

Dynamic Full-body Motion Agent with Object Interaction via Blending Pre-trained Modular Controllers

Supplementary Material

A. Supplementary Video

The submitted video qualitatively illustrates how our framework generates dynamic and physically valid HOI motions, comparing its planning and execution with multiple baselines.

- **Prior-Blending for HOI Planning.** Under identical text prompts, Ours_P generates HOI motion plans with more feasible hand-object contacts than baselines.
- **Improving Physical Plausibility.** Physical artifacts in Ours_P plans—such as object-ground penetration—are corrected in the composer-based execution stage.
- **Composer-based Execution.** Our composer achieves the highest task success rate and the most stable HOI imitation among all baselines.
- **More Qualitative Results.** Our full pipeline demonstrates dynamic and scalable HOI behaviors across a wide range of motion styles and object categories.

B. Setup for Physics Simulation

Our HOI simulation environment is built on Isaac Gym [31], following InterMimic [56] to convert the SMPL [25]-based human model from reference motions into rigid-body representations suitable for physics simulation. Each body part is approximated with box or cylinder primitives, while object meshes are converted using convex decomposition as illustrated in Fig. A.

Although our framework utilizes both PHC [26] and InterMimic [56] as imitation experts, the two policies were originally trained under different physics parameter settings. To ensure consistency across all HOI imitation experiments, we adopt the physics parameters used in PHC as our default configuration. We verified that InterMimic maintains comparable performance under PHC’s physics configuration. Consequently, all imitation experiments are conducted using the PHC physics parameters. Detailed physical configurations are provided in Tab. A.

The joint Range of Motion (RoM) follows the biomechanical constraints specified by InterMimic [56]. Finger flexion and extension are fully allowed to support natural grasping behaviors, while the Metacarpophalangeal joints are restricted to prevent finger interpenetration. Although PHC [26] also limits the RoM of several body joints to avoid body interpenetration, we relax these constraints to preserve the capability of the InterMimic agent. The full RoM specifications for each joint are listed in Tab. B.

Table A. Simulation parameters for Isaac Gym [31] used in this paper.

Simulation Parameter	Value
Sim dt	1/60 s
Control dt	1/30 s
Number of envs	1024
Episode length	300
Number of substeps	2
Solver	TGS
Number of position iterations	4
Number of velocity iterations	0
Contact offset	0.02
Rest offset	0.0
Bounce threshold velocity	0.2
Max depenetration velocity	10.0
Ground friction	1.0
Ground restitutions	0.0
Object density	200
Object max convex hulls	64

Table B. Range of Motion of the humanoid robot used in this paper. *Body* denotes body joints—hip, knee, ankle, toe, torso, spine, chest, neck, head, thorax, shoulder, elbow, and wrist.

Joint	x -axis		y - and z -axis	
	min	max	min	max
Body	-180.000	180.000	-180.000	180.000
Thumb1	-55.625	55.625	-55.625	55.625
Thumb2	-5.625	5.625	-5.625	5.625
Thumb3	-5.625	90.000	-5.625	5.625
Index1,2	-55.625	55.625	-5.625	5.625
Index3	-5.625	90.000	-5.625	5.625
Middle1,2	-55.625	55.625	-5.625	5.625
Middle3	-5.625	90.000	-5.625	5.625
Ring1,2	-55.625	55.625	-5.625	5.625
Ring3	-5.625	90.000	-5.625	5.625
Pinky1,2	-55.625	55.625	-5.625	5.625
Pinky3	-5.625	90.000	-5.625	5.625

C. Motion Styles and Interaction Categories

We evaluate our method against baselines across four dynamic motion styles and five interaction categories to assess its scalability to diverse motions and object types.

C.1. Dynamic Motion Styles

The following outlines the expected agent behavior and success criteria for each dynamic motion style. For SR evalua-

tion, an episode is considered success if the agent meets the style-specific criteria and keeps a target object within 0.3 m of the reference path without dropping it.

- **Run Forward.**

Success is achieved if the agent’s horizontal pelvis velocity reaches within 0.5 m/s of the reference’s maximum horizontal pelvis velocity:

$$\max_n \dot{p}_{\text{pelvis,xy}}[n] \geq \max_n \dot{\bar{p}}_{\text{pelvis,xy}}[n] - 0.5,$$

where $\dot{p}_{\text{pelvis,xy}}[n]$ and $\dot{\bar{p}}_{\text{pelvis,xy}}[n]$ denote the horizontal pelvis velocity at timestep n of the rollout and the reference, respectively.

- **Jump Forward.**

Succeed if the pelvis height gain reaches at least 50% of the reference max pelvis height and the number of false foot-floor contacts remains below 10. A false foot-floor contact occurs when a foot touches the floor in the rollout but not in the reference at time step n .

$$\frac{\max_n (p_{\text{pelvis,z}}[n]) - p_{\text{pelvis,z}}^T}{\max_n (\bar{p}_{\text{pelvis,z}}[n]) - p_{\text{pelvis,z}}^T} \geq 0.5,$$

$$\sum_{n=0}^{N-1} c_{\text{foot}}[n](1 - \bar{c}_{\text{left foot}}[n]) \leq 10,$$

where $p_{\text{pelvis,z}}[n]$ and $\bar{p}_{\text{pelvis,z}}[n]$ denote the pelvis height of the rollout and the reference, respectively. $p_{\text{pelvis,z}}^T$ denotes the initial pelvis height when the agent is standing with a T-pose. $c_{\text{foot}}[n]$ and $\bar{c}_{\text{foot}}[n]$ are binary foot-floor contact states for the rollout and the reference, respectively.

- **High Kick.**

Succeed if the foot height reaches within 5 cm of the reference maximum foot height.

$$\max_n p_{\text{foot,z}}[n] \geq \max_n \bar{p}_{\text{foot,z}}[n] - 0.05.$$

- **Dance.**

An episode is considered success if pelvis joint height does not fall below 0.3 m.

$$\min_n p_{\text{pelvis,z}}[n] \geq 0.3.$$

where $p_{\text{pelvis,z}}[n]$ denote the pelvis height of rollout.

C.2. Interaction Categories

We instantiate the above motion styles on three representative objects—*smallbox*, *clothesstand*, and *largetable*—from the FullBodyManip [22] dataset, as shown in Fig. A. These objects represent a range of manipulation difficulty: *smallbox* requires an easy two-hand manipulation, *clothesstand* involves a one-hand interaction, and *largetable* represents a challenging two-hand manipulation with a large object.



Figure A. Visualization of *clothesstand*, *largetable*, and *smallbox* from FullBodyManip [22], each preprocessed into 64 convex hulls following InterMimic [56].

We evaluate five interaction categories using the three objects: (i) **smallbox** involves lifting a smallbox with two hands; (ii) **clothesstand left** and (iii) **clothesstand right** involve lifting a clothesstand with the left and right hand, respectively; (iv) **largetable carry** denotes carrying a largetable at pelvis height using two hands; and (v) **largetable lift** represents lifting a largetable over the head with two hands. Left- and right-hand *clothesstand* interactions are treated as distinct tasks due to their asymmetric motion dynamics.

C.3. Evaluation Metrics

We assess text-to-dynamic HOI generation with respect to HOI quality [22], physical plausibility [59], and motion quality [49]. To compare with DAViD [20], we evaluate only motion segments starting at the onset of object contact.

- **HOI quality.**

- $C_{\%} \uparrow$ (**Contact Percentage**). Percentage of frames in which any body part is in contact with the object.
- $C_{\text{cons}} \downarrow$ (**Contact Consistency**). Standard deviation of object-vertex positions expressed in the hand’s local coordinate frame. A lower value indicates more consistent relative hand–object pose, reflecting more stable interaction.

- **Physical plausibility.**

- $\text{Pene}_{\text{obj}} \downarrow$ (**Object-floor Penetration, cm**). Measures interpenetration depth between the object and the floor, indicating violations of physical plausibility.
- $\text{Skate} \downarrow$ (**mm**). Mean horizontal displacement of foot–ground contact points during contact.
- $\text{Float} \downarrow$ (**mm**). Minimum distance between the ground and the lowest body joint.
- $\text{Jitter}_{\text{pos}} \downarrow$ (m/s^3). Mean third-order derivative of joint positions, measuring motion smoothness.

- **Human motion quality.**

- $\text{R-Precision} \uparrow$. Retrieval-based relevance metric checking whether the ground-truth text–motion pair appears within the top-3 retrieved candidates.
- $\text{Diversity} \uparrow$. Measures variability across the generated human motions.

Table C. Hyperparameters for two imitation experts and our composer training.

	Hyperparameter	Value
PHC	input size	574
	action size	69
	number of primitives	3
	architecture (primitive)	[2048, 1536, 1024, 1024, 512, 512]
	architecture (composer)	[2048, 1536, 1024, 1024, 512, 512]
InterMimic	input size	3198
	action size	153
	architecture	[1024, 1024, 512]
Composer	input size	3630
	action size (w)	153
	action size (r)	153
	action size (μ)	4
	architecture	[1024, 1024, 512, 512]
	PCA buffer size B	16
	PCA subspace dim S	4
	Extrapolation coefficient ρ	0.2
	Subspace coefficient σ	0.08
Training	Discount Factor	0.99
	Generalized adv. estimation	0.95
	Entropy reg. coefficient	0.0
	Optimizer	Adam
	Actor learning rate	2e-5
	Actor learning rate (for InterMimic _{FT})	2e-6
	Critic learning rate	1e-4
	Minibatch size	16384
	Horizon length H	32
	Max episode length N	300

We evaluate dynamic HOI imitation using the metrics introduced in InterMimic [56], along with the training time.

- **$SR \uparrow$ (Success Rate).** Fraction of episodes that satisfy the success criterion in Sec. C.1 without dropping the object.
- **T (Training Time, h).** Training time until the learning curve converges.
- **$D \uparrow$ (Mean Duration, s).** Average uninterrupted interaction time per episode without fall or drop.
- **$E_{\text{HOI}} \downarrow$** Mean per-frame Euclidean distance between the planned and executed HOI motions across 21 SMPL joints (excluding the pelvis) and the object. The metric is computed after globally aligning the pelvis joint at $n=0$.
- **$\text{Jitter}_{\text{DoF}} \downarrow$ (rad/s³)** Third-order derivative of joint rotations, measuring motion smoothness in the joint-rotation space. Each DoF corresponds to a rotational axis of a joint.

D. Additional Implementation Details

The interaction onset is defined as 1.5 seconds after the initial contact, and we observed that the method is robust to this choice. After the onset, the local rotations of interaction-related joints \mathcal{J}_{int} — thorax, shoulder, elbow, and wrist — are fixed to maintain consistent contacts.

All pretrained imitation experts [26, 56] used in this paper follow their publicly available implementations. InterMimic serves as our full-body controller that predicts PD targets for the entire body joints including the fingers. It receives both the observation $\mathbf{s}[n]$ and reference states $\mathbf{g}[n]$ as input and uses a 3-layer MLP followed by a final linear layer producing PD actions. PHC is a versatile expert specialized for high-dynamics whole-body imitation, producing PD targets for all body joints except the fingers. We use the PHC+ variant with three primitive agents, each implemented as a 6-layer MLP trained on progressively more difficult subsets, and a 6-layer composer network that blends their outputs. We follow the settings of the original paper. All InterMimic and PHC components are kept frozen during our composer training.

We implement several baselines for comparison. PPO denotes a policy trained entirely from scratch using the same observation and action spaces as our method. PHC_R and InterMimic_R attach an additional residual MLP that receives the same $\mathbf{s}[n]$ and $\mathbf{g}[n]$ as inputs and outputs an additive correction term; residual outputs are directly added to those of the pretrained policy. InterMimic_{FT} corresponds to training the pretrained InterMimic model without freezing any parameters. We exclude PHC_{FT} since PHC does not receive object observations.

Our proposed composer network is a lightweight MLP designed to blend the two experts spatio-temporally. The composer network is a 4-layer MLP followed by a final linear layer that predicts a 310-dimensional weight vector. The first 153 dimensions correspond to the per-DoF interpolation weights $\mathbf{w}[n]$, the next 153 dimensions to the bounded extrapolation weights $\mathbf{r}[n]$, and the remaining 4 dimensions to the exploration coefficients $\boldsymbol{\mu}[n]$. These components are then passed through separate nonlinear heads: Sigmoid(\cdot) for $\mathbf{w}[n]$, $\rho \tanh(\cdot)$ for $\mathbf{r}[n]$, and $\sigma \tanh(\cdot)$ for $\boldsymbol{\mu}[n]$. For PCA-based low-dimensional action subspace exploration, at each timestep we store the body-only action difference between the two experts, $\Delta \mathbf{a}_{\text{body}}[n] = \mathbf{a}_{\text{body}}^{\text{IM}}[n] - \mathbf{a}^{\text{PHC}}[n]$, into a buffer. PCA is performed on the most recent B samples, and the top- S eigenvectors form an orthogonal basis for the exploration subspace. The detailed hyperparameters are summarized in Tab. C.

E. Additional Qualitative Results

Prior-blending planning comparisons are shown in Fig. E. The execution-stage comparisons of our composer-based

Table D. Additional motion quality results. For clarity, we repeat the main quantitative results of planning-only methods.

Method	Phys.	HOI quality		Physical plausibility				Motion quality	
		$C\%$ \uparrow	C_{cons} \downarrow	Pene. \downarrow	Skate \downarrow	Float \downarrow	Jitter _{pos} \downarrow	R-Prec \uparrow	Div. \uparrow
Planning-only									
DAViD		0.848	10.9	4.842	0.261	20.4	6.69×10^4	0.310	6.70
HOI-Diff		0.358	22.8	2.785	0.633	17.8	1.56×10^4	0.257	4.66
Ours _P		1.000	0.906	4.196	0.217	30.1	5.93×10^4	0.332	5.56
Planning+Execution									
InterMimic _{FT}	✓	0.999	2.50	0.000	2.66	7.98	7.05×10^4	0.256	2.99
InterMimic _R	✓	0.999	3.07	0.000	0.512	19.6	5.43×10^4	0.216	3.40
Ours _{P+E}	✓	0.999	2.95	0.009	0.786	9.95	4.53×10^4	0.316	3.97



Figure B. FullBodyManip objects used in Ours₊.

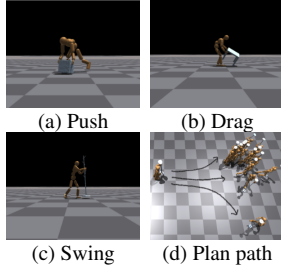


Figure C. Additional motion styles.

Table E. Additional ablation studies. The final row denotes Ours₊ which is our method trained and evaluated on a larger set of objects shown in Fig. B.

Method		More Obj.	$SR \uparrow$	$D \uparrow$	$E_{\text{HOI}} \downarrow$
Planning	Execution				
DAViD			–	0.555	13.47
HOI-Diff	Ours _{MLP+PCA}		–	0.437	23.09
Ours _P	Heuristic _{Hand}		0.365	1.069	13.04
	Heuristic _{Arm}		0.472	1.603	17.18
	Hard MoE		0.383	2.748	20.48
	Hard MoE (Joint)		0.383	2.535	26.45
Ours _P	Ours _{MLP+PCA}		0.591	4.607	11.67
Ours _P	Ours _{MLP+PCA}	✓	0.453	4.348	11.99

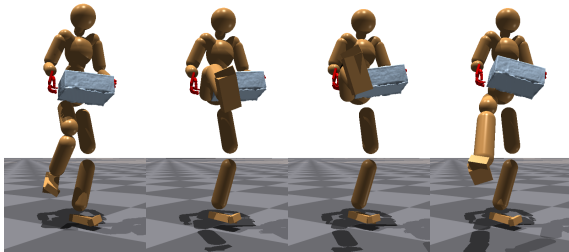


Figure D. The HOI motion planner does not enforce explicit physical constraints, causing non-physical poses such as body-object penetration and unstable hand object contacts.

imitation against baselines are presented in Fig. F. We compare PPO, InterMimic_R, InterMimic_{FT}, and Ours_{MLP+PCA} against the reference motions produced by our planning stage, evaluated across additional object categories and dynamic motion styles. Our method achieves the highest imitation performance across object types and motion styles.

Failure Cases. At the planning stage, the lack of explicit physical constraints leads to non-physical poses such as object-body penetration or unstable contacts, as illustrated in Fig. D.

F. Additional Ablation Studies

F.1. Prior-Blending for Dynamic HOI Planning

Tab. D (rows 1–3) compares our Prior blending Ours_P with HOI motion planners, DAViD and HOI-Diff. We measure motion-dynamic style text alignment by R-Prec and contact stability by $C\%$ and C_{cons} . Ours_P achieves the best for the metrics, indicating stable contacts while generating dynamic motions. Tab. E (rows 1–2,7) measures performance of our composer, using HOI plans generated by each method; ours performs the best, showing that Prior-blending is essential for composer training.

F.2. Composer-based Dynamic HOI Execution

Tab. E (rows 3–7) compares our composer with alternative blending baselines: Heuristic_{Hand} (hands from InterMimic, others from PHC), Heuristic_{Arm} (hands+arms from InterMimic, others from PHC), Hard MoE (global InterMimic/PHC switching), and Hard MoE (Joint) (per-joint switching). Our composer-based blending achieves the highest SR, while Hard MoE (Joint) is even worse than Hard MoE despite finer control, suggesting that soft blending is more important than finer discrete switching. Tab. D (rows 4–6) reports quality of the executed results for the next best baselines in the main paper, InterMimic_{FT} and InterMimic_R; Ours achieves higher R-Prec, that shows how the method preserves the dynamics of reference motion.

G. Discussions

G.1. related works

OmniGrasp [27] focuses primarily on trajectory-conditioned control, yet supports neither text-guided nor dynamic HOI. TokenHSI [32] learns primitive skills such as walking, carrying a box, and climbing on a box, then composing them using a transformer architecture for complex tasks. However, reliance on primitive skills for specific objects (e.g., box) and motion style limits its scalability to novel and dynamic HOI tasks.

G.2. Scalability

Tab. E (rows 7–8) reports Ours₊, where we expand object categories with different geometries and weights shown in Fig. B. Ours₊ achieves comparable performance to Ours, indicating geometry- and weight-agnostic execution of our blending strategies. Fig. Ca–Cc further show that our approach generalizes to diverse interaction styles, push, drag, and swing.

G.3. Plug-and-play Capability

Each module in this work can be replaced in a plug-and-play manner. Fig. Cd shows that replacing the prior-blending backbone with GMD [18] enables path planning. Similarly, InterMimic can be replaced by dataset-specific HOI experts; e.g., a ParaHome [21]-trained agent can be plugged into the InterMimic branch for kitchen-centric interactions.

G.4. Inpainting Strategy in Dynamic HOI Planning

For efficient dynamic HOI agent training, reference motions should preserve coherent whole-body kinematics and contact consistency. Our prior-blending preserves kinematic coherence through iterative denoising [43] and enforces contact consistency through inpainting. The resulting “glued” artifacts are refined by the execution stage, yielding natural and physically-plausible grasping.

G.5. Error Accumulation of Two-stage Framework

CLoSD [50] tackles error accumulation between the planning and execution stages via an autoregressive closed loop. Instead, our work prioritizes dynamic HOI motion generation. Incorporating the autoregressive planner and our composer in closed-loop is a promising direction to mitigate error accumulation, which we leave for future work.

G.6. Limitations

While our method enables diverse and dynamic HOI generation and imitation, it has several limitations. First, integrating the two experts used in this paper is challenging because they were originally trained under different humanoid

kinematics, making it difficult to preserve their full capabilities. Second, the planning stage does not enforce explicit physical constraints, such as preventing human–object penetration or dynamically infeasible human poses; therefore, physical plausibility is guaranteed only after the execution stage.

G.7. Future Work

Future work includes expanding motion-style and interaction-category combinations to enable a broader range of dynamic HOI behaviors. Furthermore, we plan to incorporate explicit physical feasibility constraints directly into the planning stage, reducing reliance on the execution stage for correction. Additionally, our planning–execution framework can be extended in an autoregressive manner to enable real-time dynamic HOI manipulation.

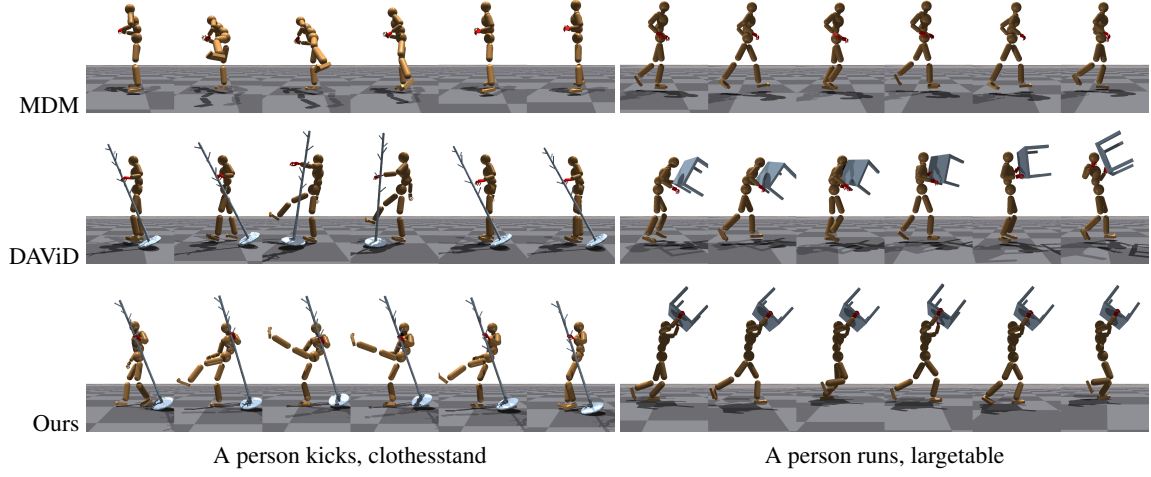


Figure E. **Qualitative comparisons of prior-blending for HOI planning.** MDM [49] does not account for object interaction, often failing to establish hand-object contacts. DAVID [20] produces unstable and inconsistent hand-object alignment. In contrast, our planning stage maintains consistent hand-object interaction while preserving the intended motion style.

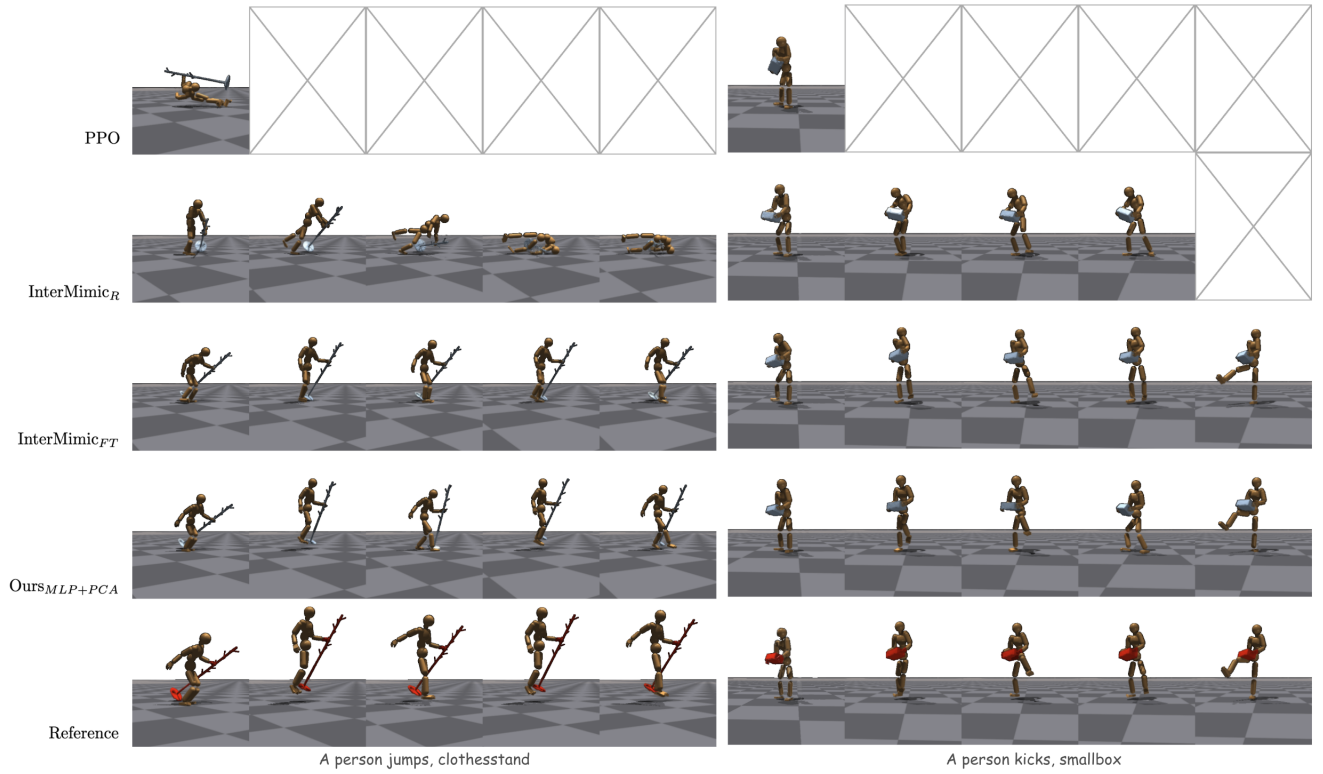


Figure F. **Qualitative comparisons of HOI execution.** Our method generates more physically plausible and dynamic motions than baseline controllers. During jumps, InterMimic_{FT} approximates the motion through small repetitive stepping patterns, whereas Ours produces an actual two-foot lift-off. During kicks, our foot trajectories reach higher and align more closely with the reference. Frames marked with X indicate early termination due to object drops or robot falls.