PRIORHOSI: COMPOSING MOTION PRIORS FOR HUMAN–OBJECT–SCENE INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human–object–scene interaction (HOSI) generation aims to synthesize coordinated human and object motion compatible with the surrounding scene. Existing HOSI datasets remain scarce, while high-quality, diverse human–object interaction (HOI) and human–scene interaction (HSI) datasets offer complementary supervision. The challenge is to adapt human–object motion to scene geometry while preserving the contact relationships that coordinate the interaction. We propose PriorHOSI, a framework that learns a human–object interaction prior (HOIPrior) and a human–scene interaction prior (HSIPrior) separately and composes them at inference time. The key idea is to use HSIPrior's scene-adapted human motion to guide joint updates of the body and object through HOIPrior. Diffusion noise optimization realizes these updates by fitting the generated motion to the scene-adapted target while constraining hand positions in the object's coordinate frame. To support repeated updates, HOIPrior uses relation-aware consistency distillation to reduce sampling steps and retain hand–object geometry. To better exploit scene geometry, HSIPrior uses body-part-aware residual refinement to propose scene-aware motion adjustments during editing. Experiments on the InfBaGel benchmark demonstrate that PriorHOSI achieves strong task performance and hand–object contact while improving human–scene compatibility.

1 INTRODUCTION

Human activities often combine object interaction with movement through cluttered environments. Carrying a table across a room, for example, requires the body and object to move together while clearing nearby furniture. Synthesizing such behavior, known as human–object–scene interaction (HOSI) generation, supports applications in character animation, simulation, and embodied AI (Zou et al., 2026). Its central difficulty lies in coordinating three coupled relationships: human–object contact, human–scene interaction, and object–scene clearance. A change that improves one relationship can compromise another, such as a body adjustment that avoids furniture but separates the hands from the object.

Existing interaction datasets capture complementary aspects of this problem. GRAB (Taheri et al., 2020) records whole-body grasping of everyday objects, while OMOMO (Li et al., 2023) captures interactions with larger movable objects, including tables, chairs, and suitcases. These resources support learning the close coordination of body pose, hand contact, and object motion (Taheri et al., 2022; Li et al., 2024; Diller & Dai, 2024). Scene-oriented datasets such as TRUMANS (Jiang et al., 2024b) and LINGO (Jiang et al., 2024a) instead emphasize human behavior in furnished environments, covering locomotion, reaching, and interaction with scene elements. Their geometry and motion annotations provide supervision for adapting body movement to the environment. Broad HOSI coverage requires these aspects to occur together across combinations of objects, instructions, and scene layouts, making comprehensive capture particularly demanding (Zou et al., 2026).

Recent HOSI methods address this data gap through structured generation and shared learning across interaction sources. HOSIG (Yao et al., 2026) combines grasp synthesis, navigation, and scene-guided motion generation. InfBaGel (Zou et al., 2026) combines supervision from OMOMO and LINGO, using dynamic scene perception and consistency-based refinement to generate interactions in unseen scenes. This hybrid supervision establishes a strong baseline. Its reported data-mixture study also reveals a balance between objectives: adding HSI supervision improves several scene-

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.

penetration measures, while task success and hand–object contact can decrease. This observation motivates an alternative use of the same complementary resources: learning specialized HOI and HSI priors separately, then coordinating their predictions for the target interaction.

Composing these priors requires reconciling their different outputs and geometric responsibilities. An HSI prior proposes how the human should move within a scene, whereas an HOI prior generates the body and object together. Motion-prior composition can combine learned forms of control (Shafir et al., 2024; Ruiz-Ponce et al., 2025), but this setting also requires transferring a body-only scene adjustment to a coupled human–object sequence. The central question is how to realize that adjustment while retaining the contact relationships that define the interaction. Expressing hand positions in the object's coordinate frame allows us to constrain their positions relative to the object while permitting coordinated changes in the global placement of the body and object.

We propose PriorHOSI, a framework that learns a human–object interaction prior (HOIPrior) and a human–scene interaction prior (HSIPrior) separately and composes them at inference time. HOIPrior learns coupled human–object motion from OMOMO without scene conditioning, while HSIPrior learns human motion from LINGO without object conditioning. Given a current human–object interaction, HSIPrior proposes a scene-adapted human motion. Diffusion noise optimization (DNO) (Karunratanakul et al., 2024) fits this target through the frozen HOIPrior sampler, jointly updating the body and object under constraints on object-relative hand positions. We refresh the scene target after each update and retain contact references from the initial interaction. This turns the two specialized predictions into a coupled editing process.

We adapt each prior to its role in this repeated editing process. For HOIPrior, consistency distillation (Song et al., 2023) reduces the sampling steps traversed by each noise update. Our relation-aware supervision additionally transfers the teacher's object-relative hand positions at contact frames, making interaction geometry an explicit distillation target. For HSIPrior, body-part-aware residual refinement queries local scene distances and gradients around the predicted body. These queries describe the geometry encountered by the torso and limbs, complementing the broader occupancy context. Motion supervision and recorded distances to scene surfaces train the residual to account for both obstacle clearance and intended surface proximity. Together, the two designs support efficient repeated sampling and geometrically informed scene targets.

We evaluate on the InfBaGel benchmark using its data processing and evaluation protocol. Relative to InfBaGel, PriorHOSI improves hand–object contact from 76.96% to 87.54% and task success from 81.45% to 83.16%, while reducing all three human–scene penetration metrics. Ablations examine how contact constraints complement scene guidance and how the prior-specific designs affect the final interaction. The results support composing specialized priors for scene adaptation while retaining human–object coordination. Our contributions are:

- A framework that composes separately trained HOIPrior and HSIPrior through diffusion noise optimization, using scene-adapted human targets and object-centric contact constraints to guide joint motion updates.
- Relation-aware consistency distillation that transfers hand–object geometry to HOIPrior for repeated motion sampling during editing.
- Body-part-aware residual refinement for HSIPrior that uses local scene geometry and existing human–scene motion supervision to improve its scene-adapted editing targets.

2 Related Work

2.1 Human–Object and Human–Scene Interaction

Human–object interaction (HOI) research studies the coordination of body and object motion through contact. GRAB (Taheri et al., 2020) captures whole-body grasping with articulated hands, and GOAL (Taheri et al., 2022) generates a whole-body grasp pose and the motion needed to reach it. For larger movable objects, OMOMO (Li et al., 2023) generates human motion from object trajectories, while CHOIS (Li et al., 2024) jointly generates human and object motion from language and sparse object waypoints. Explicit contact modeling further strengthens this coordination: InterDiff (Xu et al., 2023) predicts future interactions with physics-informed correction based on

contact-relative motion, and CG-HOI (Diller & Dai, 2024) jointly models motion and contact, using predicted contact for inference-time guidance.

Human–scene interaction (HSI) research emphasizes how body configurations and actions relate to environmental geometry. SAMP (Hassan et al., 2021) generates stochastic scene-aware motion, COINS (Zhao et al., 2022) synthesizes interaction poses with compositional semantic control, and DIMOS (Zhao et al., 2023) combines motion primitives with scene-aware control for navigation and interaction. Scene affordances also connect language instructions to geometrically compatible motion (Wang et al., 2024). Extending this line to dynamic interactions, TRUMANS (Jiang et al., 2024b) combines large-scale capture with local occupancy-conditioned generation and object-motion optimization. LINGO (Jiang et al., 2024a) coordinates locomotion, reaching, and interaction through text and goal conditioning, with object motion following hand-relative attachment and release.

These lines of work provide complementary supervision for contact coordination and scene compatibility, with dynamic interaction methods covering aspects of both. We learn HOIPrior from OMOMO and HSIPrior from LINGO. Object-relative hand geometry serves as a reference for preserving the interaction, while body-part-aware scene geometry informs the human motion adjustments proposed by HSIPrior.

2.2 Human–Object–Scene Interaction Generation

Human–object–scene interaction (HOSI) generation combines object coordination and scene compatibility within a single task. HOSIG (Yao et al., 2026) organizes generation around grasp-pose synthesis, navigation, and scene-guided motion generation. UniHM (Geng et al., 2025) integrates human motion generation and object interaction in indoor scenes. InfBaGel (Zou et al., 2026) generates interactions through iterative refinement with dynamic scene perception and bump-aware collision guidance. To address the scarcity of HOSI data, it combines LINGO with OMOMO motions embedded in synthetic scenes. These scenes are constructed to accommodate the recorded human and object trajectories, with obstacles placed outside their joint motion volume. Such pairs capture compatibility with existing motions but offer limited direct supervision for obstacle-induced adjustments, such as turning sideways or taking a detour during object manipulation.

In contrast to InfBaGel's hybrid training, PriorHOSI learns contact coordination from OMOMO and scene adaptation from LINGO in separate priors, then composes them at inference time. HSIPrior supplies a scene-adapted human target, and HOIPrior realizes it through joint human–object updates constrained by object-relative hand positions. This transfers scene-dependent adjustments to the coupled interaction while preserving hand–object coordination as an explicit editing objective.

2.3 Motion Prior Composition and Editing

Learned motion priors support the reuse and combination of generative capabilities. Building on diffusion-based generation such as MDM (Tevet et al., 2023), PriorMDM (Shafir et al., 2024) explores sequential, parallel, and model composition. Composable diffusion models (Liu et al., 2022) combine learned distributions for visual generation, while MixerMDM (Ruiz-Ponce et al., 2025) learns to combine motion denoiser outputs according to the motion, conditions, and diffusion timestep. SceneAdapt (Cho et al., 2025) uses motion inbetweening to transfer scene knowledge between disjoint text–motion and scene–motion resources. Our composition connects priors with different outputs: a scene-conditioned human sequence and a coupled human–object sequence. The human target links these two motion spaces.

Optimization-based editing provides a mechanism for realizing this link. Guided Motion Diffusion (Karunratanakul et al., 2023) introduces spatial guidance for controllable motion synthesis, and diffusion noise optimization (DNO) (Karunratanakul et al., 2024) propagates task-loss gradients through a frozen sampler to its initial noise. HOIDiNi (Ron et al., 2025) applies DNO to HOI through an object-centric planning phase, including hand–object contact locations, followed by a human-centric motion refinement phase. We use DNO to compose the two specialized priors: each iteration jointly updates human and object motion toward a refreshed HSI target, while retaining contact references from the initial interaction. This couples repeated scene adjustments to the geometry of the original human–object interaction.

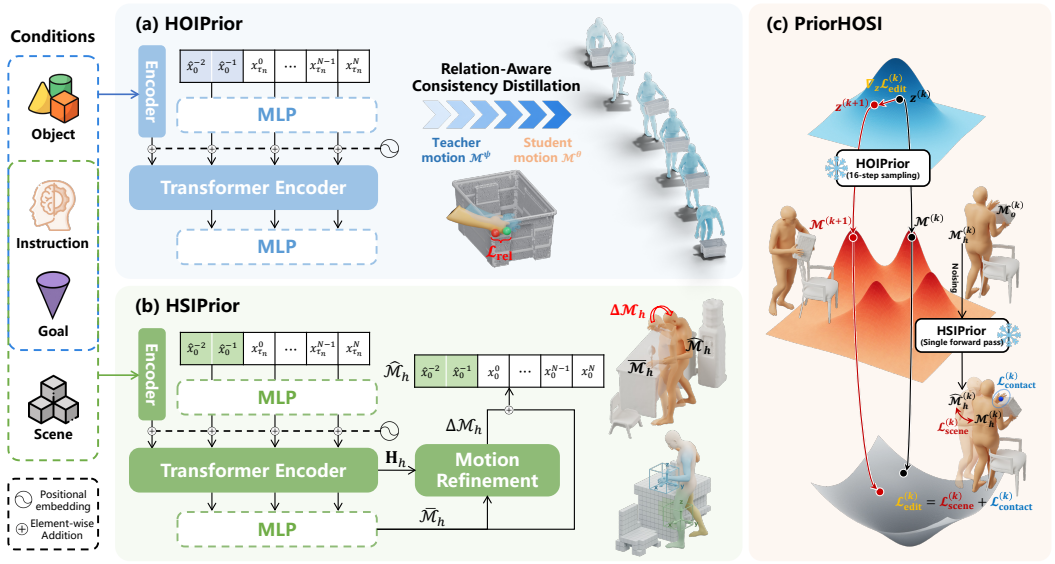


Figure 1: Overview of PriorHOSI. (a) HOIPrior learns coupled human–object motion through relation-aware consistency distillation. (b) HSIPrior refines human motion using body-part-aware scene geometry. (c) At inference time, scene-adapted targets from HSIPrior and object-centric contact constraints guide noise optimization through HOIPrior, jointly updating human and object motion.

3 Method

PriorHOSI generates human–object motion in a target scene by composing two specialized motion priors (Figure 1). HSIPrior proposes a scene-adapted human motion, and HOIPrior realizes this target through joint human–object updates constrained by contact geometry. We first define the two priors and their representations, then describe the training designs that support their respective roles, and finally explain the inference-time composition through noise optimization.

3.1 Problem Formulation

Problem formulation. Given a 3D scene $\mathcal{S}$, movable-object geometry $\mathcal{O}$, a text instruction $\mathcal{T}$, and task goals $\mathcal{G} = (\mathcal{G}_h, \mathcal{G}_o)$ for the human pelvis and object centroid, we seek a coupled motion sequence $\mathcal{M} = (\mathcal{M}_h, \mathcal{M}_o)$ that performs the instructed interaction while maintaining hand–object contact and respecting scene geometry. We model the target distribution $p_{\text{HOSI}}(\mathcal{M}_h, \mathcal{M}_o \mid \mathcal{S}, \mathcal{O}, \mathcal{T}, \mathcal{G})$ by learning two complementary priors separately from HOI and HSI data:

$$p_{\text{HOI}}(\mathcal{M}_h, \mathcal{M}_o \mid c_O), \qquad p_{\text{HSI}}(\mathcal{M}_h \mid c_S), \tag{1}$$

where $c_O = (\mathcal{O}, \mathcal{T}, \mathcal{G})$ and $c_S = (\mathcal{S}, \mathcal{T}, \mathcal{G}_h)$. We compose these priors at inference time to adapt coupled human–object motion to the scene.

Human and object motion. At each frame t, human motion $\mathcal{M}_h = (T_h, R_h)$ is represented by the root translation $T_h(t) \in \mathbb{R}^3$ and 6D rotations $R_h(t) \in \mathbb{R}^{22\times 6}$ of 22 body joints. We use SMPL-X (Pavlakos et al., 2019) to reconstruct the human mesh from these parameters. Object motion $\mathcal{M}_o = (T_o, R_o)$ is represented by the centroid translation $T_o(t) \in \mathbb{R}^3$ and rotation matrix $R_o(t) \in \mathrm{SO}(3)$, which transform the centered reference object geometry $\mathcal{O}$ into world coordinates.

Input conditions. Both priors receive the text instruction and human pelvis goal; HOIPrior additionally receives the object centroid goal and object geometry, while HSIPrior receives scene geometry. We encode the object using a basis point set (BPS) (Prokudin et al., 2019) of 1,024 vectors from fixed basis points to its surface. For HSIPrior, local occupancy grids provide scene context to the first-stage denoiser, while a signed distance field (SDF) $d_{\mathcal{S}} : \mathbb{R}^3 \to \mathbb{R}$, positive outside scene surfaces, provides distances and gradients around the predicted body for residual refinement.

3.2 HOIPRIOR: RELATION-AWARE CONSISTENCY DISTILLATION

HOIPrior learns coordinated human–object motion and contact relationships from OMOMO (Li et al., 2023). We first train a conditional diffusion teacher f_ψ to generate coupled human–object motion $\mathcal{M}$ given c_O. Since noise optimization repeatedly differentiates through motion sampling, we distill the teacher into a few-step consistency model f_θ to reduce the cost of each editing update. Alongside consistency matching, we introduce relation-aware supervision that transfers the teacher's hand–object geometry to the student, preserving the interaction structure needed for subsequent edits. Teacher training and distillation details are provided in Appendix C.1.

Consistency distillation. Consistency distillation (Song et al., 2023) trains f_θ to map noisy states on the same teacher denoising trajectory to a common clean motion. Given a noise-corrupted human–object motion x_{τ_n}, the frozen teacher f_ψ takes one DDIM solver step (Song et al., 2021) to obtain $\widehat{x}_{\tau_{n-1}}$ at the next, lower noise level. We match the student's prediction at x_{τ_n} to the prediction of a target consistency network $f_{\bar{\theta}}$ at $\widehat{x}_{\tau_{n-1}}$:

$$\mathcal{L}_{\mathrm{CD}} = \mathbb{E}\Big[\big\| f_\theta(x_{\tau_n}, \tau_n; c_O) - f_{\bar{\theta}}(\widehat{x}_{\tau_{n-1}}, \tau_{n-1}; c_O)\big\|_2^2\Big], \tag{2}$$

where $\bar{\theta}$ is an exponential moving average of the student parameters θ. The teacher transition and target-network prediction are held constant during the student update.

Object-centric relation distillation. We compare complete student and teacher samples generated from the same initial noise z and condition c_O. With G_θ and G_ψ denoting their respective sampling procedures, these motions are $\mathcal{M}^\theta = G_\theta(z; c_O)$ and $\mathcal{M}^\psi = G_\psi(z; c_O)$, respectively. To identify hand–object contacts, we use four query points: the wrist and the base of the middle finger on each hand. Thresholding each point's distance to the nearest object vertex in the teacher motion yields the contact set $\mathcal{C}^\psi$, where $(t, j) \in \mathcal{C}^\psi$ indicates that point j is in contact at frame t.

Using world-space positions from forward kinematics (FK), we express query point j at frame t in the object frame as $r_{t,j}(\mathcal{M}) = R_o(t)^\top \big(\mathrm{FK}(\mathcal{M}_h)_{t,j} - T_o(t)\big)$, where $T_o(t)$ is the object centroid in world coordinates and $R_o(t)$ maps the centered reference geometry to the world frame. We match the student and teacher relations over the contact set:

$$\mathcal{L}_{\mathrm{rel}} = \mathbb{E}_{z, c_O}\left[\frac{1}{\max(1, |\mathcal{C}^\psi|)} \sum_{(t,j)\in\mathcal{C}^\psi} \big\| r_{t,j}(\mathcal{M}^\theta) - r_{t,j}(\mathcal{M}^\psi)\big\|_2^2\right]. \tag{3}$$

This loss transfers the teacher's contact geometry in the object frame. A shared rigid transformation of the human and object leaves these relations unchanged, so the supervision preserves a geometric reference while allowing coordinated changes in global placement.

The complete distillation objective is

$$\mathcal{L}_{\mathrm{HOI}} = \mathcal{L}_{\mathrm{CD}} + \lambda_r \mathcal{L}_{\mathrm{rel}}. \tag{4}$$

We freeze the distilled prior and use its sampler G_θ for noise optimization.

3.3 HSIPRIOR: BODY-PART-AWARE RESIDUAL REFINEMENT

HSIPrior learns a scene-conditioned motion prior from LINGO (Jiang et al., 2024a). To provide motion-dependent scene guidance for noise optimization, we formulate scene adaptation as residual refinement: the model uses body-aligned scene geometry to correct its own motion predictions. The refinement accounts for differences in local geometry across body parts and learns from paired motion–scene data, with the corrections supervised through the refined motion predictions.

Predicting body-part-aware corrections. HSIPrior operates in two stages within a single denoising evaluation. Given noisy motion $\mathcal{M}_{h,\tau}$, the first-stage Transformer predicts a clean motion estimate $\bar{\mathcal{M}}_h$ and provides the associated latent representation $\mathbf{H}_h$, with $(\bar{\mathcal{M}}_h, \mathbf{H}_h) = f_\phi^{(1)}(\mathcal{M}_{h,\tau}, \tau; c_S)$. The second stage queries SDF values and gradients around $\mathrm{FK}(\bar{\mathcal{M}}_h)$ and encodes them into torso, arm, and leg features, preserving joint identities and aggregating temporal

changes in local geometry. These body-aligned features, denoted by $\mathcal{Q}_{\mathcal{S}}(\mathrm{FK}(\bar{\mathcal{M}}_h))$, complement the broader scene context encoded by occupancy features. The second-stage predictor $f_\phi^{(2)}$ combines these features with $\mathbf{H}_h$ to predict a residual motion correction:

$$\Delta\mathcal{M}_h = f_\phi^{(2)}\left(\mathbf{H}_h, \mathcal{Q}_{\mathcal{S}}(\mathrm{FK}(\bar{\mathcal{M}}_h))\right), \qquad \widehat{\mathcal{M}}_h = \bar{\mathcal{M}}_h + \Delta\mathcal{M}_h. \tag{5}$$

During editing, we obtain $\mathcal{M}_{h,\tau}$ by adding noise to the human motion in the current HOI candidate. The refined prediction $\widehat{\mathcal{M}}_h$ then provides its scene-adapted editing target.

Learning corrections from motion supervision. After pretraining the first-stage denoiser, we train both stages jointly against ground-truth motion $\mathcal{M}_h^*$, applying all losses to the refined prediction $\widehat{\mathcal{M}}_h$. The residual update in Equation 5 lets the motion reconstruction loss supervise the correction from $\bar{\mathcal{M}}_h$ toward $\mathcal{M}_h^*$. We further introduce a geometry loss to guide these corrections using ground-truth distances to nearby scene surfaces:

$$\mathcal{L}_{\mathrm{geo}} = \frac{1}{\max(1, |\mathcal{V}|)} \sum_{(t,j)\in\mathcal{V}} \ell_{\mathrm{sL1}}\left(d_{\mathcal{S}}\left(\mathrm{FK}(\widehat{\mathcal{M}}_h)_{t,j}\right), d_{\mathcal{S}}(\mathrm{FK}(\mathcal{M}_h^*)_{t,j})\right), \tag{6}$$

where $d_{\mathcal{S}}(\cdot)$ denotes the signed distance from a spatial point to the scene surface, ℓ_{sL1} is a scaled smooth-L_1 loss, and $\mathcal{V}$ selects valid joint queries near scene surfaces. Matching recorded surface distances encourages intended proximity, such as support during sitting, while supervising corrections to the predicted motion. The query construction and training details are given in Appendix C.2.

3.4 PRIORHOSI: PRIOR COMPOSITION THROUGH NOISE OPTIMIZATION

Diffusion noise optimization. Diffusion noise optimization (DNO) (Karunratanakul et al., 2024) edits motion by optimizing the initial noise through a pretrained diffusion sampler. In PriorHOSI, HOIPrior generates the human–object motion to be edited, while HSIPrior provides scene guidance for noise optimization. Starting from $z^{(0)} \sim \mathcal{N}(0, I)$, we generate the candidate at iteration k using G_θ, with all intermediate sampling noise held fixed:

$$\mathcal{M}^{(k)} = G_\theta(z^{(k)}; c_O) = (\mathcal{M}_h^{(k)}, \mathcal{M}_o^{(k)}). \tag{7}$$

The editing criterion combines HSIPrior's candidate-dependent scene guidance with constraints that preserve hand–object relations in the initial interaction $\mathcal{M}^{(0)}$:

$$\mathcal{L}_{\mathrm{edit}}^{(k)} = \mathcal{L}_{\mathrm{scene}}^{(k)} + \mathcal{L}_{\mathrm{contact}}^{(k)}. \tag{8}$$

We differentiate this objective through the HOIPrior sampler to update the initial noise:

$$g^{(k)} = \nabla_z \mathcal{L}_{\mathrm{edit}}^{(k)}(G_\theta(z; c_O))\Big|_{z=z^{(k)}}. \tag{9}$$

Since G_θ jointly generates human and object motion, each noise update can adapt both components. We refresh the scene target from the updated interaction and repeat for K iterations, returning $\mathcal{M}^{(K)} = G_\theta(z^{(K)}; c_O)$.

Scene guidance from HSIPrior. The scene term measures how closely the current human motion matches HSIPrior's scene-adapted target. We add Gaussian noise to $\mathcal{M}_h^{(k)}$ at a fixed diffusion level τ_S and obtain $\widehat{\mathcal{M}}_h^{(k)}$ with a single forward denoising pass through HSIPrior. We use this prediction as an editing target in global joint coordinates, averaging over the set $\mathcal{I}$ of frame–joint pairs used for scene guidance:

$$\mathcal{L}_{\mathrm{scene}}^{(k)} = \frac{1}{|\mathcal{I}|} \sum_{(t,j)\in\mathcal{I}} \left\| \mathrm{FK}(\mathcal{M}_h^{(k)})_{t,j} - \mathrm{FK}(\widehat{\mathcal{M}}_h^{(k)})_{t,j} \right\|_2^2. \tag{10}$$

We hold the target fixed during each gradient evaluation and recompute it from the updated interaction. Gradients pass through HOIPrior, allowing the human target to guide both components of its coupled output. The target directly supervises human motion; object motion responds through this joint generator and the contact constraints below.

Interaction-preserving criterion. The contact term links scene-driven body changes to compatible object motion. We therefore constrain the object-relative hand positions $r_{t,j}$ defined in Section 3.2. We extract a contact set $\mathcal{C}$ from the initial interaction $\mathcal{M}^{(0)}$ using the same distance criterion, retaining temporally persistent contacts. For each $(t, j) \in \mathcal{C}$, we seek to maintain $\|r_{t,j}(\mathcal{M}^{(k)}) - r_{t,j}(\mathcal{M}^{(0)})\|_2 \leq \delta$, where δ is the allowed relative-hand displacement. The contact set and reference relations remain fixed throughout editing, allowing coordinated changes in global placement while limiting hand–object drift. We optimize these constraints with an augmented-Lagrangian penalty, whose motion-dependent part is an adaptive squared-hinge loss:

$$\mathcal{L}_{\text{contact}}^{(k)} = \frac{\rho_k}{2} \sum_{(t,j)\in\mathcal{C}} \max\Big(\Big\| r_{t,j}(\mathcal{M}^{(k)}) - r_{t,j}(\mathcal{M}^{(0)}) \Big\|_2 - \delta_{t,j}^{(k)},\, 0\Big)^2, \tag{11}$$

where $\rho_k > 0$ controls the penalty strength and $\delta_{t,j}^{(k)} = \delta - \mu_{t,j}^{(k)}/\rho_k$ is an adaptive threshold determined by the nonnegative contact multiplier $\mu_{t,j}^{(k)}$. Sampling, optimization, and multiplier updates are detailed in Appendix C.3.

4 Experiments

4.1 Experimental Setup

Implementation details. We train HOIPrior on OMOMO and HSIPrior on LINGO, following the data processing of InfBaGel (Zou et al., 2026). All training uses four NVIDIA RTX 3090 GPUs, a batch size of 512, and the Adam optimizer, with learning rates of 10^{-4} for HOIPrior and 2×10^{-4} for HSIPrior. We pretrain HOIPrior for 500 epochs and distill it for 200 epochs. We pretrain HSIPrior for 250 epochs, followed by 150 epochs of joint training with body-part-aware residual refinement. Generation is performed offline with 150 editing iterations, each differentiating through the 16-step HOIPrior sampler.

Benchmark and baselines. We evaluate on the InfBaGel benchmark, which contains 469 tasks with seven object categories in 67 indoor scenes. Each task specifies an initial interaction, scene and object geometry, a text instruction, and human and object goals. We compare with TRUMANS (Jiang et al., 2024b), LINGO (Jiang et al., 2024a), CHOIS (Li et al., 2024), HOSIG (Yao et al., 2026), and InfBaGel (Zou et al., 2026) under the HOSI evaluation protocol. All baselines are trained or fine-tuned under InfBaGel's hybrid-data setting, combining pseudo-HOSI samples from OMOMO augmented with synthetic scene occupancy and HSI samples from LINGO with real scene geometry.

Metrics. Under the InfBaGel evaluation protocol, we measure task completion using human and object goal errors (T_h, T_o) and success rate ($S\%$), where both final goal errors must be below $10\,\text{cm}$. We evaluate motion and interaction quality with foot sliding (FS), contact percentage ($C\%$), and body–object penetration (P_{body}). Human–scene (HS) and object–scene (OS) penetration are measured by P_{mean}, P_{max}, and the percentage of penetrating frames ($P_f\%$). The first two are the temporal mean and maximum of per-frame sums of vertex penetration depths, following the benchmark definition.

4.2 Comparison with Existing Methods

Quantitative comparison. PriorHOSI improves human–object interaction quality and human–scene compatibility while achieving the highest task success on the InfBaGel benchmark (Table 1). Compared with InfBaGel, contact rate increases from 76.96% to 87.54%, and body–object penetration decreases by 37.2% (5.05 to 3.17), yielding the best results on both metrics among the compared methods. These gains accompany a reduction in object goal error from 7.94 to 6.05 and an increase in task success from 81.45% to 83.16%. PriorHOSI also achieves the lowest foot sliding and HS penetration across all three metrics, with a 10.4% reduction in HS P_{mean} relative to InfBaGel. Object–scene results show a trade-off: OS penetrating frames decrease from 22.32% to 21.62%, while P_{mean}/P_{max} increase from 12.45/93.37 to 13.75/94.11. Together, these results show that PriorHOSI improves contact quality and human–scene compatibility while achieving more accurate and successful task completion.

Table 1: Comparison on the InfBaGel benchmark. HO, HS, and OS denote human–object, human–scene, and object–scene pairs. P_{body}, P_{mean}, and P_{max} are aggregate vertex-penetration depths in meters.

Method	Task accuracy			FS↓	HO interaction		HS penetration ↓			OS penetration ↓		
	$T_h \downarrow$	$T_o \downarrow$	$S\% \uparrow$		$C\% \uparrow$	$P_{body} \downarrow$	P_{mean}	P_{max}	$P_f\%$	P_{mean}	P_{max}	$P_f\%$
TRUMANS	–	73.52	1.38	0.72	45.12	6.38	9.35	68.41	90.82	37.32	173.12	32.60
LINGO	**3.04**	17.97	39.45	0.57	53.43	7.58	7.09	57.45	91.96	22.55	123.56	**20.56**
CHOIS	–	22.43	0.43	0.40	74.66	3.40	27.66	132.62	45.60	30.62	175.43	21.91
HOSIG	9.37	–	11.73	0.91	47.82	3.84	74.14	403.60	80.06	126.22	789.25	31.80
InfBaGel	4.37	7.94	81.45	0.15	76.96	5.05	3.17	37.09	26.74	**12.45**	**93.37**	22.32
PriorHOSI	3.13	**6.05**	**83.16**	**0.10**	**87.54**	**3.17**	**2.84**	**34.38**	**24.52**	13.75	94.11	21.62

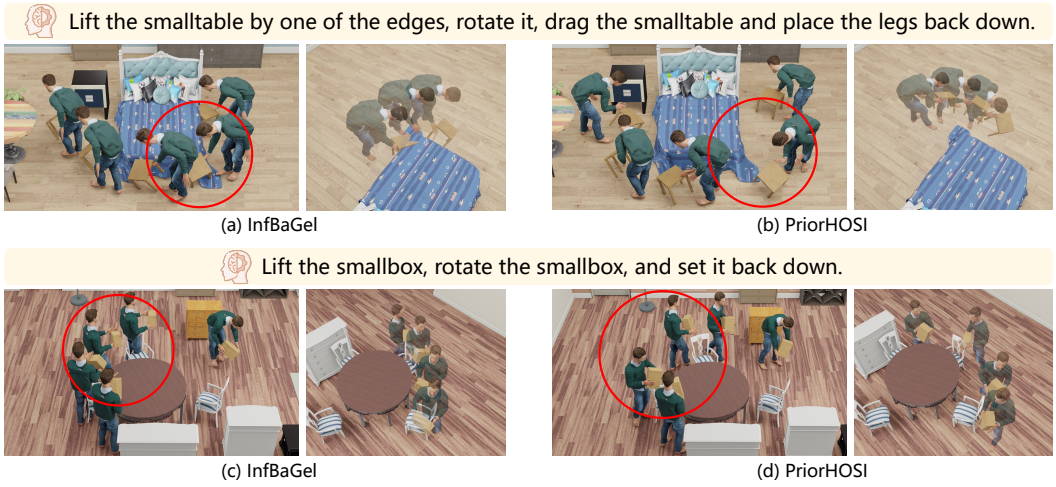


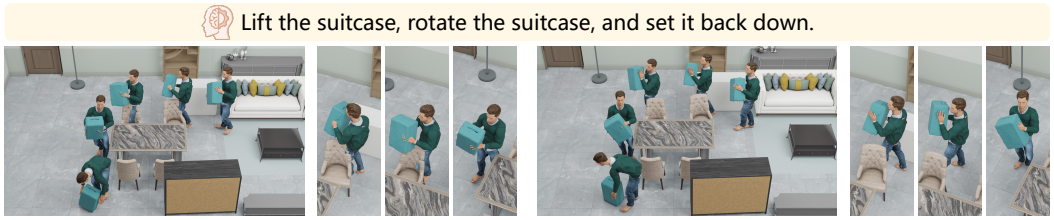
Figure 2: Qualitative comparison with InfBaGel. Each row shows the text instruction and motion snapshots for InfBaGel (left) and PriorHOSI (right), with an additional view of each interaction.

Qualitative comparison. Figure 2 illustrates how PriorHOSI coordinates body posture and object placement near scene obstacles. In the small-table task (panels a–b), the person must lift an edge, rotate and drag the table, and lower its legs to the floor. InfBaGel places the moving table partly within the bed region in the highlighted snapshots, whereas PriorHOSI keeps it farther from the bed as the person bends and turns with the table. In the box task (panels c–d), InfBaGel's body overlaps a dining chair while carrying the box around the round table; PriorHOSI positions the body farther from the chair with the box remaining in front of the hands. The additional views clarify these spatial relationships, illustrating coordinated changes in body posture and object placement near furniture.

4.3 Ablation Study

Do scene guidance and contact constraints improve interaction quality? Scene guidance improves scene compatibility, while contact constraints preserve human–object coordination (Table 2). Relative to direct outputs from HOIPrior, scene guidance alone reduces HS/OS P_{mean} from 6.48/25.84 to 3.52/18.73, but lowers contact rate from 89.06% to 75.82%. Adding contact constraints recovers contact rate to 87.54%, further reduces HS/OS P_{mean} to 2.84/13.75, and raises task success from 81.66% to 83.16%, with a slight increase in HS penetrating frames (23.78% to 24.52%). Figure 3 illustrates the resulting coordination: noise optimization adjusts body posture and suitcase placement to improve clearance from a nearby chair while maintaining hand–object alignment.

Do prior-specific designs improve interaction quality? Table 3 evaluates the two prior-specific designs with both terms active in the noise optimization criterion. Adding $\mathcal{L}_{\text{rel}}$ during HOIPrior



(a) HOIPrior (w/o noise optimization) (b) PriorHOSI

Figure 3: Qualitative ablation of noise optimization. Compared with the direct output of HOIPrior (a), PriorHOSI (b) improves scene compatibility while preserving human–object coordination.

Table 2: Ablation of scene guidance and contact constraints in noise optimization. Scene guidance reduces scene penetration, while contact constraints recover human–object contact. The first row reports direct HOIPrior outputs; – indicates that contact constraints are inactive without scene guidance.

Noise optimization terms		Task accuracy			FS↓	HO interaction		HS penetration ↓			OS penetration ↓		
$\mathcal{L}_{\text{scene}}$	$\mathcal{L}_{\text{contact}}$	$T_h \downarrow$	$T_o \downarrow$	$S\% \uparrow$		$C\% \uparrow$	$P_{body} \downarrow$	P_{mean}	P_{max}	$P_f\%$	P_{mean}	P_{max}	$P_f\%$
×	–	3.48	6.54	80.60	0.14	89.06	2.86	6.48	61.82	31.26	25.84	148.36	28.71
✓	×	3.04	6.37	81.66	0.12	75.82	5.26	3.52	41.86	23.78	18.73	126.58	25.34
✓	✓	3.13	6.05	83.16	0.10	87.54	3.17	2.84	34.38	24.52	13.75	94.11	21.62

Table 3: Ablation of prior-specific designs. Relation supervision improves human–object contact, while residual refinement reduces human–scene penetration. $\widetilde{\mathcal{M}}_h$ and $\widehat{\mathcal{M}}_h$ denote scene guidance targets before and after refinement.

HOIPrior	HSIPrior	HO interaction		HS penetration ↓		OS penetration ↓		
Relation supervision	Scene guidance target	$C\% \uparrow$	$P_{body} \downarrow$	P_{mean}	$P_f\%$	P_{mean}	P_{max}	$P_f\%$
w/o $\mathcal{L}_{\text{rel}}$	$\widehat{\mathcal{M}}_h$	82.76	4.09	3.63	24.18	13.42	99.86	21.72
w/ $\mathcal{L}_{\text{rel}}$	$\widetilde{\mathcal{M}}_h$	88.12	3.09	4.46	26.13	13.68	102.34	21.87
w/ $\mathcal{L}_{\text{rel}}$	$\widehat{\mathcal{M}}_h$	87.54	3.17	2.84	24.52	13.75	94.11	21.62

distillation raises contact rate from 82.76% to 87.54% and reduces body–object penetration from 4.09 to 3.17, demonstrating the effectiveness of relation supervision in preserving human–object contact. Applying residual refinement during HSIPrior inference lowers HS P_{mean} by 36.3% (4.46 to 2.84), with a slight contact-rate decrease (88.12% to 87.54%). This highlights the value of refining the scene guidance target at inference time.

5 Conclusion

We present PriorHOSI, a framework that composes separately learned HOIPrior and HSIPrior for human–object–scene interaction generation. The key idea is to use scene-adapted human motion to guide joint human–object updates through diffusion noise optimization, while object-relative contact constraints preserve the local interaction geometry. Relation-aware consistency distillation supports efficient repeated sampling, and body-part-aware residual refinement improves the scene guidance targets. Experiments on the InfBaGel benchmark show higher task success and hand–object contact rates, along with reduced human–scene penetration. These results demonstrate a practical approach to combining the complementary supervision available in existing interaction datasets.

This composition currently operates in an offline setting and retains contact references extracted from the initial interaction. Its scene guidance acts directly on human motion, while object–scene compatibility is addressed indirectly through the coupled generator and contact constraints. Improving the efficiency of iterative editing and incorporating direct object–scene guidance are natural next steps toward broadening the framework's applicability.

AI USE STATEMENT

In this work, we used generative AI tools to help identify research topics and potential gaps in the literature and to polish the writing of this paper, including grammar, word choice, and sentence fluency. All AI-assisted text was manually reviewed and edited by the authors. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Jungbin Cho, Minsu Kim, Jisoo Kim, Ce Zheng, Laszlo A. Jeni, Ming-Hsuan Yang, Youngjae Yu, and Seonjoo Kim. SceneAdapt: Scene-aware Adaptation of Human Motion Diffusion. *arXiv preprint arXiv:2510.13044*, 2025. URL https://arxiv.org/abs/2510.13044.

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 (CVPR)*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. URL https://arxiv.org/abs/2010.11929.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Mian. UniHM: Universal Human Motion Generation with Object Interactions in Indoor Scenes. *arXiv preprint arXiv:2505.12774*, 2025. URL https://arxiv.org/abs/2505.12774.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic Scene-Aware Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, 2024a. URL https://arxiv.org/abs/2410.03187.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling Up Dynamic Human-Scene Interaction Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

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*, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing Diffusion Noise Can Serve As Universal Motion Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object Motion Guided Human Motion Synthesis. *ACM Transactions on Graphics (TOG)*, 42:1–11, 2023. URL https://arxiv.org/abs/2309.16237.

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 (ECCV)*, 2024. URL https://arxiv.org/abs/2312.03913.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional Visual Generation with Composable Diffusion Models. In *European Conference on Computer Vision (ECCV)*, pp. 423–439, 2022. URL https://arxiv.org/abs/2206.01714.

Xinpeng Liu, Haowen Hou, Yanchao Yang, Yong-Lu Li, and Cewu Lu. Revisit Human-Scene Interaction via Space Occupancy. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2312.02700.

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 (CVPR)*, 2019.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient Learning on Point Clouds With Basis Point Sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Roey Ron, Guy Tevet, Haim Sawdayee, and Amit H. Bermano. HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization. *arXiv preprint arXiv:2506.15625*, 2025. URL https://arxiv.org/abs/2506.15625.

Pablo Ruiz-Ponce, Germán Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. MixerMDM: Learnable Composition of Human Motion Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12380–12390, 2025.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human Motion Diffusion as a Generative Prior. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2303.01418.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.02502.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. In *International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

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 (ECCV)*, 2020. URL https://arxiv.org/abs/2008.11200.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14916.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (ICCV)*, 2023.

Wei Yao, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. HOSIG: Full-Body Human-Object-Scene Interaction Generation with Hierarchical Scene Perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. doi: 10.1609/aaai.v40i14.38177. URL https://doi.org/10.1609/aaai.v40i14.38177.

Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional Human-Scene Interaction Synthesis with Semantic Control. In *European conference on computer vision (ECCV)*, 2022. URL https://arxiv.org/abs/2207.12824.

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing Diverse Human Motions in 3D Indoor Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yude Zou, Junji Gong, Xing Gao, Zixuan Li, Tianxing Chen, and Guanjie Zheng. InfBaGel: Human-object-scene interaction generation with dynamic perception and iterative refinement. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2604.04843.

A Interaction Data and Evaluation Protocol

A.1 Coverage of Existing Interaction Data

Learning HOSI requires synchronized human and object motion together with scene geometry and task descriptions. The central data challenge is coverage: a given interaction can require different body configurations and object trajectories as the scene layout changes. Capturing each object and action in isolation therefore provides only part of the supervision needed for scene-dependent adaptation. Existing datasets provide complementary parts of this supervision. OMOMO (Li et al., 2023) captures approximately ten hours of whole-body interaction with 15 objects, including boxes, tables, chairs, and suitcases. It provides tracked object motion and geometry, but lacks the surrounding scene annotations needed to learn scene-dependent adaptation directly.

Scene-oriented capture expands environmental coverage, with different emphases on action and object supervision. TRUMANS (Jiang et al., 2024b) provides more than 15 hours of interaction in 100 indoor scenes, including object dynamics and frame-wise action labels. LINGO (Jiang et al., 2024a) provides 16 hours in 120 scenes, with language annotations spanning 40 motion types, including locomotion, reaching, and scene interaction. These resources support learning how human motion depends on the environment, while OMOMO supplies complementary demonstrations of whole-body interactions with large movable objects. This complementary coverage motivates learning specialized priors and coordinating their outputs during inference.

Synthetic scene augmentation. InfBaGel (Zou et al., 2026) converts OMOMO motions into pseudo-HOSI examples using synthetic occupancy, following the occupancy-based data construction of Liu et al. (2024). For each interaction, it computes the volume swept by the human and object, encloses that volume in a bounding box, and places occupied voxels outside the box. The resulting scene is compatible with the recorded motion and allows the sample to be expressed as a scene–object–instruction–goal tuple. This construction provides spatial context consistent with the recorded trajectory. Learning how an interaction responds to a new obstacle arrangement requires additional scene-dependent motion supervision, which motivates our use of LINGO through HSIPrior.

Motivation for separate priors. InfBaGel combines these augmented HOI examples with LINGO to learn human–object and human–scene interactions in one model. Its hybrid-data ablation reports a balance between the two: adding LINGO at an OMOMO-to-LINGO ratio of $1{:}0.5$ reduces several scene-penetration measures, while task success changes from 83.16% to 81.45% and contact rate from 78.18% to 76.96% (Zou et al., 2026). This observation motivates our choice to learn the priors separately and combine their predictions during editing. The inference objective then couples scene adaptation to explicit constraints on hand–object coordination.

A.2 Training Data and the InfBaGel Benchmark

Training data. We use the augmented OMOMO and processed LINGO data released by InfBaGel (Zou et al., 2026), with its data splits, motion processing, text annotations, and goal construction. HOIPrior is trained on the human and object motion in augmented OMOMO, conditioned on object geometry, text, and goal locations. Its input excludes the synthetic scene fields. HSIPrior is trained on LINGO human motion with scene geometry, text, and the human goal; object geometry and object-goal conditions are inactive.

Benchmark construction. We use the released InfBaGel benchmark, which comprises 469 tasks with seven categories of movable objects in 67 indoor scenes from TRUMANS (Jiang et al., 2024b; Zou et al., 2026). The benchmark combines initial interactions and instructions from OMOMO-test with start and goal locations in the scenes. Its construction checks collision-free initial and goal configurations and a traversable passage between them; the interaction may still require a detour around obstacles. The tasks include lifting, dragging, kicking, and rotating objects. All evaluations use the tasks in this released test set.

Baseline adaptations. TRUMANS and LINGO are adapted to the HOSI interface following InfBaGel (Zou et al., 2026). TRUMANS receives text, object geometry, and goal conditions, with addi-

Table 4: Conditions supplied to the two motion priors. A checkmark denotes an active condition.

Condition	HOIPrior	HSIPrior
Text instruction	✓	✓
Human pelvis goal	✓	✓
Object centroid goal	✓	–
Object geometry	✓	–
Scene occupancy	–	✓

tional heads for object motion. LINGO receives object geometry and object-goal inputs and predicts object translation and rotation alongside human motion. CHOIS and HOSIG use their native motion representations with scene-derived waypoints for goal conditioning on the Hybrid Dataset. CHOIS is trained from scratch; HOSIG's control branch is fine-tuned while its pretrained motion and grasp priors remain frozen.

Evaluation metrics. We follow InfBaGel's HOSI protocol. Human and object goal errors, T_h and T_o, measure final distances to the prescribed goals, and $S\%$ is the fraction of tasks for which both distances are below $10\,\mathrm{cm}$. Foot sliding (FS), contact percentage ($C\%$), and body–object penetration (P_{body}) measure motion and interaction quality. For human–scene and object–scene penetration, the benchmark first sums penetration depths over mesh vertices within each frame, then computes the temporal mean (P_{mean}) and maximum (P_{max}) of these sums. These aggregate depth measures are reported in meters. $P_f\%$ is the percentage of frames with penetration. InfBaGel also uses summed vertex penetration depths, in meters, for P_{body} (Zou et al., 2026).

B Architecture and Condition Encoders

Each prior receives conditions matched to its task, as summarized in Table 4. The relation-aware distillation objective for HOIPrior and the residual refinement module for HSIPrior are the prior-specific designs described in Section 3.

Motion representation and coordinates. The human state at each frame contains the 3D root translation and 6D rotations of 22 body joints, giving $3 + 22 \times 6 = 135$ channels. The object state contains three centroid coordinates and nine rotation-matrix entries, giving 147 channels for coupled HOI motion. Forward kinematics yields 28 joint positions from these parameters: the 22 body joints and two additional points on each hand and the head. Geometric losses and scene queries use the joint positions specified below. Motion and goals are expressed in a local frame centered at the initial pelvis on the ground plane and aligned with its heading. Position channels are scaled to $[-1, 1]$ using training-set bounds. Geometry losses use decoded world coordinates in meters, with forward kinematics (FK) computed from the pelvis position, joint rotations, and rest-pose skeleton. When transferring motion between priors, we decode it and apply the receiving prior's normalization.

Motion backbone. The backbone and condition encoders are based on InfBaGel (Zou et al., 2026). An MLP embeds each motion frame into a 512-dimensional token. Motion and condition tokens are processed by an eight-layer Transformer encoder with 16 attention heads, and an output MLP predicts the clean motion. Diffusion-time embeddings are added to the condition tokens, and positional encodings preserve temporal order. The feed-forward width is 2048 and dropout is 0.1. HOIPrior predicts human and object motion jointly, while HSIPrior predicts human motion.

Text encoder. The instruction is encoded by CLIP (Radford et al., 2021) into a 768-dimensional feature. A text MLP maps this feature to a 512-dimensional condition token. Both priors use this encoder so that the interaction and its scene refinement are conditioned on the same instruction.

Goal encoders. Separate MLPs map the human pelvis goal and object centroid goal to 512-dimensional embeddings. The human encoder takes the two horizontal coordinates of the pelvis goal, leaving the motion model to determine body height from the action and scene. The object encoder takes all three coordinates of the target centroid. HOIPrior receives both goals; HSIPrior receives the human goal.

Object geometry encoder. The basis point set (BPS) representation (Prokudin et al., 2019) encodes the centered reference object as 1,024 vectors from fixed basis points to the object surface. An MLP projects this 1024×3 representation to a 512-dimensional object token for HOIPrior. Object geometry is omitted from HSIPrior.

Scene encoder. HSIPrior encodes scene context using five local occupancy grids: two cover the start and goal regions, and three follow pelvis locations sampled along the motion. Each grid contains 32^3 voxels in a cube of side length $1.2\,\mathrm{m}$, giving a voxel width of $3.75\,\mathrm{cm}$. A shared vision Transformer (Dosovitskiy et al., 2020) with six layers and 16 attention heads encodes each grid into a 512-dimensional feature. The feature is fused with the positional encoding of the grid center, preserving the spatial location of the observed geometry. The five tokens provide occupancy context to the first-stage denoiser. During editing, the three intermediate crop centers follow the current HOI candidate. The residual module adds SDF observations around the predicted body, described in Appendix C.2. HOIPrior receives no scene tokens.

Motion refinement module. The residual module encodes body-aligned SDF observations, constructed as described in Appendix C.2, to refine the first-stage motion prediction. A shared linear encoder and learned joint embeddings produce 48-dimensional features, pooled separately over the torso, both arms, and both legs. Each group uses layer normalization, a linear layer, and SiLU. The residual MLP receives the five group features, their first temporal differences (zero at the first frame), the pelvis distance, the first-stage latent, and a 16-dimensional sinusoidal frame encoding. A hidden layer of width 256 with SiLU and a zero-initialized output layer predict the 135-channel motion correction. The correction is added in the normalized motion space and decoded for geometric supervision.

C Training and Inference Details

All training stages use four NVIDIA RTX 3090 GPUs and Adam with moment coefficients $(0.9, 0.999)$ and zero weight decay. The learning rates are 10^{-4} for HOIPrior and 2×10^{-4} for HSIPrior. We denote the set of generated frames by $\mathcal{F}$.

C.1 HOIPrior

Teacher objective. The forward process is

$$x_\tau = \sqrt{\bar{\alpha}_\tau}\, x_0 + \sqrt{1-\bar{\alpha}_\tau}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{12}$$

where $\bar{\alpha}_\tau = \prod_{s=0}^{\tau}(1-\beta_s)$, with 500 levels and linearly spaced β_s from 10^{-4} to 0.02.

We sample τ uniformly and train the clean-motion prediction $\widetilde{x}_0 = f_\psi(x_\tau, \tau; c_O)$ with

$$\mathcal{L}_{\text{teacher}} = \mathbb{E}[\mathrm{MSE}_{\mathcal{F}}(\widetilde{x}_0, x_0) + 50\mathcal{L}_{\mathrm{FK}} + 50\mathcal{L}_{\mathrm{obj}}], \tag{13}$$

where $\mathrm{MSE}_{\mathcal{F}}$ averages over predicted frames and channels. $\mathcal{L}_{\mathrm{FK}}$ is the mean squared error between decoded and ground-truth joint positions, and $\mathcal{L}_{\mathrm{obj}}$ is the mean squared error between corresponding object surface points transformed by the predicted and recorded poses. The human term uses 28 FK points: 22 body joints and two additional points on each hand and the head. The object term uses 1,024 fixed surface points. Both geometric losses use positions in meters and average over frames, points, and coordinates. The teacher is trained for 500 epochs with batch size 512.

Consistency and relation distillation. We use the 50-level grid $u_n = \mathrm{round}(499n/49)$ for $n = 0, \ldots, 49$, with $u_{-1} = -1$ and $\bar{\alpha}_{-1} = 1$. At a sampled noise level, one DDIM teacher step constructs the adjacent state for the consistency loss in Equation 2. We use the boundary parameterization $f_\theta(x, \tau) = c_{\mathrm{skip}} x + c_{\mathrm{out}} h_\theta(x, \tau)$, where $c_{\mathrm{skip}} = \sigma_{\mathrm{data}}^2/(\widetilde{\tau}^2 + \sigma_{\mathrm{data}}^2)$, $c_{\mathrm{out}} = \widetilde{\tau}/\sqrt{\widetilde{\tau}^2 + \sigma_{\mathrm{data}}^2}$, $\widetilde{\tau} = 10(\tau+1)$, and $\sigma_{\mathrm{data}} = 0.5$. This gives $f_\theta(x, -1) = x$ at the clean boundary. The student is initialized from its trained HOI teacher, and the target network from the student. Teacher and target predictions are detached; the target weights are updated as $\bar{\theta} \leftarrow 0.95\bar{\theta} + 0.05\theta$.

For relation supervision, the frozen teacher generates a complete motion with 50 DDIM steps, and the student uses the 16-step sampler with the same initial noise and conditions. Teacher motions

are generated online and detached; gradients traverse the complete student trajectory. Each update combines a consistency batch and a relation batch with $\lambda_r = 1$. We distill for 200 epochs with an effective batch size of 512 and use the final EMA weights for HOI sampling. Let $\mathcal{J}_{\text{hand}}$ index the four contact query points: the wrist and the base of the middle finger on each hand. The teacher contact set is

$$\mathcal{C}^{\psi} = \left\{ (t,j) \in \mathcal{F} \times \mathcal{J}_{\text{hand}} : \min_{v \in \mathcal{O}_t^{\psi}} \| \operatorname{FK}(\mathcal{M}_h^{\psi})_{t,j} - v \|_2 \le 0.03\,\text{m} \right\}, \tag{14}$$

where $\mathcal{O}_t^{\psi}$ denotes the transformed object vertices. Relations use the decoded object frame and are averaged over contacting pairs and then over the batch.

Few-step sampling and decoding. Starting from $x_{\tau_0} = z$, the student uses $\tau_s = u_{49-\lfloor 50s/16 \rfloor}$, $s = 0, \ldots, 15$, and

$$y_s = f_\theta(x_{\tau_s}, \tau_s; c_O), \qquad x_{\tau_{s+1}} = \sqrt{\bar{\alpha}_{\tau_{s+1}}}\, y_s + \sqrt{1 - \bar{\alpha}_{\tau_{s+1}}}\, \xi_s \quad (s < 15). \tag{15}$$

The final prediction y_{15} is decoded to motion, defining the sampler G_θ used in Section 3.4. The 15 Gaussian tensors ξ_s are drawn once per sample and held fixed throughout noise optimization, making the sampler deterministic with respect to z. All 16 evaluations remain differentiable. Dropout is disabled during relation supervision and inference-time editing.

For geometric queries, the nine object-rotation channels are decoded from the first two matrix columns a_1, a_2 as

$$b_1 = \frac{a_1}{\|a_1\|_2}, \quad b_2 = \frac{a_2 - (b_1^\top a_2) b_1}{\|a_2 - (b_1^\top a_2) b_1\|_2}, \quad R_o = [b_1, b_2, b_1 \times b_2]. \tag{16}$$

This differentiable conversion gives a proper rotation shared by object vertices and object-relative hand relations.

C.2 HSIPRIOR

Local scene geometry. The residual module queries a scene SDF sampled on a world-aligned grid with spacing $h = 0.02\,\text{m}$. Trilinear interpolation provides continuous values at the predicted joint locations. The field represents the static environment; the query locations change with the motion.

Body-aligned SDF queries. We query 24 FK points, comprising the 22 body joints and one additional point on each hand. At each first-stage prediction $q_{t,j}$, we sample the SDF and its six axis-aligned neighbors. Central differences with spacing h give $\widetilde{\nabla} d_{\mathcal{S}}$, and the observation is

$$\varphi_{t,j} = \left[v_{t,j}\, d_{\mathcal{S}}(q_{t,j})/a, \quad v_{t,j} A^\top \widetilde{\nabla} d_{\mathcal{S}}(q_{t,j}), \quad v_{t,j} \right], \qquad a = 0.25\,\text{m}, \tag{17}$$

where A rotates the local motion frame into world coordinates and $v_{t,j}$ indicates that all seven locations lie within the SDF domain. These observations are encoded by the motion refinement module described in Appendix B.

Joint training objective. We first train the HSI denoiser from scratch on LINGO, then add the residual module and jointly optimize both stages on the refined prediction. Using the same 500-level corruption schedule as HOIPrior, we minimize

$$\mathcal{L}_{\text{HSI}} = \mathbb{E}\left[\underbrace{\text{MSE}_{\mathcal{F}}(\widehat{x}_h, x_h^*) + 50\mathcal{L}_{\text{FK}}}_{\mathcal{L}_{\text{motion}}} + \mathcal{L}_{\text{geo}} \right], \tag{18}$$

where $\widehat{x}_h$ and x_h^* contain all 135 normalized human channels, and $\mathcal{L}_{\text{FK}}$ is defined as above. First-stage pretraining uses $\mathcal{L}_{\text{motion}}$ on its own prediction for 250 epochs, followed by 150 epochs of joint training, both with effective batch size 512. At inference time, scene guidance uses the final jointly trained weights. During joint training, gradients reach the first stage through the residual sum, its latent features, and differentiable SDF queries. Geometry supervision uses predicted-frame query points whose ground-truth SDF magnitude is at most $a = 0.25\,\text{m}$, with both ground-truth and predicted points within the SDF domain. This defines $\mathcal{V}$ in Equation 6. The scaled smooth-L_1 penalty is $\ell_{\text{sL1}} = e^2/2$ for $|e| < 1$ and $|e| - 1/2$ otherwise, with $e = (d_1 - d_2)/a$. Active-set membership is fixed during each gradient evaluation.

C.3 Inference-Time Prior Composition

Scene guidance. At editing iteration k, the current human motion is normalized for HSIPrior and corrupted by Equation 12 at $\tau_S = 199$, with fresh Gaussian noise. A single two-stage denoising pass provides the scene target, which is decoded to world coordinates and detached. Equation 10 averages over $\mathcal{I} = \mathcal{F} \times \mathcal{J}_{24}$, including the wrists and additional hand points. The scene gradient traverses the coupled HOI sampler; both priors' parameters remain frozen.

Contact constraints. We construct $\mathcal{C}$ from the initial motion $\mathcal{M}^{(0)}$ using the same four query points $\mathcal{J}_{\text{hand}}$ and the $0.03\,\text{m}$ distance threshold. For each point, we retain only contacts lasting at least three consecutive frames. The contact set $\mathcal{C}$ and reference relations $r_i^{(0)}$ remain fixed throughout editing. With displacement tolerance $\delta = 0.01\,\text{m}$ and $c_i(\mathcal{M}) = \|r_i(\mathcal{M}) - r_i^{(0)}\|_2 - \delta$, the contact term is the motion-dependent part of

$$\mathcal{A}(\mathcal{M}; \mu, \rho) = \sum_{i \in \mathcal{C}} \frac{[\mu_i + \rho c_i(\mathcal{M})]_+^2 - \mu_i^2}{2\rho}, \qquad [b]_+ = \max(b, 0). \tag{19}$$

The $-\mu_i^2/(2\rho)$ terms are constant during a noise update, so the motion-dependent form in Equation 11 has the same gradient. Starting from $\mu_i = 0$ and $\rho = 1$, every five completed noise updates we set

$$\mu_i \leftarrow [\mu_i + \rho c_i(\mathcal{M}^{(k)})]_+. \tag{20}$$

Every 25 updates, we evaluate $v_k = \max_{i \in \mathcal{C}} [c_i(\mathcal{M}^{(k)})]_+$ and double ρ (up to 100) if the violation exceeds $0.75 v_{\text{prev}}$. The first check at $k = 25$ establishes v_{prev}; subsequent checks compare with the preceding check and refresh this reference. When both updates are due, multipliers use the current ρ before its adjustment. An empty contact set contributes zero penalty.

Noise optimization. We optimize the initial noise for $K = 150$ iterations using normalized-gradient Adam with moment coefficients $(0.9, 0.999)$ and 10^{-8} for both Adam's numerical constant and gradient normalization. The learning rate increases linearly to 0.05 over the first 25 updates, then follows a cosine decay over the remaining optimization budget. Each iteration refreshes the HSI target and differentiates the editing objective through all 16 HOI evaluations, with sampling perturbations, contact references, and the current multipliers held fixed. After the last noise update, we generate $\mathcal{M}^{(K)}$ with the same sampler.

D HOI EXPERIMENTS

Evaluation setup. We evaluate HOIPrior on OMOMO to assess the quality of the human–object motions used to initialize prior composition. Table 5 reports results for HOIPrior with 16-step consistency sampling and published results for LINGO, CHOIS, ROG, and InfBaGel, reproduced from Zou et al. (2026). Under this HOI protocol, LINGO and InfBaGel operate in an empty scene, while CHOIS and ROG are adapted to human waypoints and an object goal.

Evaluation metrics. We evaluate condition matching, motion quality, interaction quality, and agreement with ground-truth motion. Object endpoint error T_e and human waypoint error T_{xy} measure condition matching, while foot sliding (FS) measures motion quality. Interaction metrics comprise contact precision, recall, and F1-score (C_{prec}, C_{rec}, C_{f1}), contact fraction ($C\%$), and body–object penetration (P_{body}). Agreement with ground truth is measured by mean per-joint position error (MPJPE), root translation error T_{root}, and object translation and rotation errors (T_{obj}, O_{obj}).

Table 5: HOI comparison on OMOMO. Reference results and baseline timings follow Zou et al. (2026). Position errors are in cm and P_{body} is in m; contact metrics, including $C\%$, are fractions. The subscript in InfBaGel$_{16}$ denotes its sampling steps.

Method	Condition matching ↓		FS↓	Interaction					Ground-truth error ↓				FPS↑
	T_e	T_{xy}		C_{prec} ↑	C_{rec} ↑	C_{f1} ↑	$C\%$ ↑	P_{body} ↓	MPJPE	T_{root}	T_{obj}	O_{obj}	
Real motion	0.00	0.00	0.26	–	–	–	0.66	1.29	0.00	0.00	0.00	0.00	–
LINGO	17.57	3.76	0.50	0.70	0.46	0.50	0.41	4.80	16.39	10.28	27.64	1.33	28.86
CHOIS	5.28	5.73	**0.32**	0.78	0.63	0.66	0.53	2.52	13.93	8.42	17.74	**0.92**	28.01
ROG	26.68	19.43	0.49	0.80	0.68	0.70	0.54	5.86	13.37	21.28	31.89	1.14	9.07
InfBaGel$_{16}$	3.06	3.70	**0.32**	0.78	**0.73**	**0.73**	**0.60**	**2.49**	12.11	7.93	15.93	1.02	29.31
HOIPrior	**2.71**	**3.55**	0.35	**0.81**	0.65	0.68	0.52	2.75	**11.20**	**7.88**	**14.05**	0.99	**482.82**

Results. HOIPrior combines accurate goal matching with efficient motion generation. It achieves the lowest condition-matching and positional errors among the generated-motion methods in Table 5, reducing object endpoint error from 3.06 to 2.71 cm and MPJPE from 12.11 to 11.20 cm relative to InfBaGel$_{16}$. It also achieves the highest contact precision (0.81), with a trade-off in contact recall, F1-score, contact fraction, and body–object penetration relative to InfBaGel$_{16}$. With 16-step consistency sampling, HOIPrior reaches 482.82 FPS in our evaluation, which supports efficient repeated sampling during inference-time prior composition.