. Object meshes are centered by their vertex mean, with translations adjusted to preserve world geometry; ADE follows this center's trajectory. InterDiff uses the author SMPL diffusion branch without interaction correction and 50-step denoising diffusion implicit model (DDIM) sampling. HOIMotion's discrete cosine transform and graph convolution networks consume past joints, fitted head direction, and object box corners, and predict body joints. We evaluate fixed final checkpoints, averaging errors over future frames and test windows within each seed. The fixed 3k source-only HOIMotion checkpoints, evaluated with zero BEHAVE updates, achieve 93.27 ± 1.61 mm MPJPE. Extrapolating the mean of the last two observed joint displacements gives 123.73 mm; a static-body reference gives 127.67 mm.

Transfer gains and reference conditions. Using the reported seed means in Table 4, InterDiff's object ADE decreases from 157.39 to 141.32 mm at the short budget and from 155.24 to 139.36 mm at the long budget, reductions of 16.07 and 15.88 mm. HOIMotion's body MPJPE decreases from 67.93 to 56.59 mm and from 67.73 to 55.43 mm, reductions of 11.34 and 12.30 mm. The table reports sample SD across eight seeds for InterDiff and three for HOIMotion. Both methods use matched total budgets of 6k and 12k updates. Pretraining or BEHAVE warm-up lasts 3k updates, followed by 3k or 9k updates on BEHAVE, respectively. Conditions share initial weights and target batches, with optimizers reset before target training.

At the long budget, InterDiff's static-object reference is 138.32 mm and its full-target-training reference (6,741 windows) is 131.16 mm. These references distinguish the pretraining benefit from absolute forecasting performance.

E.1 LANGUAGE AND SCENE-CONTEXT ABLATIONS

TMR (Petrovich et al., 2023) evaluates text-motion retrieval and SceneDiffuser (Huang et al., 2023) scene-conditioned prediction. Removed retains only object-name text or removes background scene points; Shuffled permutes the corresponding training context within object categories.

Table 9: Language and scene-context ablations on SituHOI HOLDOUT. Values are means ± sample SD over three runs; bold denotes the best mean within each row.

Method	Context	Metric	Updates	Full	Removed	Shuffled
TMR	Text	R@1 (%) ↑	6k	**8.91±1.21**	3.11±0.47	3.42±0.94
SceneDiffuser	Scene	MPJPE (mm) ↓	12k	**792.40±11.15**	803.74±25.11	804.93±20.24
	Scene	MPJPE (mm) ↓	50k	**760.11±18.05**	809.53±31.61	802.52±22.29

Semantic retrieval. TMR uses one entry per fitted segment under the recording-level partition, encoding 64 frames across the complete segment with 22 body joints and eight object corners. A common 20 m range screen relative to the initial root excludes one TRAIN and three VAL segments, retaining 3,432/1,365/954 TRAIN/VAL/HOLDOUT segments. All conditions train from scratch for 6k updates with three seeds. Positives share the normalized full description and object category; R@1 measures the probability of a positive top-ranked motion, averaging tied ranks. Shuffling affects training pairs only. Full descriptions improve HOLDOUT R@1 over object-name and shuffled-text controls by 5.80/5.49 percentage points (Table 9), demonstrating information beyond object identity. Joint seed/recording intervals are $[2.16, 9.64]$/$[1.93, 9.19]$, and seed-t intervals are $[1.63, 9.97]$/$[1.72, 9.25]$. VAL R@1 is 3.47/1.42/1.76% in the same condition order, with gain intervals crossing zero. Uniform same-object retrieval reaches 2.47% on HOLDOUT.

Scene-conditioned forecasting. SceneDiffuser uses PointNet to encode background/object points and predicts body joints and object corners from motion history. Under the recording-level split, HOLDOUT contains 2,360 forecasting windows, taking at most three non-overlapping four-second windows per segment. Full scene context yields the lowest mean MPJPE at both budgets, reducing it by 6.10%/5.29% over Removed/Shuffled at 50k updates (Table 9). These separate ablations support the value of aligned language and geometry.