HOI-UP: ADAPTING HUMAN-OBJECT INTERACTION MOTION TO UNSEEN OBJECT PROPERTIES WITHOUT PROPERTY-LABELED MOTION DATA

Anonymous authors
Paper under double-blind review

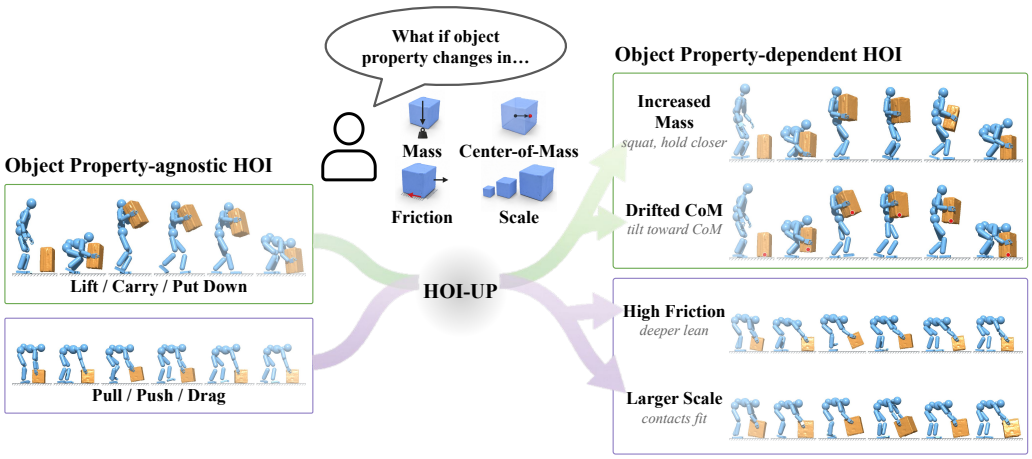


Figure 1: **HOI-UP.** Given an existing human-object interaction (HOI) reference (left) and changes in the object's physical properties (top), HOI-UP adapts the reference into property-dependent HOI motions (right). The red dot in *Drifted CoM* marks the shifted center-of-mass (CoM).

ABSTRACT

How humans interact with an object depends on its physical properties, yet widely used Human-Object Interaction (HOI) datasets provide little explicit supervision for property-dependent adaptation. We aim to adapt an existing HOI reference to unseen object properties without property-labeled motion data. Existing property-aware methods such as FORCE and PA-HOI require additional motion data for new property conditions, while motion editing approaches such as CoMo rely on discrete pose-code editing that compromises the precise contacts required for physical HOI execution. Biomechanical objectives, used as the sole adaptation guidance in motion optimization or physics-based imitation, yield similar motion strategies across object-property levels rather than human-like variations. In this paper, we introduce HOI-UP, a framework for adapting existing HOI references to unseen object properties. A vision-language model (VLM) predicts property-dependent changes to compact, literature-informed and empirically validated HOI-codes. Kinematic optimization adapts the reference toward these targets in an anatomically plausible HOI motion space. RL-based tracking then refines the resulting plans in physics simulation for execution under the target properties. We further introduce an evaluation protocol that jointly assesses property response, reference preservation, naturalness, biomechanical demand, and physical executability across object property changes. Our motions adapted to unseen properties follow human-like posture changes and vary monotonically along the property axis, while achieving the highest tracking success rate and the lowest tracking errors under constrained actuator powers.

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.

1 INTRODUCTION

Realistic human–object interaction (HOI) motion is essential for high-fidelity virtual humans in animation, virtual reality, and embodied AI. Existing HOI datasets (Taheri et al., 2020; Bhatnagar et al., 2022; Li et al., 2023; Lv et al., 2024) and generative models (Tevet et al., 2023; Chen et al., 2023; Shafir et al., 2024; Xu et al., 2023; Diller & Dai, 2024; Li et al., 2024; Nam et al., 2026) have enabled the synthesis of diverse interaction motions from semantic and spatial conditions. However, how humans interact with an object also depends on the object's physical properties such as mass, scale, center of mass, and object-floor friction. For example, carrying a heavier object leads a person to hold it closer to the body and take shorter strides (Knapik et al., 2004; Mexi et al., 2023; Schipplein et al., 1995). Widely used HOI datasets are largely agnostic to these properties, providing no explicit supervision for property-dependent motion changes and limiting data-driven generation and reference adaptation under unseen physical conditions.

Recent property-aware approaches (Zhang et al., 2025; Wang et al., 2025) directly capture interactions under varying object properties and train generators on the collected data, but extending them to unseen property axes or ranges requires additional data collection and training. CoMo (Huang et al., 2024) avoids data collection via LLM-based pose code editing, yet its human-centric codes omit human-object relations and do not explicitly preserve the interaction structure of the reference. Biomechanical optimization (Xiang et al., 2010) explicitly reduces joint loading, but a joint-loading objective alone does not determine how a human-like interaction strategy should vary across object property variations. Physics-based imitation (Wang et al., 2023; Xu et al., 2025b) improves physical feasibility, but tracks the given reference rather than determining how that reference should adapt.

Our key observation is that property-agnostic datasets already contain abundant natural HOI references with human and object motion, temporal organization, and contacts. HOI-UP therefore synthesizes property-adapted motions for unseen object properties, specified only by their property values without any motion data captured under them, by inferring how an existing reference should change. It represents the reference using compact HOI-codes that capture human motions, object states, and human-object relations. A VLM predicts continuous, phase-wise HOI-code targets, and interaction-aware optimization realizes these targets while retaining the reference task and discouraging excessive joint loading. A physics-based controller is then fine-tuned under the target properties to produce the final motion. All three stages, including target-condition controller fine-tuning, form part of the synthesis pipeline. The same synthesis procedure can be applied to new reference–property inputs without collecting corresponding property-labeled motions.

To evaluate adaptation without paired ground-truth target motions, we compare adaptations of the same reference across controlled property sweeps. The protocol jointly assesses physical executability, motion naturalness, biomechanical demand, property response, and preservation of the reference interaction structure. Because these criteria capture different and potentially competing aspects of adaptation, we interpret them as complementary evidence rather than collapsing them into a single score. Experiments across multiple properties show that HOI-UP produces consistent and balanced adaptations under this joint evaluation.

Our contributions are summarized as follows:

- **Unseen-Property HOI Adaptation:** We formulate the task of adapting existing property-agnostic HOI references to unseen object-property conditions without property-labeled motion supervision, while preserving the underlying interaction structure.
- **HOI-UP framework:** We propose HOI-UP to address this task, combining VLM-predicted phase-wise changes to continuous HOI-codes with interaction-aware motion optimization and target-condition controller fine-tuning to synthesize property-adapted HOI motions. HOI-UP demonstrates human-like adaptation across the evaluated property types and levels, achieving higher physical executability than the compared baselines under a common actuation profile.
- **Evaluation Protocol:** We introduce an evaluation protocol for this task that compares adaptations of the same reference across multiple levels of a single object property without requiring paired ground-truth target motions. The protocol jointly assesses property response, reference preservation, naturalness, biomechanical demand, and physical executability.

2 Related Works

Physical Property-Aware Human-Object Interaction. Existing HOI datasets (Taheri et al., 2020; Bhatnagar et al., 2022; Li et al., 2023; Jiang et al., 2023; Lv et al., 2024) and generators (Xu et al., 2023; Diller & Dai, 2024; Li et al., 2024; Nam et al., 2026) capture and synthesize whole-body interactions, but most do not explicitly model object physical properties. Recent approaches have begun to incorporate object properties into HOI generation. MACS (Shimada et al., 2024) considers mass-conditioned hand–object interactions, while PA-HOI (Wang et al., 2025) captures whole-body interactions across weights, sizes, and shapes, but uses coarse size and weight categories with limited within-object variation. FORCE (Zhang et al., 2025) captures and generates interactions under varying resistance, while other physical conditions, such as object-floor friction, remain fixed. Extending these methods to new property types or ranges requires additional motion data, whereas we adapt property-agnostic HOI references to unseen conditions without property-labeled supervision.

Foundation Models for Motion Adaptation. Foundation-model priors extend motion generation to open-vocabulary language (Tevet et al., 2022), unseen action compositions (Athanasiou et al., 2023), and novel objects or interactions (Xu et al., 2024; Li et al., 2025; Li & Dai, 2026), but these HOI generators do not explicitly model reference adaptation to quantitative property changes. Foundation models also support reward design for physics-based control (Yu et al., 2023; Ma et al., 2024; Cui et al., 2025), but RL-based reward refinement requires costly policy optimization and rollout feedback.

Motion edits can also be specified through interpretable pose and motion codes (Pons-Moll et al., 2014; Delmas et al., 2022; Yazdian et al., 2025). CoMo (Huang et al., 2024) reconstructs motion from language-edited codes using a learned decoder. Its human-centric vocabulary omits object states and human–object relations, while discretization limits fine-grained control over contact geometry. HOI-UP instead uses a VLM to predict property-conditioned, phase-wise changes to compact, continuous HOI-codes for precise interaction control.

Biomechanics and Motion Optimization. We use biomechanical findings on load-dependent posture and joint loading (van Dieën et al., 1999; Burgess-Limerick, 2003; Lavender et al., 2003; Shoji et al., 2025) and the relation between floor friction and gait (Yamaguchi et al., 2013; Cooper et al., 2008) to select HOI-codes, not to supervise their property-dependent changes.

OWAS (Karhu et al., 1977) and REBA (Hignett & McAtamney, 2000) assess postural risk, while the RNLE-based Lifting Index (RNLE LI) (Waters et al., 1993) assesses task-specific lifting demand. Inverse-dynamics joint moments provide continuous loading measures for motion optimization. Xiang et al. (2010) predict lifting motions by jointly optimizing dynamic effort and stability under physical constraints. However, minimizing biomechanical cost alone can compromise naturalness and favor similar low-load strategies across property levels. Our optimization instead combines VLM-predicted HOI-code targets with joint-loading and balance objectives to preserve naturalness during property-dependent adaptation.

Physics-Based Motion Generation and Control. Physics-based control tracks reference human motions (Peng et al., 2018; Luo et al., 2023) and contact-rich HOI (Wang et al., 2023; Xu et al., 2025b), and complements kinematic motion generation (Yuan et al., 2023; Ren et al., 2023; Tevet et al., 2025). The zero-moment point (ZMP) assesses balance relative to the foot support polygon (Vukobratović & Borovac, 2004). However, successful tracking alone does not establish executability within human strength limits. Strength measurements characterize joint-specific torque capacity (Danneskiold-Samsøe et al., 2009; Holzbaur et al., 2007), while musculoskeletal models provide state-dependent limits (Jiang et al., 2019). Joint impedance studies additionally examine stiffness and damping (Rouse et al., 2013; 2014). We use InterMimic under a fixed human-strength-informed actuation to mitigate physical artifacts and assess executability under target properties.

Evaluation of HOI and Reference Adaptation. Existing HOI-quality (Li et al., 2023; 2024) and smoothness metrics (Yi et al., 2021; Du et al., 2023) alone do not establish property-dependent adaptation with reference preservation when ground-truth target motions are unavailable. Melistas et al. (2024) evaluate counterfactual images through composition, effectiveness, minimality, and realism without ground-truth counterfactuals. Inspired by these principles, we jointly assess property response, reference preservation, HOI naturalness, biomechanical demand, and physical executability under constrained actuation.

3 METHODS

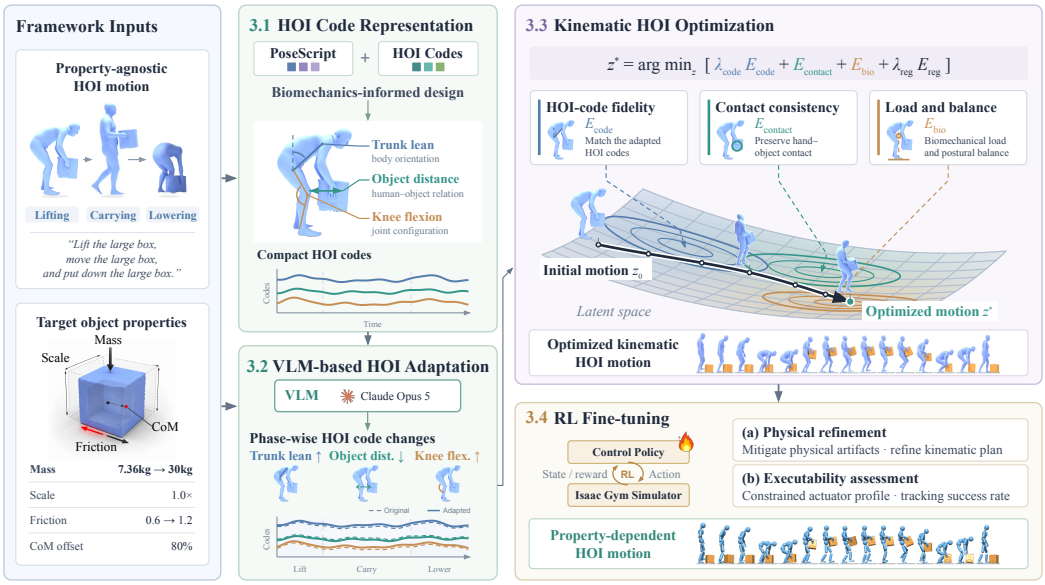


Figure 2: **Overview of HOI-UP.** Given a property-agnostic HOI reference and target object properties, HOI-UP predicts phase-wise HOI-code changes, realizes them through kinematic HOI optimization, and refines and evaluates the adapted motion using physics-based tracking.

HOI-UP adapts a reference HOI motion $\mathcal{H}$ to target object properties ϕ^{tgt} without property-labeled motion supervision. We use default object properties ϕ from the InterMimic simulation settings (Xu et al., 2025b) and consider changes in mass, object–floor friction, center-of-mass offset, and scale. The adapted motion preserves the reference's high-level task and interaction-phase sequence. As in Fig. 2, a VLM predicts phase-wise HOI-code changes, and kinematic optimization realizes them in an adapted motion $\widehat{\mathcal{H}}$ while accounting for contact, balance, and joint loading. Finally, a physics-based imitation controller is fine-tuned to track this plan under ϕ^{tgt} and a fixed human-strength-informed actuation profile, mitigating physical artifacts and producing a simulated rollout $\widehat{\mathcal{H}}^{\text{sim}}$.

3.1 HOI-CODE REPRESENTATION

Property-dependent adaptations vary across the stages of an interaction. We therefore segment each reference into interaction phases based on contact states, object motion, and human–object relations, and use these phases as the temporal units of adaptation (App. A.1).

Following CoMo (Huang et al., 2024), we represent motion with interpretable HOI-codes that a language model can edit. However, the PoseScript codes (Delmas et al., 2022) used in CoMo describe only human-body geometry and thus cannot represent object states or human–object relations. Therefore, we introduce HOI-codes for object states and human–object relations (Tab. 4), and construct a compact HOI-code vocabulary informed by biomechanics and ergonomics studies (Sec. 2). Our default $L + I$ vocabulary combines literature-motivated codes (L) with interaction codes (I) describing whole-body and object-motion states (Tab. 5). Unlike CoMo, which discretizes codes into categorical levels, we retain continuous values to preserve fine-grained reference variations.

3.2 VLM-BASED PROPERTY-DEPENDENT HOI ADAPTATION

In motion editing methods such as CoMo, a natural-language instruction directly specifies the desired change (e.g., "bend the knees more"). In contrast, an object-property change $\phi \rightarrow \phi^{\text{tgt}}$ does not directly specify how the motion should change, so the adaptation strategy must be inferred. We therefore use a pretrained VLM to infer this strategy as phase-wise changes to the HOI-codes.

Let $\mathbf{c}[t] \in \mathbb{R}^C$ denote the reference code vector at frame t, where C is the number of HOI codes, and let $\mathcal{T}_p$ denote the frame indices of phase $p \in \{1, \dots, P\}$. We summarize each phase p as $\mathbf{s}_p = \mathrm{concat}(\mathbf{c}_p^{\mathrm{bgn}}, \mathbf{c}_p^{\mathrm{mid}}, \mathbf{c}_p^{\mathrm{end}}, \mathbf{c}_p^{\mathrm{min}}, \mathbf{c}_p^{\mathrm{max}}) \in \mathbb{R}^{5C}$, concatenating the HOI-code vectors at the beginning, middle, and end of the phase with the code-wise minima and maxima. Given reference-motion images $\mathcal{V}$, phase summaries $\{\mathbf{s}_p\}_{p=1}^P$, and the default and target properties $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^{\mathrm{tgt}}$, the VLM predicts an adapted HOI-code range for each phase as

$$\{(\widehat{\mathbf{c}}_p^{\mathrm{min}}, \widehat{\mathbf{c}}_p^{\mathrm{max}})\}_{p=1}^P = \mathrm{VLM}\left(\mathcal{V}, \{\mathbf{s}_p\}_{p=1}^P, \boldsymbol{\phi}, \boldsymbol{\phi}^{\mathrm{tgt}}\right). \tag{1}$$

The prompt provides the definitions of the HOI-codes but does not prescribe their desired changes. We map the reference's code range onto the predicted range to obtain frame-wise targets as

$$\widehat{\mathbf{c}}[t] = \widehat{\mathbf{c}}_p^{\mathrm{min}} + \mathbf{g}_p \odot \left(\mathbf{c}[t] - \mathbf{c}_p^{\mathrm{min}}\right), \quad \mathbf{g}_p = \left(\widehat{\mathbf{c}}_p^{\mathrm{max}} - \widehat{\mathbf{c}}_p^{\mathrm{min}}\right) \oslash \left(\mathbf{c}_p^{\mathrm{max}} - \mathbf{c}_p^{\mathrm{min}}\right), \quad t \in \mathcal{T}_p, \tag{2}$$

where $\mathbf{g}_p \in \mathbb{R}^C$ is the gain vector for phase p, and $\odot$ and $\oslash$ denote element-wise multiplication and division. This shifts and rescales the code trajectory within each phase while retaining its fine-grained temporal variation. Since Eq. 2 retains the reference timing, we additionally stretch or compress the time scale according to the predicted changes in speed-related codes, capturing adaptations such as slower lifting of heavier objects.

3.3 Kinematic HOI Optimization

Kinematic HOI optimization adapts the reference $\mathcal{H}$ toward the code targets $\widehat{\mathbf{c}}$ while accounting for contact consistency and biomechanical requirements under $\boldsymbol{\phi}^{\mathrm{tgt}}$. We parameterize the human motion with SKEL (Keller et al., 2023) and constrain optimization to the model's anatomical joint limits. We represent changes to the human motion and object pose as bounded B-spline corrections $\mathbf{z}$, yielding the adapted interaction $\mathcal{H}(\mathbf{z})$. Given the code targets $\widehat{\mathbf{c}}$, we solve

$$\mathbf{z}^\star = \arg\min_{\mathbf{z}} \Big[\lambda_{\mathrm{code}} E_{\mathrm{code}} + \underbrace{\lambda_{\mathrm{grasp}} E_{\mathrm{grasp}} + \lambda_{\mathrm{foot}} E_{\mathrm{foot}}}_{E_{\mathrm{contact}}} + \underbrace{\lambda_{\mathrm{zmp}} E_{\mathrm{zmp}} + \lambda_{\mathrm{load}} E_{\mathrm{load}}}_{E_{\mathrm{bio}}} + \lambda_{\mathrm{reg}} E_{\mathrm{reg}}\Big], \tag{3}$$

where all terms are evaluated on $\mathcal{H}(\mathbf{z})$ with fixed weights. The target properties enter through the object model and dynamics — the scale determines the object mesh and its signed distance field (SDF) used in contact and penetration terms, while mass, center of mass, and friction determine the forces and moments in E_{bio}.

We follow the VLM targets through channel-normalized code matching (E_{code}), while contact penalties (E_{contact}) discourage deviations from reference grasp and stance geometry. To promote stable handling under the target properties, E_{bio} combines ZMP-based balance within the foot support polygon with joint-load penalties. The load penalty acts only above reference-derived thresholds, bounding excessive loading rather than favoring the lowest-load strategy at every property level. Regularization (E_{reg}) additionally discourages collisions, task deviation, and temporal irregularity. We optimize each reference–property pair independently using the Adam optimizer (Kingma & Ba, 2015), without sharing information across property levels, and pass $\widehat{\mathcal{H}} = \mathcal{H}(\mathbf{z}^\star)$ to physics-based tracking in Sec. 3.4. Implementation details are provided in App. A.4.

3.4 RL Fine-tuning

We fine-tune InterMimic (Xu et al., 2025b) in Isaac Gym to track $\widehat{\mathcal{H}}$ under $\boldsymbol{\phi}^{\mathrm{tgt}}$, producing $\widehat{\mathcal{H}}^{\mathrm{sim}}$ for physical refinement and executability assessment. We use only the InterMimic tracking reward, since adaptation is realized in the kinematic plan. To reflect human strength limits, we construct a *human-strength-informed actuation profile* (HS profile) with per-axis torque limits and proportional–derivative (PD) gains fixed across reference and target conditions. For literature-supported axes (Danneskiold-Samsøe et al., 2009), we set

$$\bar{\tau}_{j,a} = \frac{m_{\mathrm{H}}}{2} \sum_{s \in \{\mathrm{F},\mathrm{M}\}} \frac{M_{j,a}^s}{m_s}, \tag{4}$$

where j and a index joints and local axes, m_{H} is avatar mass, and m_s and $M_{j,a}^s$ denote sex-specific mean body mass and the larger opposing mean isometric moment; other axes retain heuristic caps. We set $K_{j,a} = \alpha \bar{\tau}_{j,a}$ and $D_{j,a} = \beta K_{j,a}$, anchoring α to ankle stiffness (Rouse et al., 2013; 2014). Construction parameters and training details are provided in Apps. A.5 and A.6.

4 EVALUATION PROTOCOL

Ground-truth target motions are generally unavailable for arbitrary property changes, and motion quality alone cannot establish appropriate adaptation. We therefore propose a protocol that evaluates kinematic plans and their physical executions under the target object properties, comparing adaptations of the same reference across levels of one property while holding the others fixed. Criteria and descriptors are shared across methods. Details are provided in App. A.7.

Physical Executability. Each plan is evaluated under the target properties, HS actuation profile, and evaluation criteria. **Success Rate** measures the percentage of executions that reach the final target frame, while **Completion Ratio** is the mean fraction of the plan completed before termination. Together with **HOI Error**, the mean distance between simulated and planned human joints and object surface points, these metrics assess whether the adapted plan can be completed under the target conditions and how faithfully it is physically realized.

Naturalness. We assess human–object smoothness, hand–object contact preservation, and foot–ground artifacts using **HOI Jitter** ($\mathrm{Jit}_{\mathrm{HOI}}$), **Contact F1**, and **Foot-floor**, respectively (Li et al., 2023; 2024). **Foot–floor error** is the sum of foot-float and foot–floor penetration errors.

Biomechanical Demand. **Mean Joint Moment** ($\bar{M}$) averages model-estimated joint-moment magnitudes in motion plans, whereas **Mean Simulator Torque** ($\bar{\tau}_{\mathrm{sim}}$) averages recorded torque magnitudes during execution. **High Moment Count** (C_{100}) complements the averages, which can hide excessive peaks, by measuring the mean per-frame number of evaluated joints exceeding $100\,\mathrm{N\,m}$. **OWAS** (Karhu et al., 1977) summarizes posture- and load-based action ratings, while **RNLE LI** (Waters et al., 1993) measures object mass relative to a recommended lifting limit. RNLE LI is evaluated only during lifting, carrying, and lowering.

Reference Preservation. **Composition** ($\mathrm{Comp}_{\mathrm{HOI}}$) measures human–object reconstruction error under a default intervention ($\phi^{\mathrm{tgt}} = \phi$), assessing reference fidelity when no adaptation is requested. **Minimality** measures normalized, phase-wise deviations of reference-relative descriptors from their linear trends across object-property levels, penalizing changes not explained by the property change. It uses the 12 literature-motivated codes (L) in Tab. 5 as descriptors. Together, these metrics complement the property-response measures by assessing whether unintended changes occur.

Property Response. **Effectiveness** (Eff) measures weighted agreement between mass-response slope directions of descriptors L and those derived from the FORCE dataset (Zhang et al., 2025), used only for evaluation. **Monotonicity** (Mon) evaluates pairwise ordering and default-level consistency of reference-relative changes in these descriptors, selecting the better-fitting increasing or decreasing direction. Using normalized pose, trajectory, and timing changes relative to each method's default output, we compute the fraction R^2_{prop} of within-group feature variation explained by differences between property-level means. R^2_{all} reports the observation-count-weighted mean of these fractions across valid property–phase groups, expressed as a percentage. These metrics assess how well adapted motions reflect property-level changes through directional agreement, consistent ordering, and distinguishable responses.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Data and target conditions. We use 30 OMOMO reference sequences (Li et al., 2023) for end-to-end comparison and a 20-sequence subset for kinematic ablation, covering diverse motions and objects, and vary mass (10, 20, and 30 kg), scale ($0.75\times$, $1.25\times$), CoM offset ($\pm 80\%$ along the object's longest local axis), and object–floor friction (0.1, 1.2; applied only to the 15 sequences with pushing, pulling, or dragging) one at a time while keeping the remaining properties fixed, yielding 240 reference–property pairs (App. A.8).

Physics-based execution. All methods use InterMimic (Xu et al., 2025b) under the same HS actuation profile; validation of selected strength characteristics is reported in App. A.5. Starting from a common checkpoint adapted to this profile, method-specific controllers are fine-tuned with PPO on their respective tracking targets under identical settings until convergence. Each reference–property pair is evaluated over 5 rollouts with stochastic action sampling (App. A.6).

Table 1: **End-to-end comparison under various target properties.** (a) Controller adaptation and (b) Motion adaptation followed by InterMimic fine-tuning. All methods share the same HS profile.

Method	Physical executability			Naturalness			Biomechanical demand				Preservation	Property response		
	Success ↑ (%)	Completion ↑ (%)	HOI error ↓ (mm)	$\mathrm{Jit}_{\mathrm{HOI}}$ ↓ ($\mathrm{m\,s^{-3}}$)	Foot-floor ↓ (cm)	Contact F1 ↑	$\bar{\tau}_{\mathrm{sim}}$ ↓ (N m)	C_{100} ↓ (joints)	OWAS ↓	RNLE LI ↓	Minimality ↓	Eff ↑ (%)	Mon ↑ (%)	R^2_{all} ↑ (%)
(a) Controller adaptation														
InterMimic	77.42	87.69	79.0	709.21	0.355	0.890	12.78	2.941	2.119	0.903	0.619	**63.93**	23.91	14.66
InterMimic+ZMP	78.00	87.30	81.8	**698.42**	0.289	0.889	**12.44**	2.930	**2.097**	0.883	0.693	63.00	25.79	17.71
InterMimic + HOI-code	73.83	85.56	85.3	743.11	0.415	0.866	13.61	2.958	2.116	0.891	0.798	62.94	23.73	15.86
(b) Motion adaptation														
CoMo-HOI	23.25	51.67	104.4	793.20	0.456	0.726	17.98	3.092	2.203	0.933	**0.450**	51.82	30.57	15.24
HOI-$\mathrm{UP}_{\mathrm{HOICodeOnly}}$	42.83	65.74	104.1	768.00	0.533	0.825	14.99	3.002	2.195	0.848	0.686	51.51	29.03	21.63
HOI-UP (ours)	**82.75**	**91.53**	**76.7**	702.60	**0.185**	**0.916**	12.69	**2.865**	2.170	**0.834**	0.585	60.41	**31.89**	**21.91**

5.2 HOI Adaptation to Unseen Object Properties

Baselines. We distinguish controller adaptation from motion adaptation, as grouped in Tab. 1. *InterMimic* fine-tunes the controller to track the property-agnostic reference under target properties, while *InterMimic+ZMP* additionally rewards balance (Vukobratović & Borovac, 2004). *InterMimic + HOI-code* retains the original motion-tracking objective and adds a reward for matching the same VLM-predicted $L + I$ HOI-code targets without adapting the kinematic HOI plan (App. A.9.3). For motion adaptation, *CoMo-HOI* extends CoMo (Huang et al., 2024) with our $L + I$ HOI-code vocabulary to generate human–object motion plans, which are subsequently tracked by fine-tuning InterMimic under the target properties (App. A.9.5). *HOI-UP$_{HOICodeOnly}$* follows our full pipeline but optimizes only the HOI-code objective during kinematic adaptation, omitting the contact, biomechanical, and regularization objectives (App. A.9.4).

Results. Tab. 1 shows that HOI-UP improves how motions respond across object-property levels while preserving contact and physical executability.

Comparison with controller adaptation. Compared with InterMimic and InterMimic+ZMP, HOI-UP improves success rate to 82.75% (versus $77.42\%/78.00\%$), with higher completion, lower tracking error, better contacts, and the lowest foot-floor error. Despite its second-best success rate and favorable biomechanical demand scores, InterMimic+ZMP's higher Minimality and lower Mon and R^2_{all} indicate that its changes across property levels are less consistently ordered, less distinguishable, and less explained by the property change. InterMimic+HOI-code reduces success to 73.83%, consistent with competing demands from tracking the unchanged reference and realizing adapted HOI-code targets. The resulting Mon and R^2_{all} remain close to those of InterMimic, indicating that the same VLM-predicted HOI-code targets yield little property response unless realized in the plan. HOI-UP instead realizes these HOI-code targets in a contact-aware plan before controller fine-tuning.

Comparison with motion adaptation. CoMo-HOI and HOI-$\mathrm{UP}_{\mathrm{HOICodeOnly}}$ achieve higher Mon than the controller-adaptation baselines, but their success falls to 23.25% and 42.83%, and their Eff drops to 51.82% and 51.51%. HOI-UP retains the highest Mon and R^2_{all} with higher Eff, while substantially improving execution and contact preservation and lowering joint torques; contact F1 reaches 0.916, compared with $0.726/0.825$. Although CoMo-HOI obtains lower Minimality, its plans are rarely executable, as reflected in its low success and contact F1. The weaker execution and contact scores of both baselines support jointly optimizing HOI-code targets with contact and biomechanical objectives, rather than relying on HOI-code-driven motion editing alone.

5.3 Qualitative Results

Physical execution across property types. Although InterMimic+ZMP achieves a comparable success rate in Tab. 1, Fig. 3 shows that its motions remain close to the reference despite changes in object properties, failing with the $30\,\mathrm{kg}$ load and the $0.75\times$ object. HOI-UP instead adapts the motion plan while preserving the reference interaction structure; it squats deeper for the heavier load and adjusts hand placement for the resized object (circled). Meaningful adaptation is also visible where the baseline succeeds. Under increased friction, HOI-UP lowers its posture and extends its arms to apply force more horizontally; under the shifted CoM, it adjusts arm and object placement for stable carrying. These examples show that HOI-UP completes more interactions with coordi-

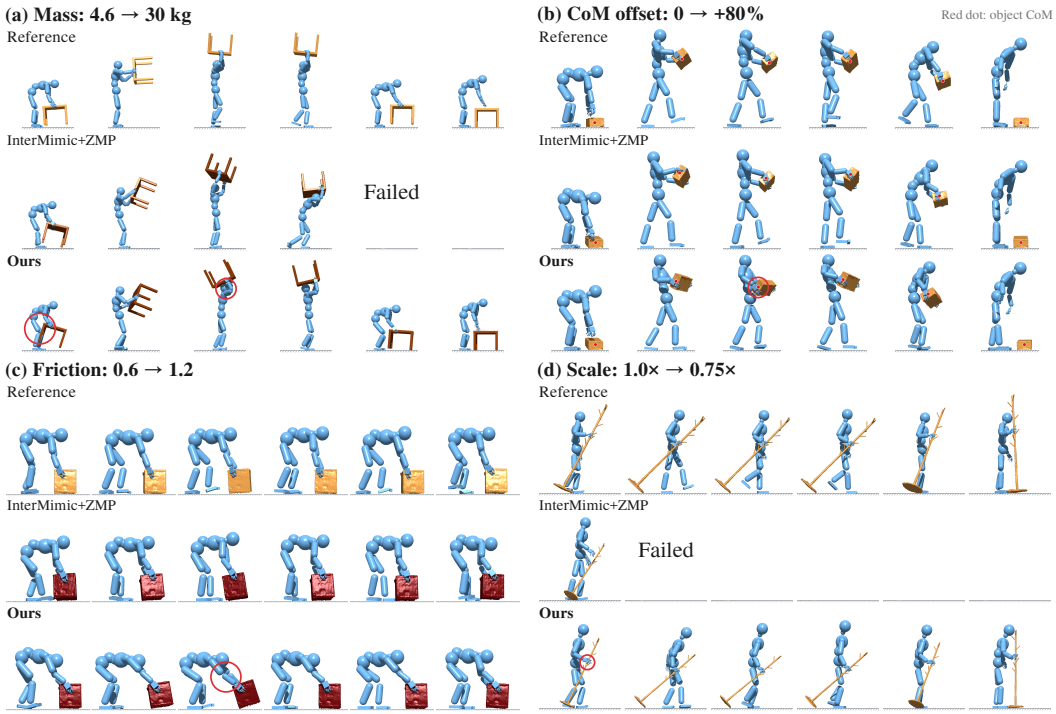


Figure 3: **Physical execution across property types.** HOI-UP vs. InterMimic+ZMP across (a) mass, (b) CoM offset, (c) friction, and (d) scale changes. Circles highlight adaptations; red dots mark the shifted CoM, darker objects are heavier, and redder objects have higher object–floor friction.

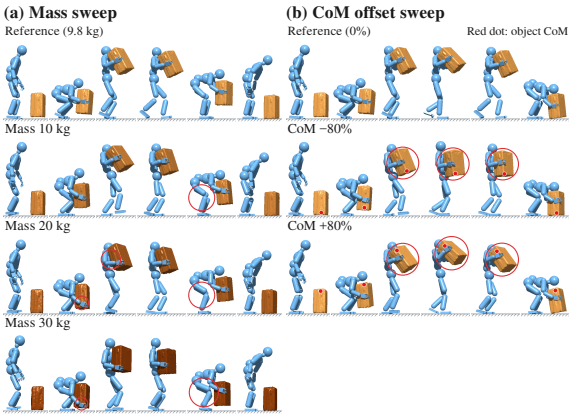


Figure 4: **Mass and CoM sweeps.** HOI-UP rollouts across mass and CoM levels.

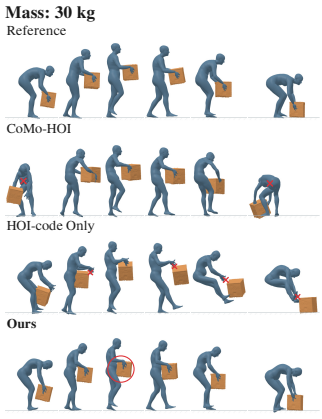


Figure 5: **Kinematic ablation.** Comparison of HOI-UP, CoMo-HOI, and HOI-code Only.

nated contacts and visible property-dependent changes, consistent with its executability, naturalness, and property-response results.

Response across object property levels. Fig. 4 shows how HOI-UP adapts the same reference across levels of a single object property, with the remaining properties fixed. At 10 kg, close to the reference mass of 9.8 kg, the motion remains nearly unchanged, whereas at 20 and 30 kg, HOI-UP progressively squats deeper and lowers its hands to support the object from underneath during lifting, and keeps the object closer to the body during carrying (circled). In the CoM sweep, arm and object placement likewise adjust to the shifted CoM. These changes retain the lifting–carrying–lowering sequence, hand–object contact, and foot-floor, consistent with the contact-related scores in Tab. 1. Thus, adaptation varies how the interaction is performed without disrupting its structure.

Table 2: **Kinematic ablation before RL fine-tuning.** Vocabulary and objective variants are evaluated on 20 reference sequences using the same evaluation descriptors.

Method	Naturalness			Biomechanical demand				Preservation		Property response		
	$\mathrm{Jit}_{\mathrm{HOI}}\downarrow$ ($\mathrm{m\,s^{-3}}$)	Foot-floor $\downarrow$ (cm)	Contact F1 $\uparrow$	$\bar{M}\downarrow$ (N m)	$C_{100}\downarrow$ (joints)	OWAS $\downarrow$	RNLE LI $\downarrow$	Minimality $\downarrow$	$\mathrm{Comp}_{\mathrm{HOI}}\downarrow$ (cm)	Eff $\uparrow$ (%)	Mon $\uparrow$ (%)	$R^2_{\mathrm{all}}\uparrow$ (%)
(a) Unadapted reference												
Reference (raw)	–	–	–	71.74	2.495	1.855	0.693	–	–	–	–	–
(b) CoMo-HOI comparison												
CoMo-HOI	296.62	3.002	0.639	76.97	2.507	1.796	0.692	0.681	31.110	47.77	20.06	8.16
(c) HOI-code vocabulary												
PoseScript	160.47	**1.111**	0.930	76.79	2.557	1.763	0.668	0.537	6.706	70.51	54.95	17.04
L	174.28	0.208	0.931	76.58	2.538	1.771	**0.663**	0.403	6.964	**71.91**	54.77	18.84
(d) Optimization objectives												
No HOI-code	170.60	1.151	0.928	76.93	2.559	**1.745**	0.679	0.457	7.757	62.75	38.13	15.38
HOI-code Only	**65.39**	2.594	0.569	**70.20**	2.525	1.836	0.684	0.411	**5.893**	62.69	**56.97**	**21.42**
Ours ($L + I$)	125.55	1.131	**0.934**	70.33	**2.420**	1.805	**0.663**	**0.371**	6.917	67.11	55.25	19.09

5.4 Ablation Studies

Tab. 2 evaluates kinematic plans before physics-based tracking. While the raw reference preserves contacts but has no property response, HOI-UP introduces property-dependent changes with the highest contact F1 and reduced joint loading. Compared with CoMo-HOI, it generates more natural and property-responsive motion. Among vocabulary variants, $L + I$ reduces jitter and joint loading and achieves lower Minimality than both L and PoseScript (Delmas et al., 2022). Although these alternatives obtain higher Eff and lower OWAS, $L + I$ better combines smoothness, joint loading, and fewer unintended changes.

Removing HOI-code guidance lowers OWAS but reduces Mon from 55.25% to 38.13%, whereas optimizing HOI-code alone lowers jitter and slightly improves Mon and R^2_{all}, but degrades contact F1 from 0.934 to 0.569. Both variants also lower Eff (62.75% and 62.69%, versus 67.11%). Despite its comparable property response, HOI-code Only shows poorer hand contact, larger foot–floor errors, and markedly weaker execution than HOI-UP in Tab. 1, supporting joint optimization for simulation-ready adaptation. Thus, HOI-code guidance promotes property-dependent changes, while the remaining objectives reconcile these changes with interaction geometry and biomechanical requirements. Fig. 5 illustrates the importance of this coordination at $30\,\mathrm{kg}$: CoMo-HOI introduces pronounced torso twisting, and HOI-code Only disrupts hand–object contact and foot support, as marked by crosses. HOI-UP instead brings the object closer to the body in the circled carrying pose while retaining coordinated contacts and the reference interaction sequence.

6 Conclusion

We formulated unseen-property HOI adaptation and introduced HOI-UP to adapt property-agnostic HOI references to unseen object properties without property-labeled motion data. HOI-UP realizes VLM-predicted phase-wise changes to compact HOI-codes through kinematic optimization that preserves the reference interaction structure, maintains balance, and limits excessive joint loading. Physics-based tracking under the HS actuation profile then refines the adapted motion. Our evaluation protocol jointly assesses property response, reference preservation, naturalness, biomechanical demand, and physical executability without ground-truth target motions. Experiments across diverse object properties show human-like, monotonic adaptations and the highest physical executability among the compared baselines. However, adaptation depends on VLM prediction quality, and our quantitative human-likeness metrics only partially capture individual variability and strategy diversity. We use simulator defaults as nominal object properties, which may differ from those underlying the reference motions. Despite these limitations, HOI-UP achieves favorable results under joint evaluation of physical executability, naturalness, biomechanical demand, and property response. These results support its potential for object-property-aware HOI motion generation and synthetic dataset construction.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, assisting with translation, interpreting results, and obtaining minor feedback on experiments. The proposed framework was proposed by the authors. For implementation, we decomposed the framework into several components with well-defined roles and prompted an LLM separately for each component. We used an LLM to implement the baseline methods such as CoMo-HOI. For writing, parts of the manuscript were first drafted by the authors and translated into English with an LLM, while other parts were written directly in English by the authors with LLM assistance limited to grammar correction. We also used an LLM to assist in interpreting experimental results. We also consulted an LLM for minor feedback on the experimental setup; the overall research methodology and experimental design were determined by the authors. Additionally, we used generative AI tools for aggregating experimental results, adjusting the layout of figures, improving the grammar and readability of the manuscript, and identifying relevant literature.

We have not used generative AI tools for generating synthetic datasets, cleaning or reformatting datasets, developing theoretical models or conceptual frameworks, proposing or refining hypotheses, or supporting qualitative or thematic data analysis. All experiments use the publicly available OMOMO dataset (Li et al., 2023) and FORCE dataset (Zhang et al., 2025), which was converted into the InterMimic representation using the official InterAct code (Xu et al., 2025a). Formulating mathematical claims and providing ingredients for or writing proofs are not applicable to this work.

We have reviewed all AI-assisted work. All LLM-generated code, including the baseline implementations, was manually reviewed by the authors, and each component of the proposed method was validated with dedicated unit tests at the time of implementation. All aggregated results were cross-checked against the raw experimental outputs. The existence of every reference suggested by an LLM was manually verified by the authors. All scientific claims in this paper were written by the authors; LLM-assisted interpretations were checked against the raw experimental results, and the translated text was reviewed by the authors to ensure it faithfully conveys the original drafts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *ICCV*, 2023.

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 *CVPR*, 2022.

R. Burgess-Limerick. Squat, stoop, or something in between? *International Journal of Industrial Ergonomics*, 2003.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023.

R. C. Cooper, L. M. Prebeau-Menezes, M. T. Butcher, and J. E. A. Bertram. Step length and required friction in walking. *Gait & Posture*, 2008.

Jieming Cui, Tengyu Liu, Ziyu Meng, Jiale Yu, Ran Song, Wei Zhang, Yixin Zhu, and Siyuan Huang. GROVE: A generalized reward for learning open-vocabulary physical skill. In *CVPR*, 2025.

B. Danneskiold-Samsøe, E. M. Bartels, P. M. Bülow, H. Lund, A. Stockmarr, C. C. Holm, I. Wätjen, M. Appleyard, and H. Bliddal. Isokinetic and isometric muscle strength in a healthy population with special reference to age and gender. *Acta Physiologica*, 2009.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: 3D human poses from natural language. In *ECCV*, 2022.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *CVPR*, 2024.

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023.

Sue Hignett and Lynn McAtamney. Rapid entire body assessment (REBA). *Applied Ergonomics*, 2000.

Katherine R. S. Holzbaur, Scott L. Delp, Garry E. Gold, and Wendy M. Murray. Moment-generating capacity of upper limb muscles in healthy adults. *Journal of Biomechanics*, 2007.

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. CoMo: Controllable motion generation through language guided pose code editing. In *ECCV*, 2024.

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

Yifeng Jiang, Tom Van Wouwe, Friedl De Groote, and C. Karen Liu. Synthesis of biologically realistic human motion using joint torque actuation. *ACM Trans. Graph.*, 2019.

Jeonguk Kang, Hanbyel Cho, Sanghyun Kang, and Donghan Koo. SplitAdapter: Load-aware humanoid loco-manipulation via factorized adaptation. arXiv preprint arXiv:2606.03297, 2026.

Osmo Karhu, Pekka Kansi, and Iikka Kuorinka. Correcting working postures in industry: A practical method for analysis. *Applied Ergonomics*, 1977.

Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3D digital humans. *ACM Trans. Graph.*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

J. J. Knapik, K. L. Reynolds, and E. Harman. Soldier load carriage: Historical, physiological, biomechanical, and medical aspects. *Military Medicine*, 2004.

S. A. Lavender, G. B. J. Andersson, O. D. Schipplein, and H. J. Fuentes. The effects of initial lifting height, load magnitude, and lifting speed on the peak dynamic L5/S1 moments. *International Journal of Industrial Ergonomics*, 2003.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 2023.

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

Lei Li and Angela Dai. HOI-PAGE: Zero-shot human-object interaction generation with part affordance guidance. In *ICML*, 2026.

Shujia Li, Haiyu Zhang, Xinyuan Chen, Yaohui Wang, and Yutong Ban. GenHOI: Generalizing text-driven 4D human-object interaction synthesis for unseen objects. arXiv preprint arXiv:2506.15483, 2025.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023.

Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, 2024.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Jim Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *ICLR*, 2024.

Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Benchmarking counterfactual image generation. In *NeurIPS Datasets and Benchmarks Track*, 2024.

A. Mexi, I. Kafetzakis, M. Korontzi, D. Karagiannakis, P. Kalatzis, and D. Mandalidis. Effects of load carriage on postural control and spatiotemporal gait parameters during level and uphill walking. *Sensors*, 2023.

Sanghyeok Nam, Byoungjun Kim, Daehyung Park, and Tae-Kyun Kim. Dynamic full-body motion agent with object interaction via blending pre-trained modular controllers. In *CVPR Findings*, 2026.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 2018.

Gerard Pons-Moll, David J. Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, Xiao Ma, Liang Pan, and Ziwei Liu. InsActor: Instruction-driven physics-based characters. In *NeurIPS*, 2023.

Elliott J. Rouse, Robert D. Gregg, Levi J. Hargrove, and Jonathon W. Sensinger. The difference between stiffness and quasi-stiffness in the context of biomechanical modeling. *IEEE Transactions on Biomedical Engineering*, 2013.

Elliott J. Rouse, Levi J. Hargrove, Eric J. Perreault, and Todd A. Kuiken. Estimation of human ankle impedance during the stance phase of walking. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2014.

O. D. Schipplein, T. E. Reinsel, G. B. J. Andersson, and S. A. Lavender. The influence of initial horizontal weight placement on the loads at the lumbar spine while lifting. *Spine*, 1995.

Na Jin Seo, Thomas J. Armstrong, and Justin G. Young. Effects of handle orientation, gloves, handle friction and elbow posture on maximum horizontal pull and push forces. *Ergonomics*, 2010.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024.

Yuto Shibata, Kashu Yamazaki, Lalit Jayanti, Yoshimitsu Aoki, Mariko Isogawa, and Katerina Fragkiadaki. Learning to assist: Physics-grounded human-human control via multi-agent reinforcement learning. In *CVPR*, 2026.

Soshi Shimada, Franziska Mueller, Jan Bednarik, Bardia Doosti, Bernd Bickel, Danhang Tang, Vladislav Golyanik, Jonathan Taylor, Christian Theobalt, and Thabo Beeler. MACS: Mass conditioned 3D hand and object motion synthesis. In *3DV*, 2024.

K. Shoji, K. Nakayama, M. Shiouchi, and Y. Manabe. Load-dependent increase in lumbar kyphosis is associated with posterior pelvic tilt during deadlift. *Frontiers in Sports and Active Living*, 2025.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In *ECCV*, 2022.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *ICLR*, 2023.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H. Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *ICLR*, 2025.

Jaap H. van Dieën, Marco J. M. Hoozemans, and Huub M. Toussaint. Stoop or squat: A review of biomechanical studies on lifting technique. *Clinical Biomechanics*, 1999.

Miomir Vukobratović and Branislav Borovac. Zero-Moment Point—thirty five years of its life. *International Journal of Humanoid Robotics*, 2004.

Ruiyan Wang, Lin Zuo, Zonghao Lin, Qiang Wang, Zhengxue Cheng, Rong Xie, Jun Ling, and Li Song. PA-HOI: A physics-aware human and object interaction dataset. In *ACM MM*, 2025.

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.

Thomas R. Waters, Vern Putz-Anderson, Arun Garg, and Lawrence J. Fine. Revised NIOSH equation for the design and evaluation of manual lifting tasks. *Ergonomics*, 1993.

Yujiang Xiang, Jasbir S Arora, Salam Rahmatalla, Timothy Marler, Rajankumar Bhatt, and Karim Abdel-Malek. Human lifting simulation using a multi-objective optimization approach. *Multibody System Dynamics*, 2010.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

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

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3D human-object interaction generation. In *CVPR*, 2025a.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025b.

T. Yamaguchi, M. Yano, H. Onodera, and K. Hokkirigawa. Kinematics of center of mass and center of pressure predict friction requirement at shoe–floor interface during walking. *Gait & Posture*, 2013.

Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, and Angelica Lim. MotionScript: Natural language descriptions for expressive 3D human motions. In *IROS*, 2025.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: Real-time 3D human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.*, 2021.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis. In *CoRL*, 2023.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *ICCV*, 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 *3DV*, 2025.

A APPENDIX

A.1 INTERACTION-PHASE SEGMENTATION

We segment each reference motion $\mathcal{H}$ into interaction phases using contact states, object-ground support, object velocity, and relative human-object states. Object-ground support is determined from the lowest object-mesh vertex $z_{\min}[t]$. Some captured sequences contain a persistent vertical artifact, causing an object that is pushed on the floor to remain above $z = 0$. Therefore, we use the 5th percentile of $z_{\min}$ within each sequence as the object's resting height. When $z_{\text{floor}} \leq 0.30$ m, a frame is considered ground-supported if $z_{\min} \leq z_{\text{floor}} + 0.03$ m. Sequences with $z_{\text{floor}} > 0.30$ m are treated as having no object-ground support.

Tab. 3 defines the interaction phases and their frame-level assignment rules. v_{xy} and v_z denote the horizontal speed and signed vertical velocity of the object, respectively. We define s_{front} as the horizontal cosine similarity between the human heading and the human-to-object direction, and s_{move} as that between the human heading and the object velocity direction. Clothesstands, floorlamps, and tripods are treated as predefined drag-only objects, since they cannot be pushed or pulled on the floor. The same thresholds are applied to all sequences.

After the segmentation, we refine the raw frame-level labels using temporal post-processing to remove short or inconsistent segments, as described in Alg. 1. Remaining unknown intervals are absorbed by their neighboring phases, while stationary contact without clear object motion is labeled as grasping.

Table 3: Definitions and frame-level assignment rules for the interaction phases. Unknown is removed in the post-processing.

Phase	Definition	Assignment rule
non-interacting	No active object interaction	No contact
lifting	Lift from the ground	Contact, no ground support, $v_z > 0.1$
carrying	Object held away from the ground	Contact, no ground support, $\lvert v_z \rvert \leq 0.1$
lowering	Put down on the ground	Contact, no ground support, $v_z < -0.1$
pushing	Push object on the ground	Contact, ground support, $v_{xy} \geq 0.1$, $s_{\text{front}} > 0.5$, $s_{\text{move}} > 0.3$
pulling	Pull object on the ground	Contact, ground support, $v_{xy} \geq 0.1$, $s_{\text{move}} < -0.3$
dragging	Non-frontal object motion on the ground	Contact, ground support, $v_{xy} \geq 0.1$, and either $s_{\text{front}} \leq 0.5$, $s_{\text{move}} > 0.3$, or a drag-only object
grasping	Contact without clear motion	Assigned from remaining unknown intervals during post-processing
unknown	None of the above	Remaining ground-supported contact

A.2 HOI-CODE DEFINITIONS

Our HOI-code candidate vocabulary contains 80 continuous channels — 46 human-body geometry codes from PoseScript (Delmas et al., 2022) and 34 codes proposed in this paper for object states and human-object relations. The 34 additional codes describe angles, distances, relative positions, orientations, ground-relative heights, and temporal rates, as listed in Tab. 4. Unlike CoMo (Huang et al., 2024) which converts PoseScript codes into categorical levels for LLM-based editing, we retain their continuous values to preserve fine variations in the reference trajectories and use them directly as targets for kinematic HOI optimization.

For the main experiments, we use the 17-channel subset in Tab. 5, which consists of 6 PoseScript channels and 11 additional HOI codes. The literature codes L describe handling posture and human–object configuration, whose load-dependent changes are reported in lifting and carrying studies (Sec. 2). The interaction codes I describe whole-body support, locomotion, and object-motion states needed to realize and preserve the interaction. These subsets are defined without property-labeled motion data.

Algorithm 1 Interaction-phase Segmentation Post-processing

Require: raw per-frame phase labels $\tilde{\mathbf{q}}$

$\mathbf{q} \leftarrow$ FILTERTINYSEGMENTS($\tilde{\mathbf{q}}$, 6) ▷ drop < 0.2 s peaks
$\mathbf{q} \leftarrow$ FILLTRIVIALUNKNOWN($\mathbf{q}$) ▷ merge noisy segments
$\mathbf{q} \leftarrow$ REPAIRNONINTERACTINGGAPS($\mathbf{q}$) ▷ clean noise at interaction start/end
$\mathbf{q} \leftarrow$ FILLTRIVIALUNKNOWN($\mathbf{q}$)
$\mathbf{q} \leftarrow$ SUPPRESSSTATICSLIDE($\mathbf{q}$, $0.2\,m$) ▷ remove non-moving slides
$\mathbf{q} \leftarrow$ FILLTRIVIALUNKNOWN($\mathbf{q}$)
$\mathbf{q} \leftarrow$ SUPPRESSFAKELIFTLOWER($\mathbf{q}$) ▷ don't mistake object wobble for lift/lower
$\mathbf{q} \leftarrow$ FILLTRIVIALUNKNOWN($\mathbf{q}$)
$\mathbf{q} \leftarrow$ SUPPRESSFAKEPUSHPULLDRAG($\mathbf{q}$) ▷ clean noise around lift/lower
$\mathbf{q} \leftarrow$ FILLTRIVIALUNKNOWN($\mathbf{q}$)
$\mathbf{q} \leftarrow$ FILTERTINYSEGMENTS($\mathbf{q}$, 15) ▷ drop < 0.5 s residuals
$\mathbf{q} \leftarrow$ FILLUNKNOWN($\mathbf{q}$) ▷ commit residuals
return $\mathbf{q}$

Function roles:
• FILTERTINYSEGMENTS($\mathbf{q}, \kappa$): merge each segment shorter than κ frames into its neighbours if they agree, else set unknown.
• FILLTRIVIALUNKNOWN($\mathbf{q}$): relabel an unknown segment from its neighbours if they agree.
• REPAIRNONINTERACTINGGAPS($\mathbf{q}$): fill a non-interacting gap flanked by a compatible motion pair (prev $\in$ {lower, push, pull, drag}, next $\in$ {lift, push, pull, drag}).
• SUPPRESSSTATICSLIDE($\mathbf{q}, \xi$): set a push/pull/drag segment to unknown if net horizontal object displacement $< \xi$.
• SUPPRESSFAKELIFTLOWER($\mathbf{q}$): set a lift/lower segment to carrying if the object does not start/end at the floor.
• SUPPRESSFAKEPUSHPULLDRAG($\mathbf{q}$): set a slide block to lifting between (non-interacting, lifting) and to lowering between (lowering, non-interacting); it removes the noise before lifting and after lowering the object.
• FILLUNKNOWN($\mathbf{q}$): commit each residual unknown to grasping if flanked by non-interacting, else its neighbour motions.

A.3 VLM-BASED ADAPTATION DETAILS

We use Claude Opus 5 to predict HOI-code adaptations under changes in object properties. Each reference motion and target property level are independently queried. The visual input consists of rendered reference-motion images sampled every 15 frames from 30 fps video, i.e., 0.5 s intervals. The textual input includes source and target object properties, interaction phase labels, phase-wise HOI-code summaries, and code definitions, units, and sign conventions. The prompt does not explicitly specify the direction of change for individual HOI-codes for object property changes. Instead, the VLM is instructed to predict phase-wise adaptations appropriate for the target physical properties while leaving the codes unchanged when the expected response is uncertain. The model outputs the predicted phase-wise code ranges and confidence for each phase in JSON format. Low-confidence changes are kept at their reference values.

The predicted code range is converted into a target code trajectory by applying a phase-wise affine transformation to the reference trajectory, as specified in Eq. 2. To ensure continuity across interaction phases, the phase-wise predictions are linearly interpolated. Since the kinematic optimizer preserves the reference length, changes in velocities are handled separately through temporal rescaling. Scaling factors are estimated from the predicted changes in velocity-related codes. This rescaling stretches or compresses the timeline, enabling slow lifting, carrying, pushing, or pulling.

A.4 KINEMATIC HOI OPTIMIZATION DETAILS

We first fit the reference human motion and body shape to SKEL, then keep the fitted shape fixed during adaptation. We optimize bounded cubic B-spline corrections to the human root, articulated pose, and object pose, subject to anatomical joint limits. Spatial rescaling and duration calibration precede optimization; the resulting timing and VLM-predicted targets remain fixed during the

Table 4: Definitions of the 34 object-related and human-object HOI-code candidates introduced in this work. p_X: 3D position of joint X; c_o: object centroid; $\mathbf{f}, \mathbf{r}, \mathbf{z}$: body-forward, body-right, world-up directions; Π_{xy}: horizontal projection; $z(\cdot)$: vertical coordinate; z_{floor}: per-sequence floor height; $\angle(a, b, c)$: interior angle at b; $\angle_{xy}$: horizontal angle; $\widehat{x}$: normalized vector; dot: time derivative; ω_o: object angular velocity; θ_{body}: body heading; L/R: left/right; sh: shoulder; spine1/spine3: lower (pelvis-side)/upper (chest-side) spine joints.

Code	Unit	Definition
Literature Codes		
object_horizontal_distance_from_feet	m	$\|\Pi_{xy}(c_o - \frac{1}{2}(p_{\text{L ankle}} + p_{\text{R ankle}}))\|$
trunk_lean	deg	$\text{atan2}(\langle u, \mathbf{f}\rangle, \langle u, \mathbf{z}\rangle)$, where $u = \widehat{p_{\text{neck}} - p_{\text{pelvis}}}$
trunk_side_lean	deg	$\arcsin\langle u, \mathbf{r}\rangle$, where $u = \widehat{p_{\text{neck}} - p_{\text{pelvis}}}$
body_twist	deg	Signed horizontal angle from the hip line to the shoulder line
object_off_center_angle	deg	$\lvert\angle_{xy}(\mathbf{f}, c_o - p_{\text{pelvis}})\rvert$
object_height_above_floor	m	$z(c_o) - z_{\text{floor}}$
Interaction Codes		
step_length	m	$\lvert\langle p_{\text{L ankle}} - p_{\text{R ankle}}, \mathbf{f}\rangle\rvert$
stance_width	m	$\langle p_{\text{R ankle}} - p_{\text{L ankle}}, \mathbf{r}\rangle$
pelvis_height_above_floor	m	$z(p_{\text{pelvis}}) - z_{\text{floor}}$
body_travel_speed	m/s	$\|\Pi_{xy}\dot{p}_{\text{pelvis}}\|$
object_vertical_speed	m/s	$\frac{d}{dt}\left(z(c_o) - z_{\text{floor}}\right)$
object_gap_to_head	m	Minimum object-surface gap to a head sphere with radius 0.09 m
object_gap_to_chest	m	Minimum object-surface gap to the spine3-to-neck capsule with radius 0.13 m
object_gap_to_belly	m	Minimum object-surface gap to the pelvis-to-spine3 capsule with radius 0.13 m
object_gap_to_left_thigh	m	Minimum object-surface gap to the left hip-to-knee capsule with radius 0.07 m
object_gap_to_right_thigh	m	Minimum object-surface gap to the right hip-to-knee capsule with radius 0.07 m
left_hip_angle	deg	$\angle(p_{\text{spine1}}, p_{\text{L hip}}, p_{\text{L knee}})$
right_hip_angle	deg	$\angle(p_{\text{spine1}}, p_{\text{R hip}}, p_{\text{R knee}})$
object_angle_from_trunk_axis	deg	$\angle(p_{\text{neck}}, p_{\text{pelvis}}, c_o)$
object_horizontal_distance_from_body	m	$\|\Pi_{xy}(c_o - p_{\text{pelvis}})\|$
object_3d_distance_from_shoulders	m	$\|c_o - \frac{1}{2}(p_{\text{L sh}} + p_{\text{R sh}})\|$
object_3d_distance_from_torso	m	$\|c_o - \frac{1}{3}(p_{\text{pelvis}} + p_{\text{spine3}} + p_{\text{neck}})\|$
object_forward_offset	m	$\langle c_o - p_{\text{pelvis}}, \mathbf{f}\rangle$
object_sideways_offset	m	$\langle c_o - p_{\text{pelvis}}, \mathbf{r}\rangle$
object_height_above_pelvis	m	$\langle c_o - p_{\text{pelvis}}, \mathbf{z}\rangle$
object_height_above_shoulders	m	$\langle c_o - \frac{1}{2}(p_{\text{L sh}} + p_{\text{R sh}}), \mathbf{z}\rangle$
hip_line_tilt	deg	$\arcsin\langle \widehat{p_{\text{R hip}} - p_{\text{L hip}}}, \mathbf{z}\rangle$
head_tilt	deg	$\arccos\langle \widehat{p_{\text{head}} - p_{\text{neck}}}, \mathbf{z}\rangle$
object_bottom_height_above_floor	m	Lowest object bounding-box height minus z_{floor}
object_horizontal_speed	m/s	$\|\Pi_{xy}\dot{c}_o\|$
object_turn_rate	deg/s	$\langle \omega_o, \mathbf{z}\rangle$
object_tip_rate	deg/s	$\|\Pi_{xy}\omega_o\|$
object_rotation_speed	deg/s	$\|\omega_o\|$
object_turn_rate_vs_body	deg/s	$\langle \omega_o, \mathbf{z}\rangle - \dot{\theta}_{\text{body}}$

Table 5: The 17 HOI-code channels used in the main experiments. Full definitions are provided in PoseScript (Delmas et al., 2022) and Tab. 4.

Set	Source	Channels	Count
Literature	PoseScript	left_shoulder_angle, right_shoulder_angle, left_elbow_angle, right_elbow_angle, left_knee_angle, right_knee_angle	6
Literature	Ours	trunk_lean, trunk_side_lean, body_twist, object_horizontal_distance_from_feet, object_off_center_angle, object_height_above_floor	6
Interaction	Ours	step_length, stance_width, pelvis_height_above_floor, body_travel_speed, object_vertical_speed	5
		Total	17

solve. HOI-code tracking uses the mean squared target error over active frames, normalized by each channel's definition scale.

Reference annotations determine hand attachments and stance intervals. Object-local hand anchors couple the object pose to the hands, while a task-space penalty maintains the foot-sole geometry

around fixed stance-episode anchors. Under scale changes, we scale the object mesh, recompute its signed distance field (SDF), and recompute the hand anchors accordingly so that the grasp remains on the scaled surface. The object trajectory is also shifted vertically so that the lowest point of the scaled mesh matches that of the original object. Contact activation is smoothed across transitions. Balance combines dynamic and gravity-only ZMP penalties with an additional support term computed from the current foot geometry. Joint loads use the target mass, center of mass, friction, and current motion. Load penalties act above fixed reference-derived thresholds, with a linear tail for large normalized violations. Additional regularization limits penetration, task deviation, and temporal variation.

Each reference–property pair is optimized independently with Adam, retaining the iterate with the lowest complete objective. Tab. 6 summarizes the shared settings and principal component weights of Eq. 3. The grouped contact and regularization objectives contain separately weighted components; weights are fixed across sequences and target property levels. The selected trajectories are exported at 30 fps for physics-based tracking.

Table 6: Kinematic optimization settings and principal objective weights. Weights apply per component; paired values follow the order of the listed terms.

Parameter	**Value**
Optimizer / iterations	Adam / 400
Learning rate: pose / root / object	7×10^{-3} / 2×10^{-3} / 1.5×10^{-3}
Gradient norm clipping	5
B-spline degree / control points	3 / 32
Duration scale / candidate count	$[0.8, 1.35]$ / 81
Code scale: angle / distance / speed	10° / 0.1 m / 0.3 m/s
Grasp / stance-sole tolerance	5 mm / 5 mm
Contact activation transition	0.25 s
Objective component	**Weight**
HOI-code tracking (each channel)	1.8×10^5
Grasp position / palm orientation (each axis)	4×10^4 / 2×10^2
Stance-sole task-space error	4×10^6
Foot position / height / sliding / floor	50 / 50 / 800 / 200
Dynamic / gravity-only ZMP	8×10^4 / 8×10^4
Current-sole support balance	8×10^4
Normalized load excess (each component)	10^4
Body–object clearance / penetration	4×10^5 / 4×10^5
Self-penetration	1.2×10^5
Reference joint position / root / pose	50 / 50 / 250
Joint-acceleration / pose-velocity smoothing	1 / 0.5
Object translation / rotation smoothing	3 / 3
Object linear / angular / CoM acceleration	10^4 / 10^4 / 10^4

A.5 Human-Strength-Informed Actuation Profile

Profile settings. Tab. 7 lists torque caps and stiffnesses for 51 actuated joints (153 DOFs). We use $\alpha = K_{\mathrm{ankle},y}/\bar{\tau}_{\mathrm{ankle},y} \approx 3.498\,\mathrm{rad}^{-1}$ for all axes, with $K_{\mathrm{ankle},y} = 444.432\,\mathrm{N\,m/rad}$. Bilateral joints share settings, the Finger row covers 30 joints, and damping is $D_{j,a} = 0.1\,\mathrm{s} \times K_{j,a}$ throughout. All methods use this fixed profile across reference motions and target object-property conditions.

Validation. Uniform uses the same torque cap on all 153 actuator axes, set to the root-mean-square HS cap to match the sum of squared torque limits, while retaining the HS stiffness and damping. We compare HS with Uniform using six upper-limb moments rescaled to avatar mass (Holzbaur et al., 2007) and pulling-force measurements under study-derived force directions and simplified contact assumptions for the bare-hand, parallel-rubber condition (Seo et al., 2010). Tab. 8 shows that HS better matches joint moments and the extended-to-flexed pulling-capacity ratio, whereas Uniform has lower normalized pull-force RMSE. These comparisons assess selected strength characteristics rather than every actuator axis.

Table 7: HS torque caps and stiffnesses along local simulator axes, rounded to three decimal places. Values apply to each of the n joints represented by a row; † marks heuristic caps.

		Torque cap $\bar{\tau}$ (N m)			Stiffness K (N m/rad)		
Joint(s)	n	x	y	z	x	y	z
Hip	2	182.028	282.729	60.428	636.743	989.000	211.378
Knee	2	40.000†	215.805	30.000†	139.922	754.896	104.941
Ankle	2	40.000†	127.052	30.000†	139.922	444.432	104.941
Toe	2	10.000†	30.000†	10.000†	34.980	104.941	34.980
Torso, Spine, Chest	3	180.000†	305.440	120.000†	629.648	1068.443	419.765
Neck, Head	2	30.000†	40.000†	25.000†	104.941	139.922	87.451
Thorax, Shoulder	4	69.734	45.320	71.773	243.934	158.531	251.066
Elbow	2	70.000†	20.000†	56.628	244.863	69.961	198.088
Wrist	2	20.427	10.000†	10.000†	71.456	34.980	34.980
Finger	30	2.000†	0.500†	2.000†	6.996	1.749	6.996

Table 8: **HS profile validation.** Normalized RMSE is $100\|\hat{\mathbf{y}} - \mathbf{y}\|_2/\|\mathbf{y}\|_2$, where $\hat{\mathbf{y}}$ and $\mathbf{y}$ are predictions and literature means for six joint moments or two pulling forces. Ratio error is the absolute relative percentage error for the extended-to-flexed pulling-capacity ratio. Within-one-SD counts use the reported standard deviations.

Metric	Uniform	HS profile
Joint-moment RMSE (%) ↓	76.19	**15.37**
Moments within one SD ↑	3/6	**5/6**
Pull-force RMSE (%) ↓	**9.71**	18.09
Ratio relative error (%) ↓	14.26	**5.74**

A.6 RL Fine-tuning Details

We fine-tune the released student policy of InterMimic (Xu et al., 2025b) in Isaac Gym to obtain physics-based tracking controllers. Since the pretrained policy was trained with an actuator profile different from our HS profile, we adopt a two-stage fine-tuning procedure to adapt it to the human-strength-informed actuation profile.

In stage 1, we retain the pretrained policy weights and training configurations while resetting the optimizer and learning-rate schedule. We then fine-tune the policy on OMOMO reference motions without modifying the objects' physical properties. All other training settings, reward functions, and termination criteria follow InterMimic without any changes. We use the checkpoint at 3,000 epochs, where the performance reaches a plateau, as the base policy for the HS actuation profile.

In stage 2, we independently fine-tune the controllers for each method from the same stage 1 checkpoint. Each controller is trained simultaneously on all target motions derived from the 30 OMOMO sequences in the evaluation set. These target motions are assigned in a round-robin manner across 8,192 parallel environments, with the corresponding target object properties applied in simulation.

All controllers are trained using PPO with the hyperparameters summarized in Tab. 9. To ensure a consistent comparison, all methods share the same base checkpoint and training configuration, and each is trained until the training reward plateaus. Motion optimization is performed using two NVIDIA RTX 4090 GPUs, while RL fine-tuning is conducted on two NVIDIA A100 80 GB GPUs.

A.7 Evaluation Protocol Details

Metrics follow Sec. 4, with evaluation descriptors shared across methods. Combined human–object scores use $A_{\mathrm{HOI}} = (22A_h + A_o)/23$, with 1,024 object-surface points for tracking and eight bounding-box corners for jitter and Composition. Loading metrics use nine selected joints, distinguishing simulator torques from model-estimated moments. Rollout-level scores are averaged over trials within each condition and then equally over valid conditions; kinematic $\bar{M}$ and C_{100} first aver-

Table 9: Training configuration for RL fine-tuning.

Parameter	Value
Simulator	Isaac Gym (PhysX)
Simulation / control frequency	60 / 30 Hz
Episode length	300 control steps
Parallel environments	8,192
Rollout horizon	32
Mini-batch size	16,384
Mini-epochs	6
Optimizer	Adam
Learning rate	2×10^{-5}
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
PPO clipping (ϵ)	0.2
Stage 1 epochs	3,000
Stage 2 epochs	3,500–4,500

age conditions within each sequence and then weight sequences equally. Simulation response scores use matched phase observations across methods and property levels: Eff includes reached phases, whereas Minimality, Mon, and R^2_{all} require completed phases. RNLE LI uses finite, in-range frames with frequency and coupling multipliers fixed to one.

Execution and naturalness. For observed/target lengths $T_e/T_e^\star$, success and completion are $100\langle \mathbf{1}[T_e = T_e^\star]\rangle_e$ and $100\langle T_e/T_e^\star\rangle_e$. Tracking error averages Euclidean distances to each method's input plan over 22 joints and 1,024 corresponding object-surface points (mm). Jitter uses $\mathrm{Jit}(x) = \langle \|\Delta^3 x\|_2\rangle/\Delta t^3$, with unsmoothed third differences, 22 joints, and eight object corners. Foot–floor error is $100\left\langle \left|\min(z_t^L, z_t^R) - h\right|\right\rangle_t$ (cm), using toe heights and each output's stationary-toe floor estimate h (Li et al., 2024). Contact F1 is $2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$, comparing either-hand contact within 5 cm of object vertices against aligned reference labels; doubly ambiguous frames are excluded.

Biomechanical demand. For nine joint magnitudes u_{tj} (torso/lumbar and bilateral shoulder, elbow, hip, knee), $\bar{U} = \langle u_{tj}\rangle_{t,j}$ and $C_{100} = \langle \sum_j \mathbf{1}[u_{tj} > 100\,\mathrm{N\,m}]\rangle_t$. $\bar{\tau}_{\text{sim}}$ uses recorded torque-vector norms on observed interaction frames; $\bar{M}$ uses target-property anatomical moment estimates on shared reference-contact frames. OWAS averages published posture–load classes over handling frames (Karhu et al., 1977). RNLE uses $\mathrm{LI} = m/(23\,HM\,VM\,DM\,AM)$ with standard geometry multipliers and $FM = CM = 1$ (Waters et al., 1993), averaging only finite, in-range values during lifting/carrying/lowering.

Preservation. Composition is default-output reconstruction error against the original reference (cm), combining 22 joints and eight object corners after temporal correspondence, without spatial alignment. For Minimality, d_ℓ concatenates phase-mean reference-relative changes in the same 12 literature descriptors L, normalized by frozen reference ranges $P_{95} - P_5$. Fit $\hat{d}_\ell = a + bv_\ell$ across $N_{\text{lv}} > 2$ physical property levels, including the default output:

$$\text{Minimality} = \frac{\sum_\ell \|d_\ell - \hat{d}_\ell\|_2^2/(N_{\text{lv}}-2)}{\sum_\ell \|d_\ell - \bar{d}\|_2^2/(N_{\text{lv}}-1)}. \tag{5}$$

Scores average valid sequence–property ladders; zero variation is undefined. Minimality measures unexplained response variation, not edit magnitude.

Effectiveness and monotonicity. Both metrics use the same 12 literature descriptors L. Eff fits mass slopes b_k at 10/20/30 kg against frozen FORCE directions b_k^F and correlation/consistency weights

w_k (Zhang et al., 2025):

$$\begin{aligned} \text{Eff} &= 100 \frac{\sum_{k \in L} w_k \, \mathbf{1}[\text{sgn}(b_k) = \text{sgn}(b_k^F)] \, \mathbf{1}[k \text{ responds}]}{\sum_{k \in L} w_k}, \\ \text{Mon} &= 100 \max_{\sigma \in \{-1,1\}} \binom{N_{\text{lv}}}{2}^{-1} \sum_{i<j} B_{ij}^{\sigma}. \end{aligned} \tag{6}$$

A response requires normalized range $> 10^{-5}$ and slope magnitude $> 10^{-10}$. For each descriptor over N_{lv} changed levels, B_{ij}^{σ} indicates that normalized changes $(v_i, d_i), (v_j, d_j)$ and the original anchor $(v_0, 0)$ are strictly monotone in direction σ (tolerance 10^{-5}). Eff averages sequences within lift/carry/lower/push/pull, then phases equally; Mon averages descriptors, phases, properties, then sequences. Simulation Eff includes shared reached-phase prefixes (at least five pooled frames per method/level/phase); Minimality, Mon, and R^2 require shared completed phases across methods and levels.

Property separation. Five blocks describe pelvis/heading-relative pose, world-space object corners, pelvis height, phase-start-aligned pelvis trajectory, and speed/duration. Positions use 30-Hz resampling, 6-Hz zero-phase filtering, and 101 samples per phase; object corners retain default size. Changes from each method's default output (log-ratios for speed/duration) use frozen reference block scales. After averaging repeats/occurrences within sequence–level–phase, for each property–phase group g,

$$R_g^2 = \frac{\sum_{\ell} n_{\ell} \|\bar{z}_{\ell} - \bar{z}\|_2^2}{\sum_{s,\ell} \|z_{s\ell} - \bar{z}\|_2^2}, \qquad R_{\text{all}}^2 = 100 \frac{\sum_g N_g R_g^2}{\sum_g N_g}. \tag{7}$$

Here $z_{s\ell}$ are normalized features, n_{ℓ} level counts, and $N_g = \sum_{\ell} n_{\ell}$. Groups require all levels and two reference sequences; zero variation gives zero. CoM uses body-relative left/right offsets within 45° of the lateral axis; insufficient shared support excludes CoM from the main-table aggregate. Mon and R^2 quantify ordering/separation; Eff additionally tests direction.

A.8 Evaluation Data Details

Evaluation set. Since each sequence is adapted into many object property variations, evaluating on every sequence is computationally limited. Therefore, we randomly sample 30 sequences from OMOMO (Li et al., 2023) dataset and evaluate every property variation of each sampled sequence. The set spans 14 subjects and 11 object categories, with no object occupying more than 20% of the evaluation set. It also covers diverse interaction scenarios such as lifting, carrying, lowering, pushing, pulling, dragging, and their combinations. The property variations include three target masses (10, 20, and 30 kg), two scale factors (0.75× and 1.25×), two center-of-mass offsets ($\pm 80\%$), and two object-floor friction coefficients (0.1 and 1.2). For center-of-mass offset variants, we compute the object's minimal bounding box and shift the center of mass along the direction of its longest edge. Friction variations are applied only to the 15 reference sequences that contain pushing, pulling, or dragging, where floor friction directly affects the interaction. This yields a total of 240 sequence-property combinations across the 30 reference HOI motions (90 mass, 60 CoM offset, 60 scale, and 30 friction). Objects and statistics of the evaluation are illustrated in Figs. 6–7.

Evaluation scale in prior works. The evaluation scale in this work is in line with prior zero-shot and out-of-distribution interaction studies. FORCE (Zhang et al., 2025) evaluates on 30 test sequences covering 8 objects with 3–6 resistance levels. SplitAdapter (Kang et al., 2026) evaluates humanoid loco-manipulation under 9 load conditions (3 masses × 3 heights). AssistMimic (Shibata et al., 2026) uses 40 clips under perturbed body dynamics. Among generative methods, HOI-PAGE (Li & Dai, 2026) uses 21 text prompts, and CoMo (Huang et al., 2024) evaluates 20 editing scenarios in its user study. Our comparison spans four object properties, namely, mass, friction, scale, and CoM offset, each varied across multiple levels on the same reference motion set.

A.9 Baseline Implementation

A.9.1 InterMimic

InterMimic adapts only the controller to the target properties ϕ^{tgt} in the physics simulator, while tracking the property-agnostic HOI reference $\mathcal{H}$. Under scale variations, the scaled object would

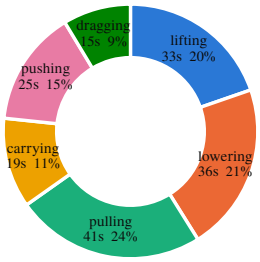


(a) 11 Objects from OMOMO dataset

Object	# Seq.	Mass (kg)
trashcan	1	0.51
floorlamp	1	1.04
plasticbox	4	1.08
monitor	1	1.11
smalltable	2	1.27
whitechair	3	1.56
woodchair	4	1.77
smallbox	2	2.63
largetable	6	4.60
largebox	2	7.36
suitcase	4	9.82
Total	30	–

(b) Sequences and mass per object

Figure 6: Objects in the 30 evaluation sequences. (a) The 11 object categories. (b) Number of evaluation sequences per object and its default mass, ordered from light to heavy (0.51–9.82 kg). No single object accounts for more than 20% of the sequences.



(a) Action tags by frame total

Subject	# Seq.	Subject	# Seq.
sub1	8	sub8	1
sub2	2	sub9	1
sub3	1	sub10	1
sub4	2	sub12	1
sub5	2	sub13	1
sub6	2	sub14	3
sub7	2	sub16	3
Total: 30 sequences, 14 subjects			

(b) Sequences per subject

Figure 7: Interactions and subjects in the 30 evaluation sequences. (a) Share of interaction time per action tag; interaction tags are well represented. (b) Number of evaluation sequences per subject.

float above the ground when shrunk or penetrate it when enlarged, since $\mathcal{H}$ retains the trajectory of the original object. Therefore, we shift the reference object height per frame so that the lowest point of the scaled mesh matches that of the original one. The reward is the multiplicative tracking reward of InterMimic (Xu et al., 2025b) with its released weights, $r_t = r_t^{\text{track}}$.

A.9.2 InterMimic+ZMP

InterMimic+ZMP tracks the same references as InterMimic, and adds a balance reward based on the zero-moment point (ZMP) (Vukobratović & Borovac, 2004) to the tracking reward,

$$r_t = r_t^{\text{track}} + w_{\text{zmp}}\,\mathbf{1}\,[\text{hold}]_t \exp\big(-(\delta_t/\sigma_{\text{zmp}})^2\big), \tag{8}$$

where δ_t is the horizontal distance from the ZMP to the foot support polygon, which is zero inside the polygon. The ZMP is computed from the gravitational and inertial forces of both the human and the object, using the object's mass, center of mass, and inertia under ϕ^{tgt}. The support polygon is the convex hull of the soles of the feet in contact with the ground. The gate $\mathbf{1}\,[\text{hold}]_t \in \{0, 1\}$ activates the reward only while the human is handling the object, i.e., when a hand is in contact with the object and the object is lifted or moving. We use $w_{\text{zmp}} = 0.25$ and $\sigma_{\text{zmp}} = 0.1$ m.

A.9.3 InterMimic+HOI-code

InterMimic+HOI-code tracks the same references as InterMimic while realizing the VLM-predicted HOI-code targets $\hat{\mathbf{c}}[t]$ from Sec. 3.2 through an additional reward instead of kinematic optimization

in Sec. 3.3,

$$r_t = r_t^{\text{track}} + w_{\text{code}} \exp\big(-(\eta_t/\sigma_{\text{code}})^2\big), \qquad \eta_t = \frac{1}{|\mathcal{A}_t|} \sum_{k \in \mathcal{A}_t} \frac{|c_k(\mathcal{H}^{\text{sim}}[t]) - \widehat{c}_k[t]|}{\rho_k}, \tag{9}$$

where $c_k(\mathcal{H}^{\text{sim}}[t])$ is the k-th HOI code computed from the simulated state, and ρ_k is the 5–95th percentile range of that code over the OMOMO dataset. The active set $\mathcal{A}_t$ contains the codes whose predicted change in the current phase is non-identity and high-confidence, and the reward is zero when $\mathcal{A}_t$ is empty. We use $w_{\text{code}} = 0.25$ and $\sigma_{\text{code}} = 0.1$. Since the tracked reference remains unchanged, the code reward conflicts with the tracking reward in proportion to the size of the predicted change.

A.9.4 HOI-UP$_{\text{HOICodeOnly}}$

HOI-UP$_{\text{HOICodeOnly}}$ follows the same pipeline as HOI-UP, except that the kinematic HOI optimization in Sec. 3.3 minimizes only the HOI-code objective,

$$\mathbf{z}^\star = \arg\min_{\mathbf{z}} \ \lambda_{\text{code}} E_{\text{code}}, \tag{10}$$

removing the contact (E_{contact}), biomechanical (E_{bio}), and regularization (E_{reg}) terms of Eq. 3. The VLM-predicted targets $\widehat{\mathbf{c}}[t]$ and all other optimization configurations are identical to those of HOI-UP in App. A.4. The optimized kinematic HOI motion $\widehat{\mathcal{H}}$ is then tracked by the same RL fine-tuning as HOI-UP. This variant isolates the effect of realizing the code targets without preserving hand-object contact, foot support, balance, or joint-loading constraints.

A.9.5 CoMo-HOI

Architecture and fine-tuning. We extend CoMo (Huang et al., 2024) with the continuous HOI codes in Tab. 5. An MLP maps these codes into the existing codebook's latent space, and the combined continuous and categorical embeddings condition the human decoder and an added object-pose decoder. Among the selected channels, only elbow and knee angles have dedicated categorical counterparts. Starting from the released CoMo decoder, we fine-tune all parameters on original OMOMO motions without property labels or optimized-motion supervision. Training uses Smooth-L1 losses on human features, object poses, joint positions, and temporal differences, with categorical-code dropout and validation-loss checkpoint selection. Preprocessing preserves metric scale and body proportions, with training-only feature standardization. Settings are listed in Tab. 10.

Inference. A frozen text-only LLM independently predicts phase-wise offsets, gains, and confidence from the original and target properties, interaction phases, and reference-code summaries. We apply affine edits around reference phase means, retain low-confidence or omitted channels, smooth phase boundaries, and update elbow and knee categories consistently. The decoders output 22-joint human positions and object poses at the reference length and frame rate, without physics or contact optimization.

Table 10: Key CoMo-HOI settings. Loss weights correspond to human features, object poses, joint positions, and temporal differences, respectively.

Setting	Value
Source sequences (train/val/test)	2,082 / 113 / 14
Continuous HOI adapter	$17 \to 256 \to 512$
Batch size / window length	32 / 96 frames
Optimizer / learning rate	AdamW / 10^{-4} to 10^{-5}, cosine decay
Loss weights	1 / 1 / 0.5 / 0.25
Categorical-code dropout	0.25
Training steps (stopped / selected)	8,000 / 4,000
LLM / decoding	Qwen2.5-7B-Instruct / greedy

A.10 OBJECT PROPERTY-WISE QUANTITATIVE RESULTS

Table 11: **End-to-end comparison by object property: Mass.** 10, 20, and 30 kg; 30 sequences, 90 conditions, and 450 trials per method (five per condition). **Bold**/underline: first/second distinct displayed scores within this property; ties share rank.

	Physical executability			Naturalness			Biomechanical demand				Preservation	Property response		
Method	Success ↑ (%)	Completion ↑ (%)	HOI error ↓ (mm)	$\mathrm{Jit}_{\mathrm{HOI}}$ ↓ ($\mathrm{m\,s^{-3}}$)	Foot-floor ↓ (cm)	Contact F1 ↑	$\bar{\tau}_{\mathrm{sim}}$ ↓ (N m)	C_{100} ↓ (joints)	OWAS ↓ mean	RNLE LI ↓ valid range	Minimality ↓	Eff ↑ (%)	Mon ↑ (%)	R^2_{prop} ↑ (%)
(a) Controller adaptation														
InterMimic	82.44	91.57	80.4	671.18	0.241	0.929	12.11	3.887	2.402	1.624	0.631	**63.93**	47.92	11.64
InterMimic + ZMP	82.89	**92.05**	**79.8**	**656.39**	**0.166**	**0.941**	**11.65**	3.882	**2.361**	1.581	0.820	63.00	42.36	13.19
InterMimic + HOI-code	79.56	90.58	84.0	703.90	0.277	0.913	12.91	3.890	2.395	1.595	0.753	62.94	44.44	12.03
(b) Motion adaptation														
CoMo-HOI	21.11	53.24	111.0	801.69	0.620	0.800	21.72	4.009	2.548	1.637	0.579	51.82	52.78	8.27
$\text{HOI-UP}_{\text{HOICodeOnly}}$	30.89	56.63	120.5	759.71	0.719	0.838	15.35	3.869	2.496	1.550	0.409	51.51	54.86	15.79
HOI-UP (ours)	**83.11**	91.35	82.5	675.58	0.225	0.934	12.27	**3.815**	2.449	**1.499**	**0.302**	60.41	**65.97**	**19.27**

Table 12: **End-to-end comparison by object property: Scale.** $0.75\times$ and $1.25\times$; 30 sequences, 60 conditions, and 300 trials per method (five per condition). **Bold**/underline: first/second distinct displayed scores within this property; ties share rank.

	Physical executability			Naturalness			Biomechanical demand				Preservation	Property response		
Method	Success ↑ (%)	Completion ↑ (%)	HOI error ↓ (mm)	$\mathrm{Jit}_{\mathrm{HOI}}$ ↓ ($\mathrm{m\,s^{-3}}$)	Foot-floor ↓ (cm)	Contact F1 ↑	$\bar{\tau}_{\mathrm{sim}}$ ↓ (N m)	C_{100} ↓ (joints)	OWAS ↓ mean	RNLE LI ↓ valid range	Minimality ↓	Eff ↑ (%)	Mon ↑ (%)	R^2_{prop} ↑ (%)
(a) Controller adaptation														
InterMimic	66.33	78.14	85.8	749.73	0.699	0.809	13.82	2.301	1.892	0.361	0.290	–	25.83	27.00
InterMimic + ZMP	67.67	77.10	91.0	744.37	0.636	0.807	13.59	2.331	**1.891**	0.357	0.368	–	28.33	39.72
InterMimic + HOI-code	59.33	73.94	94.3	781.35	0.829	0.776	14.55	2.356	1.895	0.361	**0.269**	–	34.17	35.74
(b) Motion adaptation														
CoMo-HOI	23.67	47.91	102.9	792.62	0.377	0.643	15.98	2.527	1.957	0.362	0.365	–	34.17	29.43
$\text{HOI-UP}_{\text{HOICodeOnly}}$	52.67	72.65	96.6	769.68	0.420	0.814	14.69	2.479	1.985	0.356	0.377	–	**38.33**	**40.52**
HOI-UP (ours)	**84.67**	**92.71**	**75.1**	**721.01**	**0.197**	**0.913**	**13.02**	**2.238**	1.965	**0.349**	0.479	–	35.83	27.52

Table 13: **End-to-end comparison by object property: Center of mass (CoM).** -80% and $+80\%$ interventions; 30 sequences, 60 conditions, and 300 trials per method (five per condition). **Bold**/underline: first/second distinct displayed scores within this property; ties share rank.

	Physical executability			Naturalness			Biomechanical demand				Preservation	Property response		
Method	Success ↑ (%)	Completion ↑ (%)	HOI error ↓ (mm)	$\mathrm{Jit}_{\mathrm{HOI}}$ ↓ ($\mathrm{m\,s^{-3}}$)	Foot-floor ↓ (cm)	Contact F1 ↑	$\bar{\tau}_{\mathrm{sim}}$ ↓ (N m)	C_{100} ↓ (joints)	OWAS ↓ mean	RNLE LI ↓ valid range	Minimality ↓	Eff ↑ (%)	Mon ↑ (%)	R^2_{prop} ↑ (%)
(a) Controller adaptation														
InterMimic	74.33	88.03	**73.9**	720.15	0.236	**0.922**	12.68	2.337	1.938	0.364	0.816	–	15.10	–
InterMimic + ZMP	**76.00**	87.81	79.2	**710.85**	0.181	0.910	**12.43**	2.315	**1.914**	0.364	0.797	–	19.44	–
InterMimic + HOI-code	73.33	87.24	81.4	760.88	0.296	0.899	13.71	2.376	1.934	0.365	1.030	–	6.94	–
(b) Motion adaptation														
CoMo-HOI	18.67	47.88	96.2	785.33	0.335	0.730	15.40	2.583	1.993	0.374	**0.411**	–	**23.09**	–
$\text{HOI-UP}_{\text{HOICodeOnly}}$	44.67	67.57	93.6	775.87	0.476	0.833	14.78	2.507	1.985	0.364	1.099	–	14.06	–
HOI-UP (ours)	75.67	**88.90**	74.9	719.94	**0.140**	0.915	12.84	**2.286**	1.982	**0.343**	0.857	–	13.19	–

Table 14: **End-to-end comparison by object property: Friction.** $\mu = 0.1$ and 1.2; 15 sequences, 30 conditions, and 150 trials per method (five per condition). **Bold**/underline: first/second distinct displayed scores within this property; ties share rank.

	Physical executability			Naturalness			Biomechanical demand				Preservation	Property response		
Method	Success ↑ (%)	Completion ↑ (%)	HOI error ↓ (mm)	$\mathrm{Jit}_{\mathrm{HOI}}$ ↓ ($\mathrm{m\,s^{-3}}$)	Foot-floor ↓ (cm)	Contact F1 ↑	$\bar{\tau}_{\mathrm{sim}}$ ↓ (N m)	C_{100} ↓ (joints)	OWAS ↓ mean	RNLE LI ↓ valid range	Minimality ↓	Eff ↑ (%)	Mon ↑ (%)	R^2_{prop} ↑ (%)
(a) Controller adaptation														
InterMimic	90.67	94.49	71.6	720.36	0.244	0.869	12.91	2.547	2.084	–	0.730	–	12.50	11.70
InterMimic + ZMP	88.00	92.45	74.9	**707.75**	0.175	0.853	**12.51**	**2.459**	2.087	–	0.886	–	15.62	10.94
InterMimic + HOI-code	86.67	90.40	78.7	748.72	0.240	0.836	13.66	2.490	2.086	–	1.401	–	16.67	9.30
(b) Motion adaptation														
CoMo-HOI	38.00	62.08	103.9	784.87	0.362	0.661	15.33	2.465	**2.028**	–	0.543	–	18.75	18.00
$\text{HOI-UP}_{\text{HOICodeOnly}}$	55.33	75.63	90.8	773.74	0.312	0.795	14.94	2.596	2.179	–	0.568	–	17.71	19.24
HOI-UP (ours)	**92.00**	**94.94**	**66.2**	712.20	**0.132**	**0.872**	12.97	2.460	2.127	–	**0.462**	–	**29.17**	**22.81**

A.11 Additional Qualitative Results

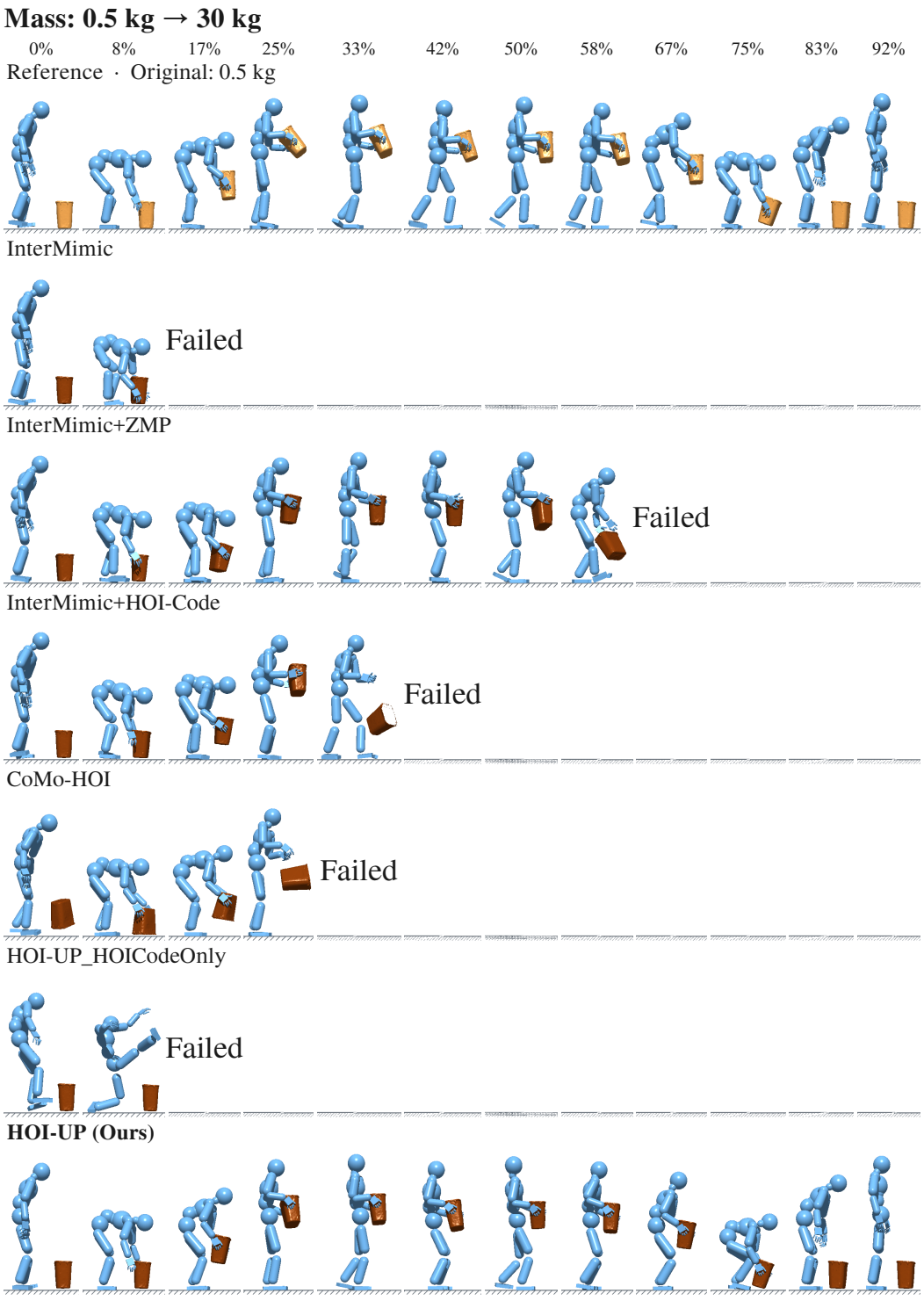


Figure 8: **Whole end-to-end baseline Mass comparison**.



Figure 9: **Whole end-to-end baseline CoM comparison**.

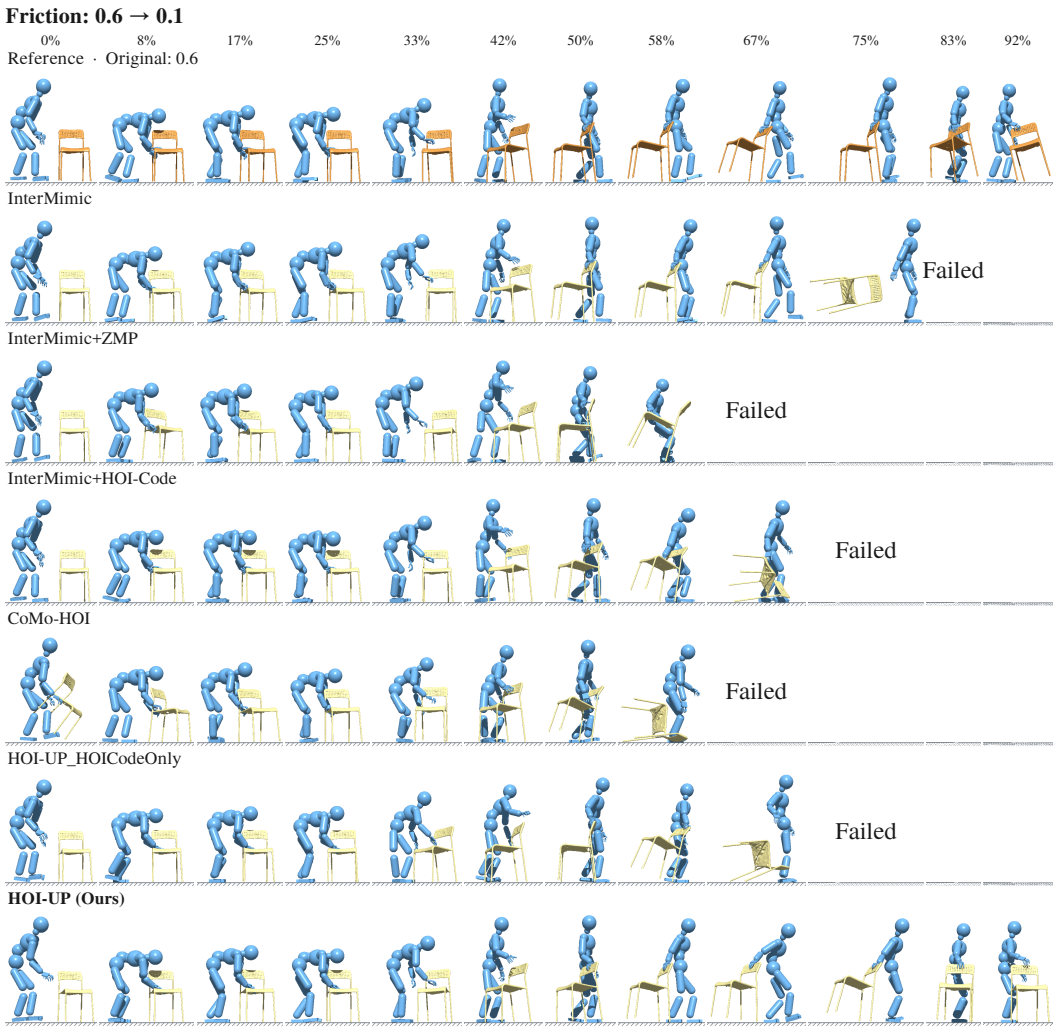


Figure 10: **Whole end-to-end baseline Friction comparison**.



Figure 11: **Whole end-to-end baseline Scale comparison**.

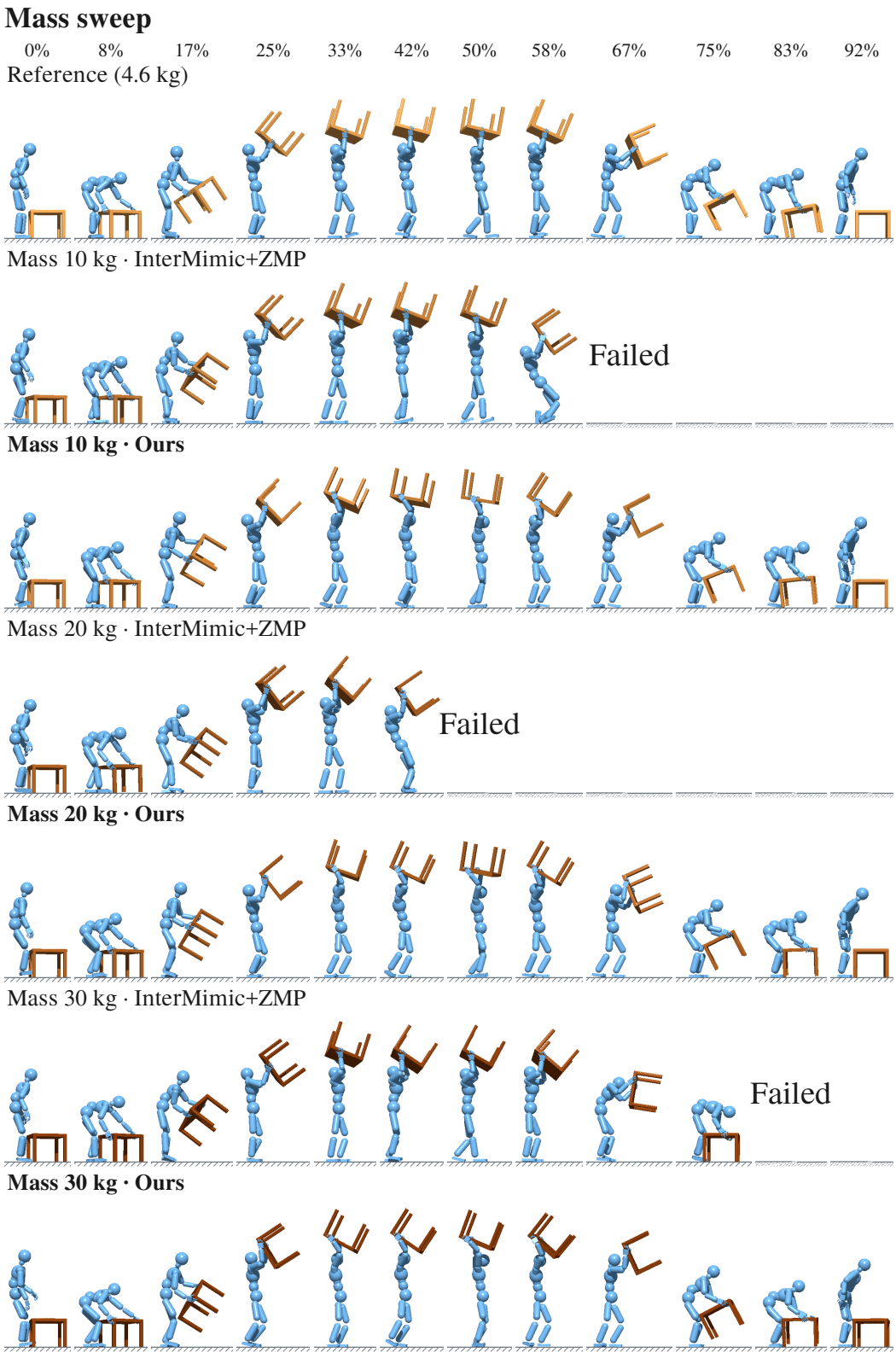


Figure 12: **Mass sweep comparison**.

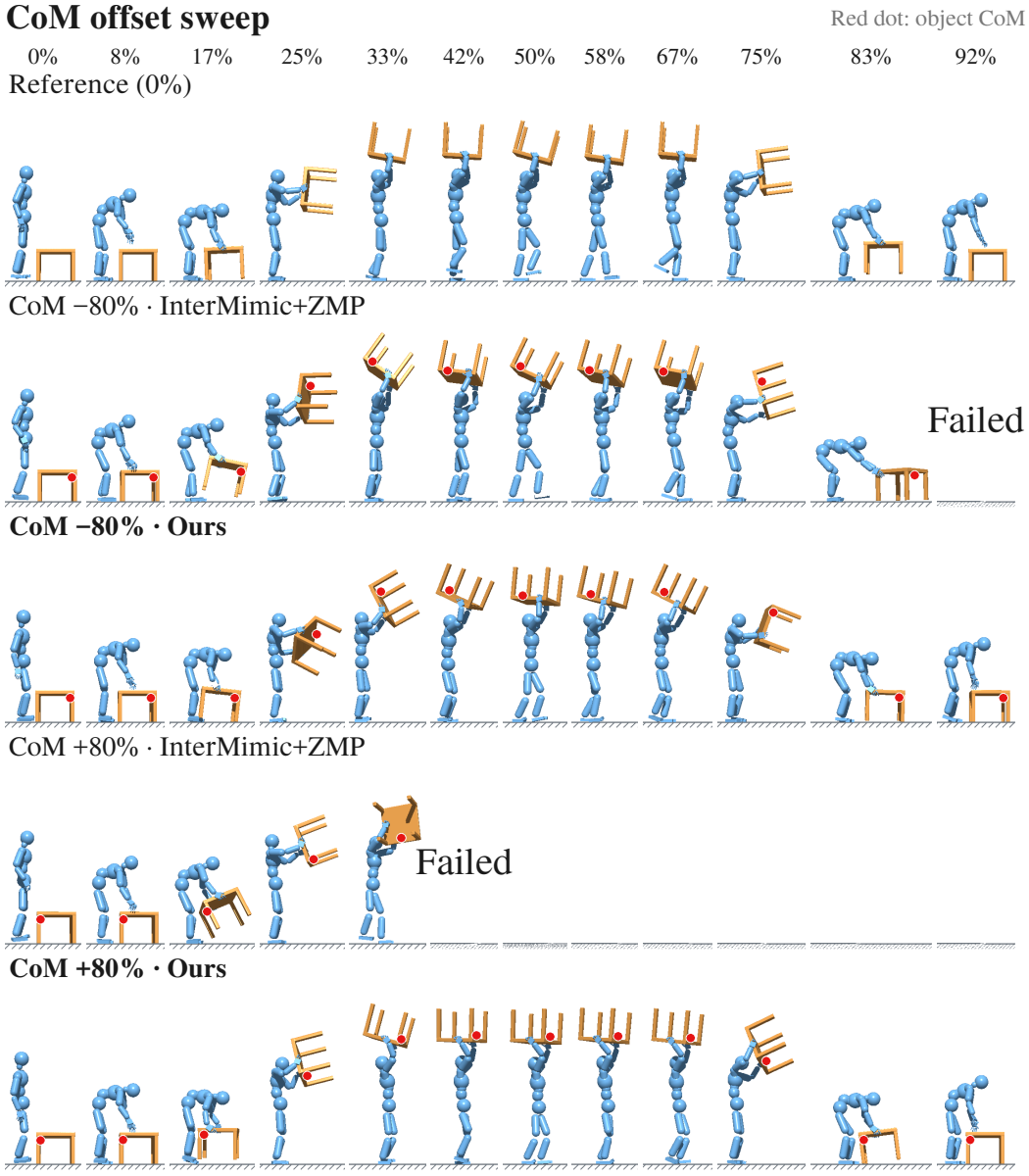


Figure 13: **CoM sweep comparison**.

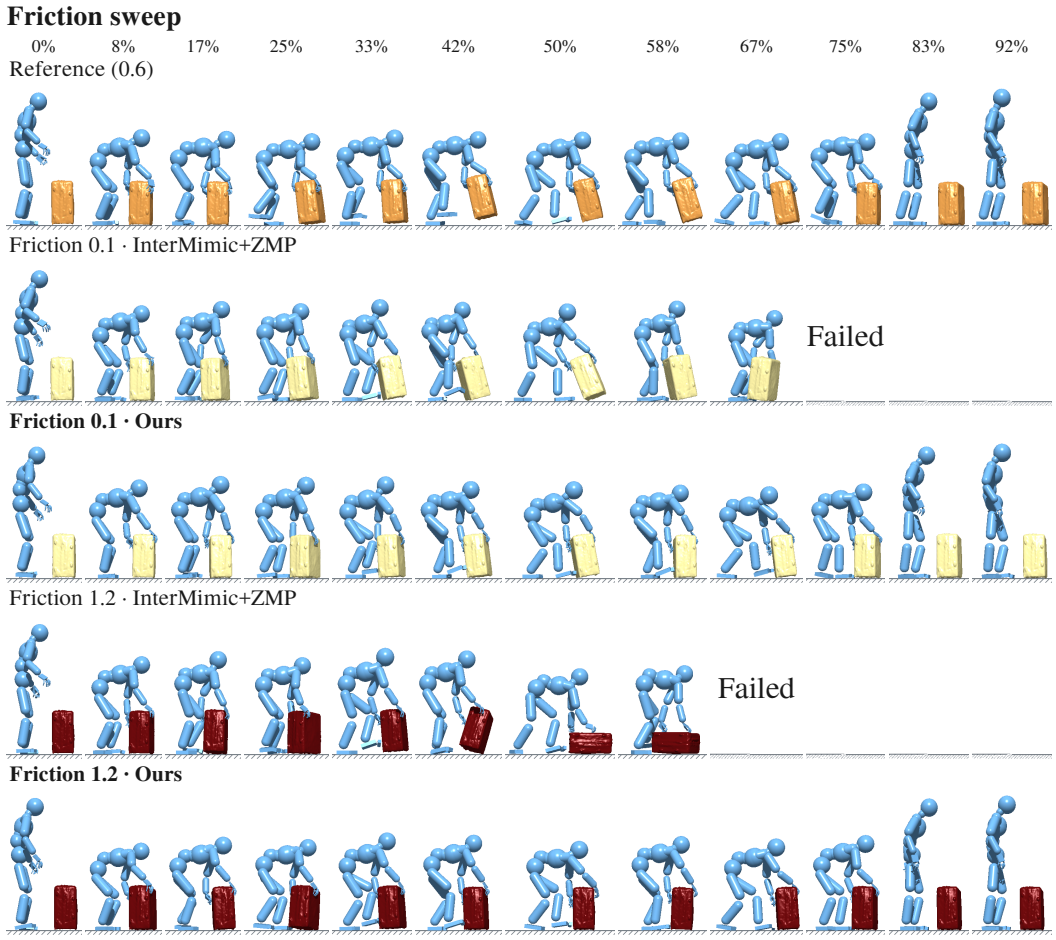


Figure 14: **Friction sweep comparison**.

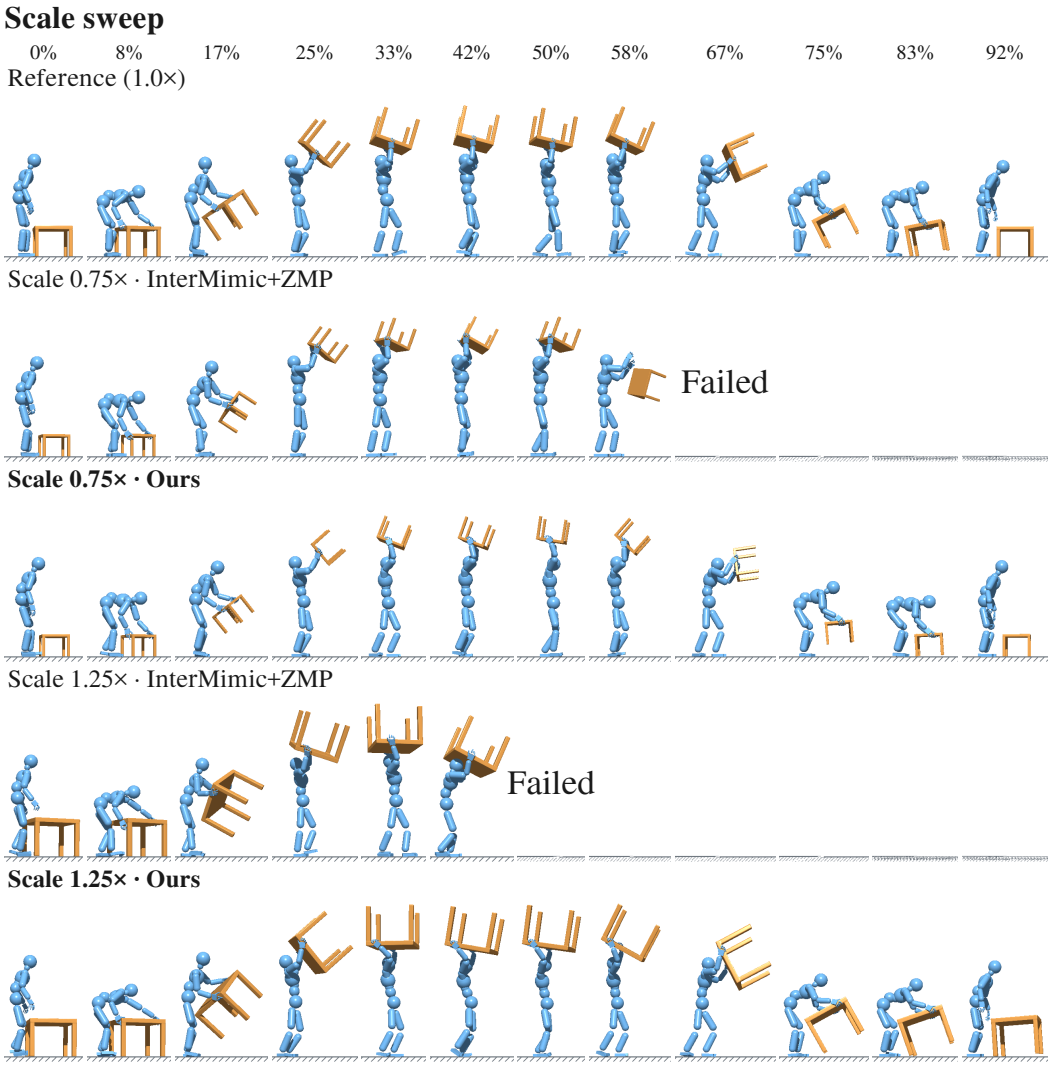


Figure 15: **Scale sweep comparison**.