FORCEHOSI: FORCE-AWARE HUMAN–OBJECT–SCENE INTERACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human–object–scene interaction generation requires motion that reaches the task goal, fits the scene, and expresses the physical demands of manipulating the object. Changing an object's weight or resistance can require different postures even when its geometry and the task remain unchanged. We present ForceHOSI, a force-aware framework for generating scene-conditioned interactions with load-dependent and resistance-dependent human motion. Starting from a task-motion estimate, our method models postural adaptation as a continuous response field along a force coordinate. Action, interaction type, and force jointly modulate the field's temporal weights through a tensor-conditioned low-rank parameterization. Task-preserving composition combines the pose response with retained support channels and manipulation refinement. We construct FORCE-Scene by grounding recorded force-annotated interactions in 3D scenes, checking ground support and penetration, and forming strategy-consistent response pairs. Under the paired-window protocol, ForceHOSI achieves the highest response-norm ordering and the lowest object endpoint error and object–scene penetration aggregate among the compared methods. On 600 sampled test force-condition groups, its posture-response descriptor passes the monotonicity check in 91.7% of groups, compared with 50.5% for InfBaGel.

1 INTRODUCTION

Moving the same object to the same destination can require different postures depending on its weight or resistance (Zhang et al., 2025). In a 3D scene, these physical responses must also preserve foot support, avoid obstacles, and accomplish the task. Human–object–scene interaction generation therefore requires both geometric compatibility and posture consistent with physical demand.

Recent interaction generators coordinate human and object motion using language, object geometry, trajectories, goals, and scene context (Li et al., 2023; 2024; Yao et al., 2026; Zou et al., 2026). FORCE (Zhang et al., 2025) studies how human motion responds to different object loads and resistances. Bringing these responses into scene-conditioned generation requires satisfying both physical and spatial demands. A motion can reach the specified goal while failing to express the effort associated with the load. Changing the pose to match a recorded response can, in turn, disturb hand–object alignment or foot support. We aim to retain the task motion's spatial structure while adapting human posture to the requested physical condition.

We introduce ForceHOSI, which integrates a learned physical response into scene-conditioned interaction generation. Starting from a task-motion estimate, it models changes with force through a continuous response field. Action, interaction type, and force jointly modulate the field's temporal weights through a tensor-conditioned low-rank parameterization. At each integration step, the field receives the current motion state and frame phase, allowing the response to vary with body configuration and the stage of the task. Integration from a reference condition produces a force-conditioned estimate. Task-preserving composition then combines the adapted pose with retained support channels and manipulation refinement, coordinating the human and object while preserving selected task components.

We construct FORCE-Scene by placing FORCE captures in existing scenes, checking ground support and penetration. Response donors use compatible manipulation strategies, and transferred targets are checked on nominal carriers, retaining annotation and placement provenance.

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.

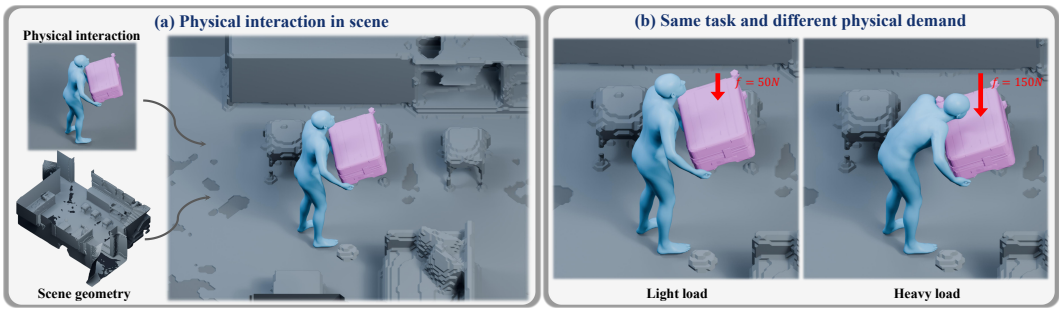

Figure 1: (a) Physical human–object interactions are grounded in 3D scenes. (b) For the same task and scene, different physical demands lead to distinct human responses, such as the posture changes induced by a light or heavy load.

Our evaluation examines force-response fidelity alongside task-goal accuracy and scene-interaction quality on the held-out test split. Paired-window comparisons measure pose reconstruction and interaction quality, while a separate group-level test examines monotonic posture changes across force conditions. Ablations assess the contributions of weight modulation, force integration, response supervision, and output composition. An added-obstacle example shows different execution choices under two loads. It illustrates the complete system's behavior without isolating the response field as the cause of route changes.

Our contributions are threefold:

- We propose a continuous force-response model with action-factorized weight modulation that adapts human posture to physical demand within scene-conditioned interaction generation.
- We introduce a task-preserving composition mechanism that combines support-channel retention with manipulation refinement, maintaining the learned pose response while coordinating human and object motion.
- We construct FORCE-Scene, a scene-grounded dataset that integrates force-annotated human–object motions with compatible 3D scenes for training and evaluating force-aware interaction generation. We will publicly release FORCE-Scene.

2 Related Work

Human–object interaction and physical response. Denoising diffusion (Ho et al., 2020) underlies motion generators such as MDM (Tevet et al., 2023). OMOMO (Li et al., 2023) conditions human motion on a complete object trajectory; CHOIS (Li et al., 2024) jointly generates human and object motion from language and sparse waypoints. InterDiff (Xu et al., 2023) corrects interaction predictions using contact structure, while InterDreamer (Xu et al., 2024) combines pretrained models with a learned interaction world model. PhysDiff (Yuan et al., 2023) projects motions through simulation during diffusion; PhysHOI (Wang et al., 2023) learns physics-based interaction imitation with contact-graph rewards. FORCE (Zhang et al., 2025) supplies recorded interactions under varying physical conditions. We use this variation to learn load- and resistance-dependent kinematic responses within scene tasks, without explicitly solving contact forces or rigid-body dynamics. Methods receiving complete object trajectories require a separate evaluation protocol.

Scene context and interaction data. PROX (Hassan et al., 2019) introduces scene contact and nonpenetration constraints for reconstruction; POSA (Hassan et al., 2021b) learns body–scene contact relationships; COINS (Zhao et al., 2022) supports compositional interaction semantics. SAMP (Hassan et al., 2021a) and DIMOS (Zhao et al., 2023) synthesize navigation and scene interactions, while LINGO (Jiang et al., 2024) integrates text-driven character behavior. HOSIG (Yao et al., 2026) uses hierarchical scene perception, and InfBaGel (Zou et al., 2026) combines dynamic perception with iterative refinement. AMASS (Mahmood et al., 2019) unifies motion captures; GRAB (Taheri et al., 2020), BEHAVE (Bhatnagar et al., 2022), and InterCap (Huang et al., 2022)

capture complementary whole-body object interactions. FORCE-Scene instead grounds FORCE recordings in constructed scenes, adding spatial context and strategy-consistent response pairs rather than new behavioral observations.

Conditional modulation and continuous transformations. Hypernetworks (Ha et al., 2017) generate conditional parameters, LoRA (Hu et al., 2022) factorizes weight updates, and FiLM (Perez et al., 2018) modulates features. Neural ODEs (Chen et al., 2018) parameterize continuous transformations using learned vector fields. Our field combines conditional low-rank weights with integration along normalized physical demand, rather than time, within iterative generation (Song et al., 2023). InfBaGel with force FiLM supplies a feature-modulation baseline for evaluating this response parameterization against direct force conditioning.

3 Force-Aware Interaction Generation

3.1 Task and representation

Let $\mathcal{S}$ denote scene geometry, $\mathcal{O}$ the manipulated object geometry, ℓ a task instruction, and $g = (g_h, g_o)$ the human and object goals. Given a motion prefix $X_{1:P}$, an action a, an interaction type p, and a physical condition f, our objective is to generate the remaining joint motion:

$$X_{P+1:T} \sim p_\Theta(X \mid \mathcal{S}, \mathcal{O}, \ell, g, X_{1:P}, a, p, f)\,. \tag{1}$$

The scalar f is expressed in newtons and denotes supported load for carrying or resistance for pushing and pulling. The type p specifies this interpretation and accompanies the scalar throughout the response model. Thus, f represents an interaction-level physical condition rather than a complete time-varying contact-force vector.

Each frame is represented by

$$x_t = [J_t, r_t, p_t^o, R_t^o, c_t] \in \mathbb{R}^{232}, \tag{2}$$

where $J_t \in \mathbb{R}^{84}$ contains 28 joint positions, $r_t \in \mathbb{R}^{132}$ contains 22 global joint rotations in a 6D representation (Zhou et al., 2019), $p_t^o \in \mathbb{R}^3$ is the object translation, $R_t^o \in \mathbb{R}^9$ encodes the relative object rotation, and $c_t \in \mathbb{R}^4$ contains contact channels. Human and object states use a consistent coordinate convention, allowing pose adaptation, hand–object alignment, and scene compatibility to be evaluated in the same spatial frame.

3.2 Continuous force-response field

Figure 2 summarizes the architecture. We build on the scene-conditioned generation framework of InfBaGel (Zou et al., 2026). At sampling step k, the task generator predicts a clean motion window,

$$\widehat{X}_k = G_\theta(Z_k, k, C_k), \tag{3}$$

where Z_k is the current sampling state and C_k contains language, goals, object geometry, and local scene context. A response field V_ψ then adapts this estimate from a reference physical condition to the requested force.

For interaction type p, we define the normalized force coordinate

$$u = \frac{f - f_{0,p}}{s_p}, \qquad s_p > 0, \tag{4}$$

where $f_{0,p}$ is the reference condition and s_p is a type-specific normalization scale. The response transformation $\mathcal{T}_\psi$ is defined by

$$\frac{dY(v)}{dv} = V_\psi(Y(v), v, \tau, a, p), \qquad Y(0) = \widehat{X}_k, \qquad \mathcal{T}_\psi(\widehat{X}_k, u) = Y(u). \tag{5}$$

Here, τ encodes frame phase within the task, whereas v indexes normalized physical demand. The field therefore describes how a motion changes across force conditions, not how a dynamical system evolves through time.

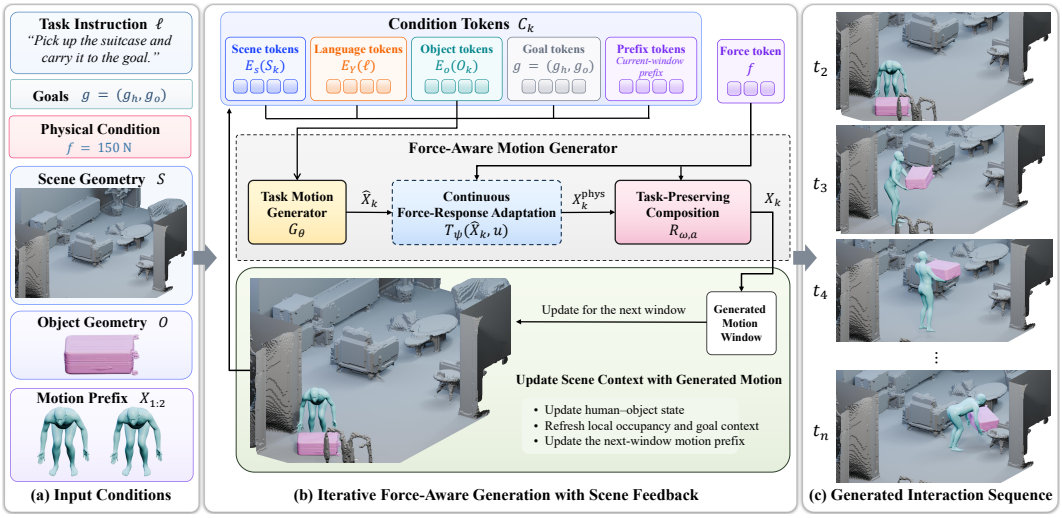

Figure 2: Overview of ForceHOSI. (a) Inputs specify the task, geometry, physical condition, and motion prefix. (b) A task motion generator, continuous force-response field, and task-preserving composition generate motion windows, which update the scene context for subsequent generation. (c) Successive windows form a force-aware human–object–scene interaction sequence.

We approximate the transformation using L explicit updates:

$$Y_{i+1} = Y_i + \Delta u\, V_\psi\big(Y_i, (i+\tfrac{1}{2})\Delta u, \tau, a, p\big)\,, \qquad \Delta u = u/L. \tag{6}$$

The reported model uses $L = 4$. Each update evaluates the current motion state at the midpoint of the corresponding force interval. At $f = f_{0,p}$, the interval vanishes, so $\mathcal{T}_\psi$ is exactly the identity. This property applies to the response transformation; the subsequent manipulation refiner may still modify the final output at the reference condition.

3.3 Action-factorized weight modulation

An equal increase in carried load and pushing resistance need not produce the same postural change. We therefore condition the response transformations jointly on an action embedding e_a, an interaction-type embedding e_p, and force features $b(v)$. For coefficient index q, a learned tensor $\mathcal{A}$ produces

$$z_q(a, p, v) = \sum_{i,j,m} \mathcal{A}_{ijmq}[e_a]_i[e_p]_j[b(v)]_m. \tag{7}$$

The force features include v, v^2, $\tanh v$, $\sin(\pi v)$, and $\cos(\pi v) - 1$, all centered such that $b(0) = 0$. A modulated linear transformation takes the form

$$h' = Wh + U\,\mathrm{diag}(z(a, p, v))Dh + b, \tag{8}$$

where $D \in \mathbb{R}^{r \times d_{\text{in}}}$ and $U \in \mathbb{R}^{d_{\text{out}} \times r}$ define an update of rank at most r. This parameterization is applied to attention and feed-forward projections in the temporal response Transformer.

The shared transformations capture common motion structure, while the low-rank updates adapt the temporal projections to the action and physical condition. The field also receives the evolving motion state and frame phase, so its response can vary with both body configuration and the stage of the task. Centering the force features removes the conditional weight update at the reference condition. The identity of the overall transport follows separately from the zero integration interval. This factorization does not enforce monotonic joint motion. Since interaction type is determined by action in this corpus, the type factor should not be interpreted as evidence of independently disentangled physical factors.

3.4 TASK-PRESERVING RESPONSE COMPOSITION

Force-dependent adaptation should modify the pose without unnecessarily disturbing the motion channels that maintain the task. For example, a carrying response can involve the trunk and arms while retaining lower-body support, whereas pushing and pulling impose different constraints through the hands and object. We use action-specific channel masks to control where the response enters the task motion.

At sampling step k, the masked response is

$$X_k^{\text{phys}} = \widehat{X}_k + \alpha_k M_a \odot \left[\mathcal{T}_\psi(\widehat{X}_k, u) - \widehat{X}_k\right], \tag{9}$$

where M_a selects the response channels and α_k controls their contribution during sampling. We use a late quadratic schedule with strength 0.35, reducing the influence of pose adaptation during early refinement.

For carrying, manipulation refinement is separated from the pose-response branch. A frozen pose carrier retains the learned physical response, while a trainable refiner adjusts the object and contact predictions. The carrier is evaluated on its own evolving state during force integration, so object corrections cannot recursively alter the state used to generate the pose response. This separation prevents the manipulation-refinement branch from directly overwriting the learned postural adaptation.

The output is composed with an action-specific task skip:

$$X^{\text{out}} = B_a \odot X^{\text{task}} + (1 - B_a) \odot \mathcal{R}_{\omega,a}(X^{\text{phys}}, f), \tag{10}$$

where X^{task} denotes the task-stream estimate, B_a selects the channels retained from that stream, and $\mathcal{R}_{\omega,a}$ denotes the action-specific refinement. Pulling also uses the task stream as a reference for subsequent motion adaptation. Appendix A summarizes the deployed action policies.

3.5 TRAINING AND SCENE-CONDITIONED GENERATION

Training uses pairs (X^0, X^f) that share a task context but differ in the target physical response. The response objective is

$$\mathcal{L}_{\text{resp}} = \mathcal{L}_{\text{pose}} + \lambda_R \mathcal{L}_{\text{rot}} + \lambda_v \mathcal{L}_{\text{vel}} + \lambda_d \mathcal{L}_{\text{dir}} + \lambda_m \mathcal{L}_{\text{mag}} + \lambda_q \mathcal{L}_{\text{preserve}}. \tag{11}$$

The terms supervise pose reconstruction, rotation agreement, temporal differences, response direction and magnitude, and preservation of non-pose channels, respectively. Direction and magnitude are defined relative to the nominal motion X^0, encouraging the model to recover the transferred response rather than only the absolute target pose. No explicit monotonicity loss is used in the reported configuration.

The manipulation refiner is trained separately on generated motions with the pose carrier frozen. This stage trains object and contact predictions while keeping the learned response-branch parameters fixed. Separating the two training stages retains the pose response during manipulation refinement. Appendix A describes the response objective and reports its loss weights.

During scene-conditioned generation, local occupancy and goal context are refreshed using the generated motion windows. Additional obstacles are inserted into the occupancy representation before sampling. The obstacle demonstrations use the same scene-conditioning mechanism together with the supplied passage constraints. When candidate routes are selected by a separate component, that selection is external to the learned response field and is distinguished from motion generation.

4 THE FORCE-SCENE DATASET

Learning force-aware scene interactions requires motion data that connects physical demand to human posture while providing compatible scene context. We construct FORCE-Scene by grounding force-annotated human–object motions from FORCE (Zhang et al., 2025) in scenes from the Inf-BaGel HOSI benchmark (Zou et al., 2026). The construction retains the physical annotations and relative human–object motion while introducing scene geometry, providing supervision for force-aware interaction generation under spatial constraints.

Table 1: **FORCE-Scene statistics.** Grounded sequences are accepted motion–scene combinations. Scene placements and overlapping windows are derived from the source captures and do not constitute additional independent recordings.

Source sequences	450	Benchmark scenes	67
Candidate combinations	30,150	Rejected combinations	18,653
Retained sequences	11,497	Frames per sequence	160
16-frame windows	137,964	Windows per sequence	12

Scene-compatible interaction grounding. We represent FORCE human motions using SMPL-X (Pavlakos et al., 2019) through the official model-transfer procedure, including articulated hands. Human and object motions are expressed in a common coordinate convention and checked for alignment before placement. Interaction-specific object offsets are held fixed across force conditions to avoid introducing alignment changes that could be mistaken for force-dependent responses.

Each human–object sequence is placed in a scene using a shared spatial transform, preserving the relative configuration of the interaction. We search for placements compatible with the scene geometry and retain those that pass checks for penetration, scene-boundary violations, and ground support. All sequences passing these checks undergo manual visual inspection as a final quality-control step. Each accepted sequence retains its physical annotations, source-motion and scene identifiers, and placement transform. The resulting dataset combines recorded force-dependent interactions with constructed scene context, allowing physical response and scene compatibility to be evaluated within the same interaction.

Dataset scale and organization. We combine 450 FORCE motion sequences with 67 benchmark scenes, yielding 30,150 candidate motion–scene combinations. Of these, 11,497 pass the grounding checks and form the retained corpus. Each sequence contains 160 frames and yields 12 overlapping 16-frame windows, producing 137,964 motion windows in total. Table 1 reports the complete corpus before the test split, distinguishing source motions, grounded sequences, and derived windows. Source and scene identifiers are retained throughout window extraction. Appendix C details grounding, response transfer, and window extraction; Appendix B describes evaluation.

5 Experiments

We evaluate force-dependent motion fidelity and scene-interaction quality by comparing ForceHOSI with existing interaction generators and a direct force-conditioning baseline. Ablations examine the response model and task-preserving composition, followed by an added-obstacle example and a group-level monotonicity test.

5.1 Setup

We evaluate on the sequence-level held-out 10% test split of FORCE-Scene (Section 4). Each 16-frame evaluation window contains two observed frames and a 14-frame generated suffix. Predictions are stitched into sequence-level outputs for interaction-quality evaluation, while physical-response metrics are computed over the generated suffixes. All comparisons use the same test examples and random seeds. Appendix B distinguishes paired-window evaluation from continuous rollout.

We compare with scene-conditioned interaction methods, including CHOIS (Li et al., 2024), HOSIG (Yao et al., 2026), and InfBaGel (Zou et al., 2026), as well as FORCE (Zhang et al., 2025) with scene grounding. We additionally evaluate force FiLM (Perez et al., 2018) as a direct-conditioning baseline. The adapted InfBaGel baseline receives force in text; Appendix D specifies its inputs and adaptation. OMOMO (Li et al., 2023) is excluded because it receives the complete object trajectory.

Physical-response metrics include pose geodesic error, response-direction cosine, mean response magnitude, and force-response ordering. Mean response magnitude is evaluated relative to the reference value of $12.23°$. All reported magnitudes fall below this reference, so a higher value indicates closer agreement within the reported comparison. Interaction metrics include human and object

Table 2: **Force-dependent motion fidelity on paired windows.** Metrics exclude the two observed prefix frames. The reference mean response magnitude is 12.23°. All reported magnitudes are below this value, so higher values indicate closer agreement in this comparison. Bold indicates the best result in each column.

Method	Magnitude (°) ↑	Ordering ↑	Direction ↑	Pose (°) ↓
CHOIS (Li et al., 2024)	9.97	0.673	0.135	19.34
HOSIG (Yao et al., 2026)	10.24	0.695	0.158	18.61
FORCE + grounding (Zhang et al., 2025)	11.38	0.836	0.229	**15.72**
InfBaGel (Zou et al., 2026)	10.30	0.712	0.162	18.67
InfBaGel + force FiLM	11.13	0.806	0.248	17.08
ForceHOSI	**11.46**	**0.852**	**0.249**	16.97

endpoint errors T_h, T_o, foot sliding, and human–scene (HS) and object–scene (OS) penetration. Appendix B further describes the response-ordering statistic.

Training proceeds by adapting the task-generation stream to FORCE-Scene, followed by learning the force-response and manipulation-refinement components with the task stream fixed. The pose carrier remains frozen during manipulation-refiner training. Appendix A specifies the response objective, loss weights, and deployed action policies. All qualitative comparisons use the official body reconstruction and a consistent camera convention.

5.2 Comparison with Existing Methods

Table 2 compares force-dependent motion fidelity on paired windows. ForceHOSI achieves the highest response-norm ordering and direction agreement, with a mean response magnitude closest to the reference. It improves all four metrics over InfBaGel and InfBaGel with force FiLM. The direction advantage over FiLM is small (0.249 versus 0.248), while ordering rises from 0.806 to 0.852. The response model thus better recovers the relative strength of postural changes than direct feature modulation.

Compared with FORCE with grounding, ForceHOSI produces a mean response magnitude closer to the reference and improves ordering and direction agreement. FORCE retains the lowest pose geodesic error (15.72° versus 16.97°), indicating better absolute pose reconstruction under this protocol. These metrics capture different aspects of fidelity: a smaller absolute pose error need not imply more accurate variation across force conditions. The gains of ForceHOSI are therefore most evident in recovering force-dependent variation within a scene-conditioned task motion.

Table 3 evaluates the resulting scene interactions. ForceHOSI obtains the lowest object endpoint error and OS penetration aggregate among the compared methods. Relative to InfBaGel, object endpoint error decreases from 11.360 to 9.752 cm, while OS penetration decreases from 0.0634 to 0.0425 m. Human endpoint error and foot sliding are also lowest, although the endpoint difference is only 0.010 cm and the sliding improvement is modest. The larger gains concern object placement and object–scene compatibility, showing that the improved response fidelity is accompanied by improvements in these task-quality measures.

HOSIG gives the lowest HS penetration rate, whereas ForceHOSI records 42.73%. The reduction in object–scene penetration does not extend to uniformly better collision measures for both the human and the object.

Figure 3 compares pushing a box and carrying a backpack under two force conditions each. The matched examples show how trunk and arm postures vary with physical demand. Such variation needs to be assessed in both magnitude and direction: correctly ordering response norms does not ensure that individual joints move in the appropriate directions. Tables 2 and 3 provide the corresponding quantitative comparisons of response fidelity and scene compatibility.

5.3 Ablation Studies

Table 4 examines conditional weight modulation, force-coordinate transport, response supervision, and task-preserving composition.

Table 3: **Scene-interaction quality of stitched paired-window outputs.** T_h and T_o are endpoint errors in centimeters. HS% is the percentage of human–scene penetration frames, and OS is the mean object–scene penetration aggregate in meters. Bold indicates the best result in each column.

Method	$T_h \downarrow$	$T_o \downarrow$	Slide $\downarrow$	HS% $\downarrow$	OS $\downarrow$
CHOIS	10.18	12.24	1.847	46.3	0.0712
HOSIG	9.58	10.63	1.691	**37.6**	0.0441
FORCE + grounding	11.52	12.91	1.963	49.6	0.0823
InfBaGel	9.236	11.360	1.744	43.10	0.0634
InfBaGel + force FiLM	9.47	10.21	1.724	43.2	0.0497
ForceHOSI	**9.226**	**9.752**	**1.688**	42.73	**0.0425**

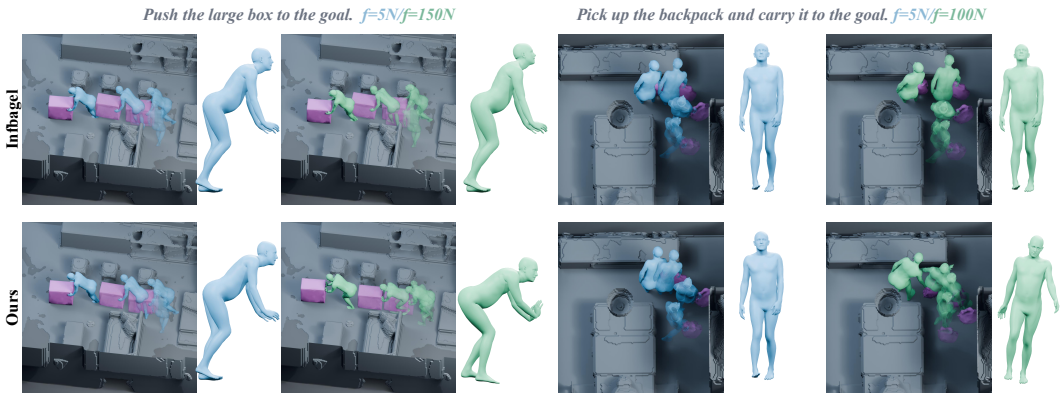

Figure 3: **Force-dependent posture generation.** Top: InfBaGel; bottom: ForceHOSI. Left: pushing a box at 5 and 150 N. Right: carrying a backpack at 5 and 100 N. Blue and green indicate the lower and higher force conditions.

Removing action factorization worsens all four reported metrics: ordering and direction agreement decrease, while pose error and OS penetration increase. Removing physical-type factorization produces smaller degradations. A single transport update also underperforms the four-step configuration across these measures, supporting successive adaptation along the force coordinate. Removing the magnitude loss gives the lowest ordering and direction agreement, highest pose error, and greatest OS penetration among the listed variants. The comparison shows that response-magnitude supervision matters for recovering both the strength and direction of the learned postural change.

The composition ablations show a tradeoff between direction agreement and the other metrics. Removing the manipulation refiner increases direction agreement from 0.249 to 0.271, but raises pose error from 16.97° to 17.23° and OS penetration from 0.0425 to 0.0527. The complete model achieves the highest ordering, lowest pose error, and lowest OS penetration among these variants. Removing the task/support skip worsens all four metrics, indicating that retaining selected task channels helps preserve both response fidelity and interaction quality. The two composition components therefore have distinct effects.

Replacing the requested force with the nominal condition changes ordering from 0.852 to 0.847 and pose error from 16.97° to 17.11°. This small gap limits the evidence that the aggregate gains depend on the requested force. The group-level test in Section 5.5 measures a different property: whether posture changes monotonically across force conditions.

5.4 Qualitative Example with an Added Obstacle

As an additional qualitative test, we insert a horizontal barrier into a suitcase-carrying scene and compare motions under supported loads of 5 N and 100 N, with the same object, task instruction, and goal. As shown in Figure 4, InfBaGel produces similar carrying postures and steps over the barrier under both conditions, showing little load-dependent adaptation in this example. In contrast, ForceHOSI steps over the barrier under the lighter load but adopts a more forward-leaning carrying

Table 4: **Model ablations.** Response metrics use paired-window suffixes; OS uses stitched outputs. Bold indicates the best result in each column.

Variant	Ordering ↑	Direction ↑	Pose (°) ↓	OS ↓
w/o action factorization	0.781	0.194	19.31	0.0579
w/o physical-type factorization	0.839	0.231	17.42	0.0448
One transport step	0.804	0.202	20.13	0.0601
w/o magnitude loss	0.771	0.162	21.41	0.0673
w/o manipulation refiner	0.829	**0.271**	17.23	0.0527
w/o task/support skip	0.818	0.217	18.92	0.0506
Nominal force at inference	0.847	0.236	17.11	0.0463
ForceHOSI	**0.852**	0.249	**16.97**	**0.0425**

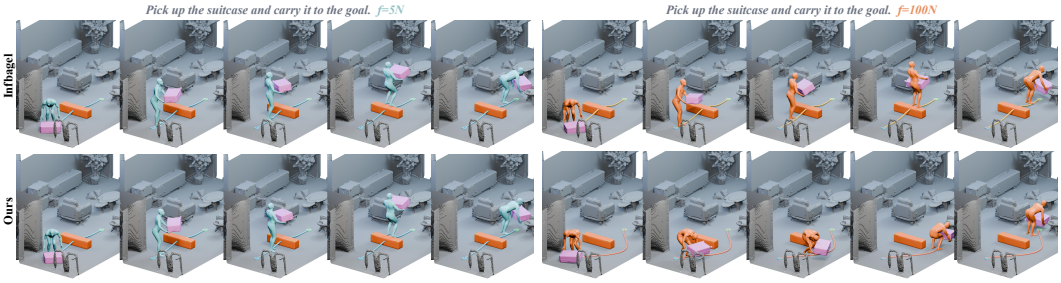

Figure 4: Qualitative comparison with an added obstacle. Top: InfBaGel; bottom: ForceHOSI. Left: $f = 5\,\mathrm{N}$; right: $f = 100\,\mathrm{N}$. InfBaGel crosses the barrier at both loads, whereas ForceHOSI crosses under the lighter load and detours with a more forward-leaning posture under the heavier load.

posture and takes a detour under the heavier load. This example illustrates the complete system under different loads; it does not isolate the response field as the cause of route changes.

5.5 Monotonicity across Force Conditions

We randomly sample 200 force-condition groups per action from the test set. Each compares the same base action and scene under increasing forces, without concatenating motions in time. The 600 groups cover different environment combinations and reuse source recordings. A group passes when its posture-response descriptor increases strictly across force levels (Appendix E). This differs from response-norm ordering. Table 5 shows that ForceHOSI passes 550/600 groups (91.7%), versus 303/600 (50.5%) for InfBaGel, improving all three actions. These results characterize the curated force groups described in Appendix C.

Table 5: Group-level monotonicity (↑), reported as pass rate (passing/total groups).

Action	InfBaGel	ForceHOSI
Carry	37.5% (75/200)	**87.5% (175/200)**
Pull	56.0% (112/200)	**87.5% (175/200)**
Push	58.0% (116/200)	**100.0% (200/200)**
Overall	50.5% (303/600)	**91.7% (550/600)**

6 Conclusion

ForceHOSI adapts human posture to object force in scene-conditioned interactions. A continuous response field supplies postural changes, while task-preserving composition retains support channels and refines manipulation. FORCE-Scene provides paired supervision from scene-grounded force recordings with separate manipulation strategies. The method improves paired-window response ordering and object-task accuracy, with remaining pose and penetration errors, and passes more group-level monotonicity checks than InfBaGel. An added-obstacle example illustrates the complete system's different postures and routes under different loads.

Reproducibility Statement

Sections 3 and 4 describe the model and dataset. Appendix A gives the training objective, loss weights, and action-specific composition policies; Appendix B specifies the sequence-level split, checkpoint selection, and paired-window evaluation. Appendix C details scene grounding, response transfer, filtering, and window extraction. Appendix D describes baseline conditioning and adaptation, and Appendix E defines the posture descriptors, thresholds, and group-level counting rule used for the monotonicity evaluation.

AI Use Statement

We used a generative AI assistant for language revision and translation, LaTeX formatting. It was not used to implement the method, generate data or images, or design the experiments. The authors are responsible for the final manuscript, methods, data, and results.

References

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 (CVPR)*, pp. 15935–15946, 2022.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

David Ha, Andrew Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://research.google/pubs/hypernetworks-2/.

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 (ICCV)*, pp. 2282–2292, 2019.

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)*, 2021a. URL https://samp.is.tue.mpg.de/.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D Scenes by Learning Human-Scene Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14708–14718, 2021b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction. In *German Conference on Pattern Recognition (GCPR)*, volume 13485 of *Lecture Notes in Computer Science*, pp. 281–299, 2022. URL https://intercap.is.tue.mpg.de/.

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*, pp. 1–11, 2024. doi: 10.1145/3680528.3687595.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object Motion Guided Human Motion Synthesis. *ACM Transactions on Graphics*, 42(6):202:1–202:11, 2023. doi: 10.1145/3618333.

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://lijiaman.github.io/projects/chois/.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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)*, pp. 10975–10985, 2019.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 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://grab.is.tue.mpg.de/.

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 (ICLR)*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-Based Imitation of Dynamic Human-Object Interaction. *arXiv preprint arXiv:2312.04393*, 2023. URL https://arxiv.org/abs/2312.04393.

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)*, pp. 14928–14940, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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*, volume 40, pp. 11901–11909, 2026. doi: 10.1609/aaai.v40i14.38177.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-Guided Human Motion Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16010–16021, 2023.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: Physics-Aware Human-Object Interaction. In *International Conference on 3D Vision (3DV)*, pp. 1473–1486, 2025. doi: 10.1109/3DV66043.2025.00139. URL https://github.com/xz6014/FORCE_dataset.

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://www.ecva.net/papers/eccv_2022/papers_ECCV/html/6216_ECCV_2022_paper.php.

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)*, pp. 14738–14749, 2023.

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 (CVPR)*, pp. 5745–5753, 2019.

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://openreview.net/forum?id=TeyHNq4WlI.

A Implementation and Training Details

Carrying uses the load/gravity type; pushing and pulling use resistance. The carrying coordinate is linear in the newton-valued annotation. The temporal field uses the centered force features and low-rank weight update in Eq. 8. Action is a supplied condition.

Transport uses four explicit updates evaluated at the current state and midpoint force coordinate (Eq. 6). It is first-order in the state update: midpoint force coordinates do not constitute a second-order Runge–Kutta midpoint state evaluation. The transport is the identity at zero force interval, whereas the final output also depends on the manipulation refiner.

The task-stream prediction serves as the initial reference carrier $Y(0)$ in Eq. 5. It defines the origin of the response transformation, but the generated carrier is not a measurement at a known physical load. The supplied action label controls the composition policy; the system does not select this label autonomously. The 232-D state contains body rotations but no separate finger-rotation channels. Using a common coordinate frame allows spatial comparisons but does not itself enforce forward-kinematic consistency between the predicted joint positions and rotations.

Response objective. Let r_t denote the 6D pose-rotation channels and w_j the joint weights. The pose term is the mean joint-weighted squared error. The additional rotation term uses

$$\mathcal{L}_{\text{rot}} = \mathbb{E}_{t,j}\left[w_j\left(1 - \frac{\text{tr}(\widehat{R}_{tj}R_{tj}^{*\top}) - 1}{2}\right)\right], \tag{12}$$

with the cosine clamped to its numerical domain. Temporal loss compares first differences of predicted and target pose channels. Define the predicted response vector $\widehat{d}$ and target response d^* relative to the same nominal motion. The direction and magnitude objectives are

$$\mathcal{L}_{\text{dir}} = \mathbb{E}[1 - \cos(\widehat{d}, d^*)], \tag{13}$$

$$\mathcal{L}_{\text{mag}} = \mathbb{E}\left[\text{SmoothL1}\left(\log\frac{\max(\left\|\widehat{d}\right\|, \epsilon)}{\max(\|d^*\|, \epsilon)}, 0\right)\right]. \tag{14}$$

The preservation term penalizes changes to the nominal joint-position and object/contact channels during response learning. The recorded weights are $\lambda_R = 0.5$, $\lambda_v = 0.03$, $\lambda_d = 0.05$, $\lambda_m = 0.03$, and $\lambda_q = 0.10$. No explicit monotonic loss is active. Monotonic supervision enters through the selected response chains, and force ordering is evaluated separately.

Action policies in the deployed system. Carrying uses the pose-carrier object/contact refiner while preserving task-stream pelvis and lower-body support. Pushing retains the physical trunk response and copies the task-arm rotations and object/contact channels. Pulling uses the task stream as a reference and retains subsequent motion adaptation in the output pipeline. Arm indices are 13, 14, and 16–21 in the 22-joint rotation layout. These policies depend on the action label and generated state; no target pose is used at inference.

The corpus uses 450 source motion sequences and 67 scenes, yielding 30,150 candidate motion–scene combinations. Grounding checks retain 11,497 sequences; extracting 12 overlapping 16-frame windows per retained sequence gives 137,964 windows. These counts describe the complete corpus before the test split. Source identifiers and placement transforms preserve construction provenance. Interaction-specific object offsets remain fixed across force conditions.

B Evaluation Protocol Details

The paired protocol conditions each 16-frame window on its first two frames and evaluates the remaining 14. Stitched outputs therefore receive repeated observations across a longer sequence. The continuous scene protocol conditions only the first two frames of a 26-frame crop and generates the rest autoregressively. Metrics from these protocols reflect different amounts of conditioning and error accumulation.

We split FORCE-Scene at the sequence level, reserving 10% as a held-out test set. Windows from the same sequence remain in the same partition. The dataset totals in Table 1 refer to the complete corpus, rather than the training or test partition alone. All compared methods are evaluated on the same test examples under the corresponding conditioning protocol.

Response checkpoint selection. The response model uses training-set reconstruction metrics for checkpoint selection, with no separate validation set for this step. This selection set is distinct from the held-out test set used for evaluation. Checkpoints minimize

$$S = \text{Pose} + 4(1 - \text{Direction}) + 2\left|\log \max(\overline{\rho}, 10^{-6})\right| + (1 - \text{Ordering}), \tag{15}$$

where $\overline{\rho}$ is the mean predicted-to-target response-norm ratio relative to the nominal carrier, with each target norm floored at 10^{-6}. It is dimensionless and differs from the angular Magnitude reported in Table 2. Selection is evaluated at epoch 1, every four epochs, and the final epoch. An improvement must be strict, so an exact tie retains the earlier checkpoint. The adapted task baseline has a separate selection rule specified in Appendix D.

Pose geodesic error measures angular disagreement with the target rather than joint-position error. Response direction and magnitude are defined relative to the nominal motion. They therefore assess the predicted response relative to a reference, rather than directly measuring the difference between two generated outputs obtained by intervening on the requested force while fixing all other inputs.

The ordering statistic compares response norms across increasing force values in compatible groups. This response-norm score differs from the action-specific descriptors used to select source chains. Results on the selected chains do not establish monotonic responses across all manipulation strategies. Joint-wise direction and pose error provide complementary measurements.

Scene placements expand contextual diversity while reusing recorded source motions; they do not constitute independent physical recordings. Stitched paired-window results do not establish initial-prefix-only continuous rollout performance. Qualitative execution choices concern the complete system and do not isolate the effect of the posture-response component on route selection.

C Grounding and Response-Pair Construction

We crop interactions to the interval between 3% and 97% of cumulative object path length. The interval is extended by $\max(4, \text{round}(0.03T_{\text{raw}}))$ frames and then by $\max(4, \text{round}(0.04T_{\text{crop}}))$ frames, clipped to source boundaries. Sequences with total object path length below 10^{-5} m retain the full interval. Captures are uniformly resampled to 160 frames, using componentwise linear interpolation for translations and spherical interpolation for body and object rotations. Transferred SMPL-X geometry is loaded from the matching cache with crop and shape checks. This aligns normalized time rather than contact events; we do not use dynamic time warping between captures.

Offsets are applied before placement in a common y-up coordinate system and remain fixed across force conditions within an interaction strategy. The vertical offset is -0.40 m for laundry-basket and two-handed-suitcase interactions and zero for other strategies. A placement is accepted only if human and object vertices have scene SDF values at least -0.003 m, all vertices are inside the scene grid, both foot-bottom minima are at least -0.001 m, and ground support is within 0.035 m in at least 80% of frames. Hand–object distance is recorded but is not part of this Boolean acceptance gate. The corpus counts are given in Table 1.

Strategy grouping and donor selection. Sources are organized by interaction family (action, object, direction, and annotated hand configuration), strategy cluster, and force. Strategy descriptors comprise late trunk tilt, trunk alignment with object displacement (absolute for carrying and signed otherwise), negative pelvis height, knee flexion, and an elbow-angle descriptor. Late descriptors use frames 80–159; knee and elbow descriptors use the full sequence. After robust standardization within each action, a descriptor z is embedded as

$$e(z) = \left[\frac{z}{\max(\|z\|_2, 10^{-6})},\ 0.35\log(1 + \|z\|_2)\right]. \tag{16}$$

Within each family, average-linkage clustering uses cosine distance with threshold 0.8; non-finite cosine distances fall back to Euclidean distance. Force ladders are selected within a strategy cluster

in increasing annotated-force order, with a minimum standardized score margin of 0.03. The score is late trunk tilt for carrying, signed trunk alignment for pushing, and its negative for pulling.

Response transfer to a nominal carrier. The reference donor is the canonical zero-force capture. Each nonzero donor is a recorded capture with the requested force label and selected strategy. Let R^0_{tj} and R^f_{tj} be donor local rotations, B_{tj} the local carrier rotation, and l_j, g_j the local and global transfer weights. Transfer applies

$$D_{tj} = R^f_{tj}(R^0_{tj})^\top, \tag{17}$$

$$\widetilde{B}_{tj} = \mathrm{Exp}(l_j \,\mathrm{Log}(D_{tj}))\, B_{tj}, \tag{18}$$

$$\widehat{G}_{tj} = \mathrm{Exp}\Big(g_j \,\mathrm{Log}(\widetilde{G}_{tj} G^\top_{tj})\Big)\, G_{tj}, \tag{19}$$

where G and $\widetilde{G}$ accumulate the carrier and transferred local rotations along the body tree. Only global body-rotation channels are replaced; joint positions, object motion, and contact channels retain carrier values. Zero-force targets are identity pairs. Thus paired response targets are transferred rotations on a carrier, rather than complete unmodified donor recordings.

Donors in a force ladder share a strategy cluster. Carriers are selected from all zero-force captures in the same family and checked after transfer; they are not required to share the donor's cluster ID. The strategy restriction therefore applies to donor selection, while the resulting carrier–response pairs must pass the checks below.

At most eight donor candidates are considered at each nonzero force. After checking increasing donor scores, the transferred chain, including its zero-force carrier, is checked with margin 10^{-4}. For carrying, the descriptor is mean pelvis-to-neck tilt from vertical, in degrees, over frames 16–159. For pushing and pulling, it is the mean normalized horizontal trunk projection along and opposite the object's net displacement, respectively, over the same frames. Displacement below 10^{-5} m gives descriptor zero. The pushing and pulling descriptors are dimensionless, so the margin has action-specific units. These checks use force order explicitly; ordering results characterize curated chains rather than the unfiltered population. Compared methods use the same paired test examples and chain groups.

For 160-frame sequences, 16-frame windows with nominal overlap two use stride 14. Zero-based starts are $0, 14, \ldots, 140$, followed by a final start at 144 to cover the endpoint, giving 12 windows. The final two windows overlap by 12 frames even though the generation prefix remains two frames. Pair metadata include family, base-source ID, target-source ID, strategy, scene, force, and start index; placement provenance is retained as described in Section 4.

D Adapted InfBaGel Comparison Protocol

The adapted InfBaGel task baseline and ForceHOSI receive the same scene occupancy, object geometry and basis-point-set (BPS) representation, timing metadata (phase and sequence duration), human pelvis goal, object position goal, and two-frame prefix. Goals are taken from the nominal carrier's endpoints; no complete future trajectory tensor is supplied. Both methods receive the same CLIP text prompt containing the task and annotated force. InfBaGel therefore receives force information through text, while ForceHOSI additionally receives the force in newtons as an explicit numeric input to the response module. These input details concern the paired-window comparison with the adapted InfBaGel baseline, rather than the additional force-FiLM variant; the obstacle example also uses the passage constraints described in Section 3.

Fine-tuning of the released InfBaGel checkpoint is configured for 60 epochs, batch size 12, learning rate 5×10^{-5}, gradient clipping at 1.0, seed 20260831, and one A100 GPU. The baseline selects the checkpoint with the lowest training loss, retaining the earlier checkpoint on an exact tie. The response branch uses a different optimization budget: although the methods share data and sampling conditions, their training compute is not matched. These settings specify the configured training budget rather than measured wall time or the epoch of the selected checkpoint.

E GROUP-LEVEL MONOTONICITY EVALUATION

This protocol defines the posture-response descriptors and pass criterion used in Section 5.5 and Table 5. Each force-condition group contains outputs for the same base action and scene under increasing forces. The outputs are compared across conditions, rather than concatenated in time.

Posture descriptors. We reconstruct joints with SMPL-X and define the trunk vector in a y-up coordinate system as

$$\mathbf{v}_t = \mathbf{J}_{\text{neck},t} - \mathbf{J}_{\text{pelvis},t}, \tag{20}$$

using joint indices 12 and 0, respectively. Let $\mathbf{v}_{t,xz}$ denote its horizontal projection. For each output, $\mathbf{d}_{xz}$ is the horizontal unit direction obtained from that output's own object-trajectory endpoint minus its start point. The action-specific descriptors are

$$D_{\text{carry}} = \text{mean}_t \left[\arccos\left(\frac{v_{t,y}}{\|\mathbf{v}_t\|} \right) \frac{180}{\pi} \right], \tag{21}$$

$$D_{\text{push}} = \text{mean}_t \left[\frac{\mathbf{v}_{t,xz} \cdot \mathbf{d}_{xz}}{\max(\|\mathbf{v}_{t,xz}\|, 10^{-8})} \right], \qquad D_{\text{pull}} = -D_{\text{push}}. \tag{22}$$

The carry descriptor is the temporal mean inclination from vertical, in degrees. The push descriptor measures mean horizontal alignment with object motion, and the pull descriptor reverses its sign to measure alignment against object motion. Both are dimensionless. Increasing force is expected to increase the corresponding descriptor.

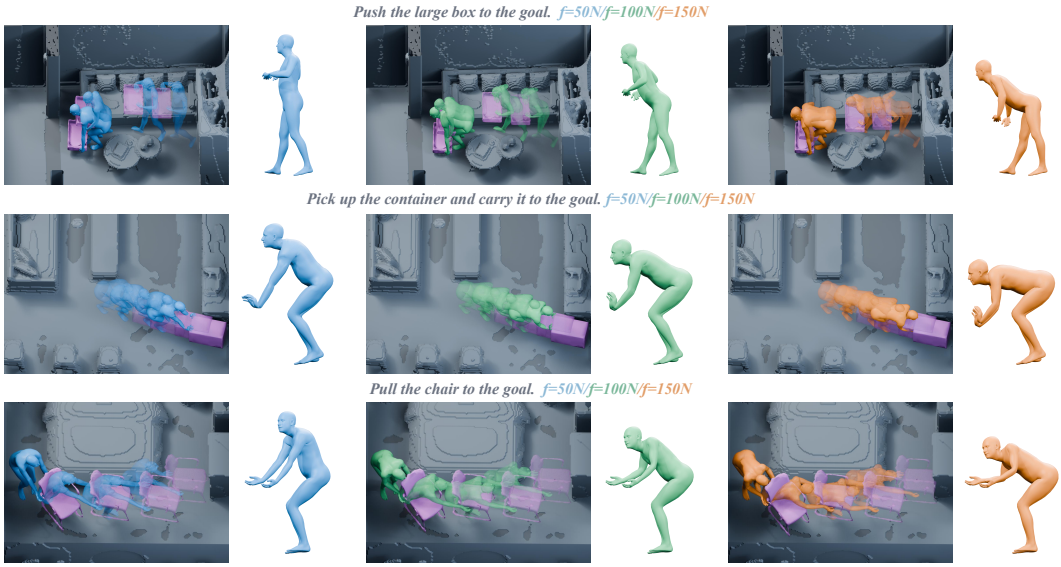

Figure 5: ForceHOSI outputs at 50, 100, and 150 N (left to right; blue, green, and orange). Rows show pushing a box, carrying a container, and pulling a chair. These examples illustrate responses at the displayed discrete force levels.

Strict monotonicity and counting. For force levels $f_0 < f_1 < \cdots < f_K$, a group passes only if every adjacent pair satisfies

$$D(f_{i+1}) - D(f_i) > 10^{-4}, \qquad i = 0, \ldots, K-1. \tag{23}$$

The threshold is 0.0001° for carry and 0.0001 for push and pull. Equal values, decreases, and positive increments no greater than the threshold all fail the test. Each complete group contributes one outcome: the pass rate is the number of passing groups divided by the number of valid groups, rather than a rate over adjacent pairs or individual windows. Table 5 uses 200 groups per action and 600 in total.

The push and pull descriptors measure lean-direction alignment, not forward or backward tilt magnitude. Normalizing the horizontal trunk vector can give the same score to small and large tilts in the

same direction. The carry descriptor measures total inclination and does not distinguish forward, backward, or lateral lean. The group-level test therefore assesses monotonicity of these specific posture descriptors, separately from response-norm ordering.

F Qualitative Responses and Limitations

Figure 5 shows ForceHOSI outputs for pushing, carrying, and pulling at three discrete force levels.

Limitations. Our current results are limited to FORCE's discrete force levels. The continuous field does not yet provide precisely calibrated motion responses at arbitrary force values.