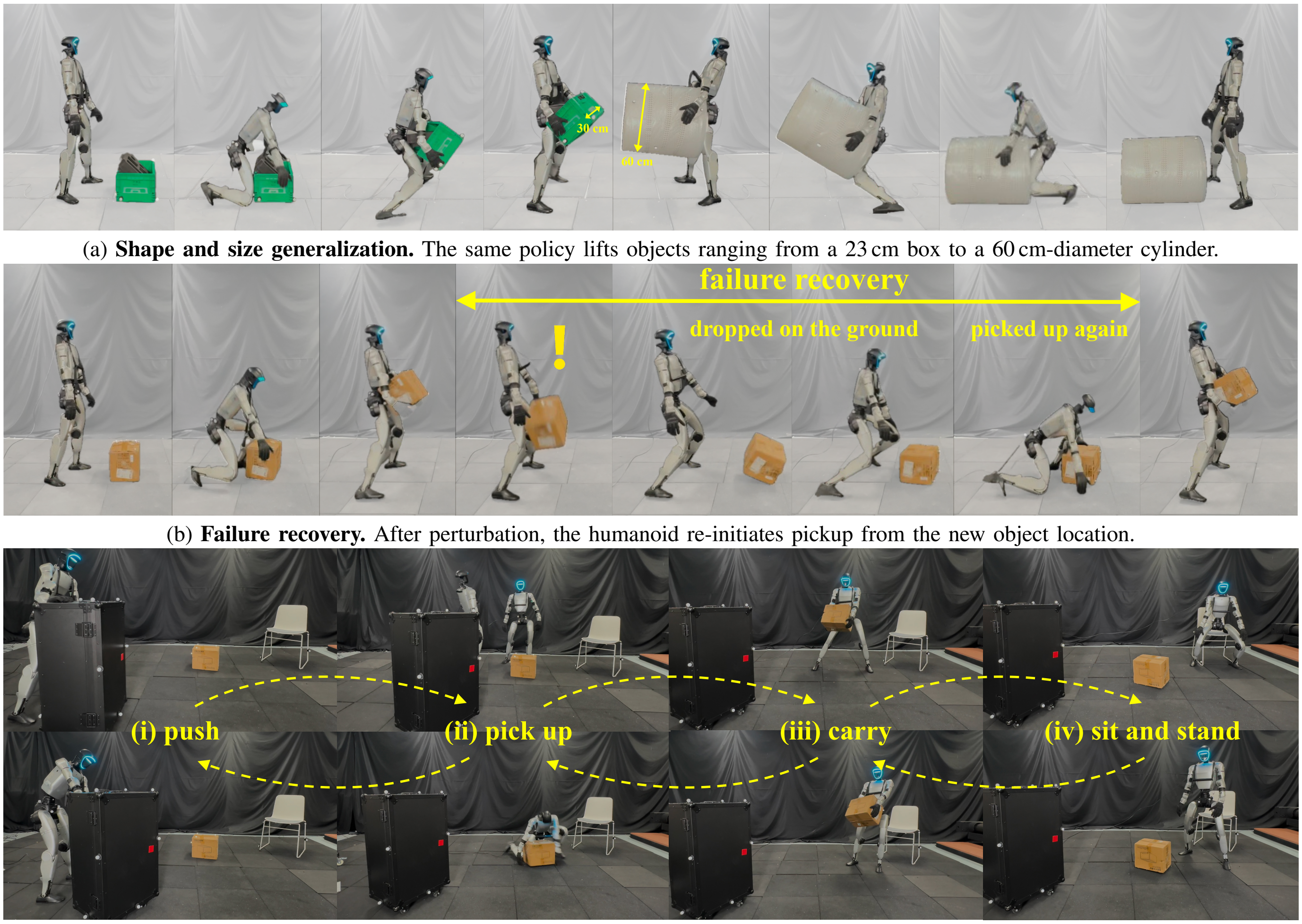
LessMimic: Long-Horizon Humanoid Interaction with Unified Distance Field Representations

Yutang Lin *,1,2,3,5,6,7 Jieming Cui *,1,2,3,5,6,7 Yixuan Li 4,2,5
Baoxiong Jia ✉,2,5 Yixin Zhu ✉,3,1,5,6,7 Siyuan Huang ✉,2,5

*Equal contribution ✉Corresponding Author Project Website: https://lessmimic.github.io
[1]Institute for AI, Peking University [2]Beijing Institute for General Artificial Intelligence (BIGAI)
[3]School of Psychological and Cognitive Sciences, Peking University
[4]School of Computer Science and Technology, Beijing Institute of Technology
[5]State Key Lab of General AI [6]Beijing Key Laboratory of Behavior and Mental Health, Peking University
[7]Embodied Intelligence Lab, PKU-Wuhan Institute for Artificial Intelligence



(a) **Shape and size generalization.** The same policy lifts objects ranging from a 23 cm box to a 60 cm-diameter cylinder.

(b) **Failure recovery.** After perturbation, the humanoid re-initiates pickup from the new object location.

(c) **Long-horizon skill composition.** A single policy executes push, pick-up, carry, and sit-stand sequentially without resets.

Fig. 1: **Generalizable long-horizon humanoid interaction via LessMimic.** A single DF-conditioned policy supports (a) online failure recovery through continuous geometric feedback, (b) generalization to unseen object shapes and scales without retraining, and (c) long-horizon composition of heterogeneous interaction skills within a single policy.

***Abstract*—Humanoid robots that autonomously interact with physical environments over extended horizons represent a central goal of embodied intelligence. Existing approaches rely on reference motions or task-specific rewards, tightly coupling policies to particular object geometries and precluding multi-skill generalization within a single framework. A unified interaction representation enabling reference-free inference, geometric generalization, and long-horizon skill composition within one policy remains an open challenge. Here we show that Distance Field (DF) provides such a representation: LessMimic conditions a single whole-body policy on DF-derived geometric cues—surface distances, gradients, and velocity decompositions—removing the need for motion references, with interaction latents encoded via a Variational Auto-Encoder (VAE) and post-trained**

using Adversarial Interaction Priors (AIP) under Reinforcement Learning (RL). Through DAgger-style distillation that aligns DF latents with egocentric depth features, LessMimic further transfers seamlessly to vision-only deployment without motion capture (MoCap) infrastructure. A single LessMimic policy achieves 80–100% success across object scales from $0.4\times$ to $1.6\times$ on PickUp and SitStand where baselines degrade sharply, attains 62.1% success on 5 task instances trajectories, and remains viable up to 40 sequentially composed tasks. By grounding interaction in local geometry rather than demonstrations, LessMimic offers a scalable path toward humanoid robots that generalize, compose skills, and recover from failures in unstructured environments.

I. Introduction

Consider a humanoid tasked with tidying a room: push a chair aside, pick up a box, carry it across the room, and sit down to rest. Each sub-task involves distinct contact patterns, object geometries, and body configurations—yet a skilled human executes the entire sequence fluidly, without consulting a motion script. Replicating this capability on hardware is an open problem in embodied intelligence [34, 21, 6]. Recent progress in whole-body humanoid control has produced impressive results in locomotion and dynamic motion [12, 27, 61, 24, 26, 7], yet translating these skills into persistent, contact-rich interaction with varied objects remains out of reach.

The core obstacle is representational. How should a robot perceive and reason about its relation to an object during interaction, and what form should that representation take to remain useful across geometries, scales, and task horizons? Current approaches split into two camps, each sacrificing something essential. **Reference-based methods** [54, 51, 62, 27, 61] achieve high-fidelity motions by conditioning policies on recorded demonstrations, but this motion-centric formulation rigidly entangles object geometry with specific reference trajectories, creating a dual limitation. First, the policy becomes a geometric specialist that memorizes training instances and fails on novel shapes. Second, and more critically, it forfeits steering flexibility: any real-time deviation from the reference trajectory is penalized as a tracking failure, not interpreted as an adaptive response. To recover this lost maneuverability, recent variants [53, 32, 8, 56, 26, 58] introduce higher-level planners or human-in-the-loop teleoperation, but these additions merely circumvent the representational bottleneck at the cost of perceptual complexity and full autonomy. **Reference-free methods** [50, 10] take the opposite stance, discarding motion references to gain flexibility, but without a principled interaction signal they resort to task-specific reward engineering and produce isolated policies that cannot compose across skills. Neither camp provides what is ultimately needed: a geometry-aware, task-unified representation that decouples interaction logic from specific motion patterns, enabling seamless skill composition without sacrificing adaptive maneuverability.

We argue that the Distance Field (DF) [35] is precisely this representation. A DF assigns to every point in space its distance to the nearest object surface, yielding a continuous, differentiable field whose gradient encodes surface normals everywhere—including during contact. This is not merely a convenient geometric abstraction; it has concrete consequences for learning. Where point clouds and voxels discretize space and lose gradient information, and where implicit neural representations are too slow for high-frequency control, the DF provides dense, directional geometric cues at negligible query cost. Crucially, the local DF structure near a hand grasping an object is largely invariant to object size and approach direction, making the representation inherently shape- and scale-agnostic. Beyond static geometry, first-order DF cues—surface gradients and velocity decompositions into normal and tangential components—capture the directional dynamics of ongoing contact, providing precisely the interaction signal needed for contact-rich whole-body coordination.

TABLE I: **Comparison of humanoid-object interaction methods.** We compare representative reference-based and reference-free methods across three axes: whether (i) a unified observation space is shared across tasks, (ii) inference requires no motion references, and (iii) a single policy supports autonomous long-horizon skill composition.

Method	**Type**	**Task-unified observation**	**No motion at inference**	**Long-horizon skill composition**
HDMI [51]	Ref-based	✗	✗	✗
ResMimic [62]	Ref-based	✗	✗	✗
CLONE [26]	Ref-based	✗	✗	✓
Op3-Soccer [10]	Ref-free	✗	✗	✗
PhysHSI [50]	Ref-free	✗	✓	✗
Ours	Ref-free	✓	✓	✓

Guided by this insight, we introduce **LessMimic**, a reference-free framework that places the DF at the center of humanoid interaction. LessMimic extracts a velocity-decomposed, history-dependent interaction feature from the DF—capturing approach intensity, surface traversal, and geometric evolution—and encodes it into a compact latent via a VAE. A single whole-body policy is then trained on this latent using a three-stage pipeline: behavior cloning from a teacher for stable initialization, RL fine-tuning with our proposed AIP for geometric generalization across randomized object geometries, and visual distillation for deployment without MoCap. At inference, **the policy requires only a target root trajectory and the current DF observation—no motion references, no task-specific rewards, no separate planners**.

The result is a single policy that simultaneously achieves what prior methods treat as competing objectives. Across four interaction tasks (PickUp, SitStand, Push, Carry) and object scales from $0.4\times$ to $1.6\times$, LessMimic attains 80–100% success on PickUp and SitStand where both reference-based and reference-free baselines degrade sharply. For long-horizon execution, the unified DF representation enables implicit skill transitions without explicit sequencing: LessMimic achieves 62.1% success on 5-task trajectories and remains viable across sequences of up to 40 heterogeneous task instances sequence—a regime where all ablated variants collapse to zero.

In summary, our contributions are threefold:

- A DF-based interaction representation that encodes local geometric relationships as lightweight, differentiable, and shape-agnostic cues, enabling a single policy to generalize across diverse object geometries without retraining.

- A three-stage training pipeline—behavior cloning, AIP-guided RL, and visual distillation—that produces a whole-body interaction policy requiring only a root trajectory command at inference, with no motion references.
- Demonstration that a single **LESSMIMIC** policy can execute, transition between, and recover from heterogeneous interaction skills over horizons of up to 40 consecutive task instances, extending the practical scope of long-horizon humanoid interaction.

II. RELATED WORK

A. Reference-based Humanoid-Object Interaction

Recent advances in RL have driven significant progress in whole-body humanoid control, producing increasingly capable systems for dynamic motion and physical interaction [33, 9, 12, 18]. A dominant paradigm formulates humanoid-object interaction through motion-centric representations, conditioning policies on reference motions as explicit supervision targets to stabilize learning and produce physically plausible behaviors [51, 27, 61, 62, 56, 1, 53]. While effective, this formulation tightly couples the learned policy to the object geometries and configurations present in the reference data [38, 37]: when object properties deviate from the training distribution, the prescribed motions become invalid, leading to brittle behavior on novel objects [21, 17].

To recover adaptability, several works introduce teleoperation or human-in-the-loop strategies that use closed-loop human guidance to generate or correct reference motions online [26, 58, 57, 32, 23, 12, 9, 13]. These approaches improve robustness under object and contact variation, but the dependence on motion references remains structural: human intervention adds supervision overhead and limits scalability, while the policy remains constrained by the topology of observed motions [2, 43, 5]. The result is a persistent trade-off between interaction fidelity, policy autonomy, and generalization that motion-centric representations cannot resolve.

B. Reference-free Humanoid-Object Interaction

Eliminating motion references at inference time has been widely recognized as a key step toward greater autonomy [31]. Early attempts drew inspiration from model-based control [40, 44, 14] and online optimization [45, 15], generating behaviors from task objectives without prerecorded motions. While these approaches offer strong controllability, their reliance on accurate system models and short planning horizons limits applicability to complex, contact-rich scenarios.

More recently, learning-based reference-free methods have gained traction [24, 25] by conditioning policies directly on task-relevant state information, enabling tighter perception-action coupling and simpler real-world deployment [50, 39, 48, 60, 22, 10]. Such formulations show promising robustness under real-world sensing conditions and greater flexibility on out-of-distribution objects and environments. However, most rely on task-specific observations, rewards, or handcrafted interaction cues [30], producing specialized policies that cannot generalize across tasks or compose skills over extended horizons. The absence of a unified interaction representation limits the robust multi-skill behavior within a single policy.

C. Representation for Human-Object Interaction

Geometric representations for human–object interaction span a broad design space, each with distinct trade-offs between expressiveness and computational cost. Voxel-based and occupancy representations [29, 16, 49, 52, 20] explicitly model geometry and collisions but incur high memory and compute costs that preclude real-time whole-body control. Point-based [63, 19, 3, 28] and mesh-based [59, 46, 36, 11] representations capture detailed surface geometry and are well-suited for perception and offline planning, yet lack the continuous distance information needed for contact-aware control. Implicit neural representations [47, 55, 41, 4] offer expressive continuous geometry, but their inference cost makes integration into high-frequency control loops impractical.

The DF [35] bypasses these limitations by encoding geometric proximity as a continuous, differentiable field that can be queried at negligible cost. Unlike discrete representations that lose gradient information or implicit neural representations that incur high inference latency, DFs provide analytical surface gradients essential for contact-aware coordination. For humanoid interaction, where interpenetration is absent, unsigned distance magnitude and local gradients constitute a sufficient geometric description—a simplification that retains the cues necessary for contact-rich control while ensuring the efficiency required for real-time deployment.

III. THE LESSMIMIC FRAMEWORK

We introduce **LESSMIMIC**, a reference-free framework that leverages Distance Field (DF) as a unified interaction representation, enabling a single policy to acquire diverse and generalizable humanoid interaction skills. Fig. 2 provides a high-level view of **LESSMIMIC** at deployment; the overall architecture, illustrated in Fig. 3, integrates geometry-aware perception with whole-body control to support long-horizon interaction across objects of varying shapes and sizes. We first formulate the DF-based interaction representation in

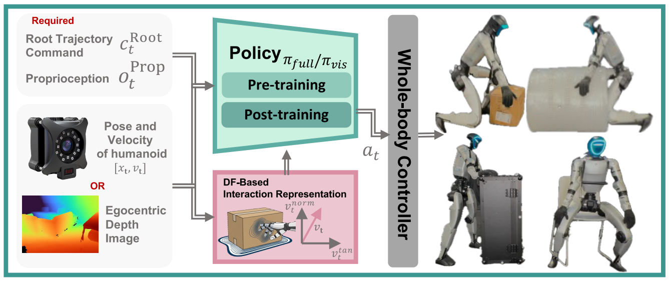

Required
Root Trajectory Command c_t^{Root}
Proprioception o_t^{Prop}
Pose and Velocity of humanoid $[x_t, v_t]$
OR
Egocentric Depth Image
Policy π_{full}/π_{vis}
Pre-training
Post-training
DF-Based Interaction Representation
v_t^{norm}
v_t
v_t^{tan}
a_t
Whole-body Controller

Fig. 2: **LESSMIMIC framework overview.** The policy takes as input a root trajectory command, humanoid proprioception, and a unified DF-based interaction representation that captures current humanoid-object spatial- and temporal-relation. The representation is constructed from MoCap or depth image and encoded into a compact latent z_t via a VAE. The policy is trained in two stages (interaction skill pre-training and discriminative post-training) and outputs actions to a whole-body controller at deployment.

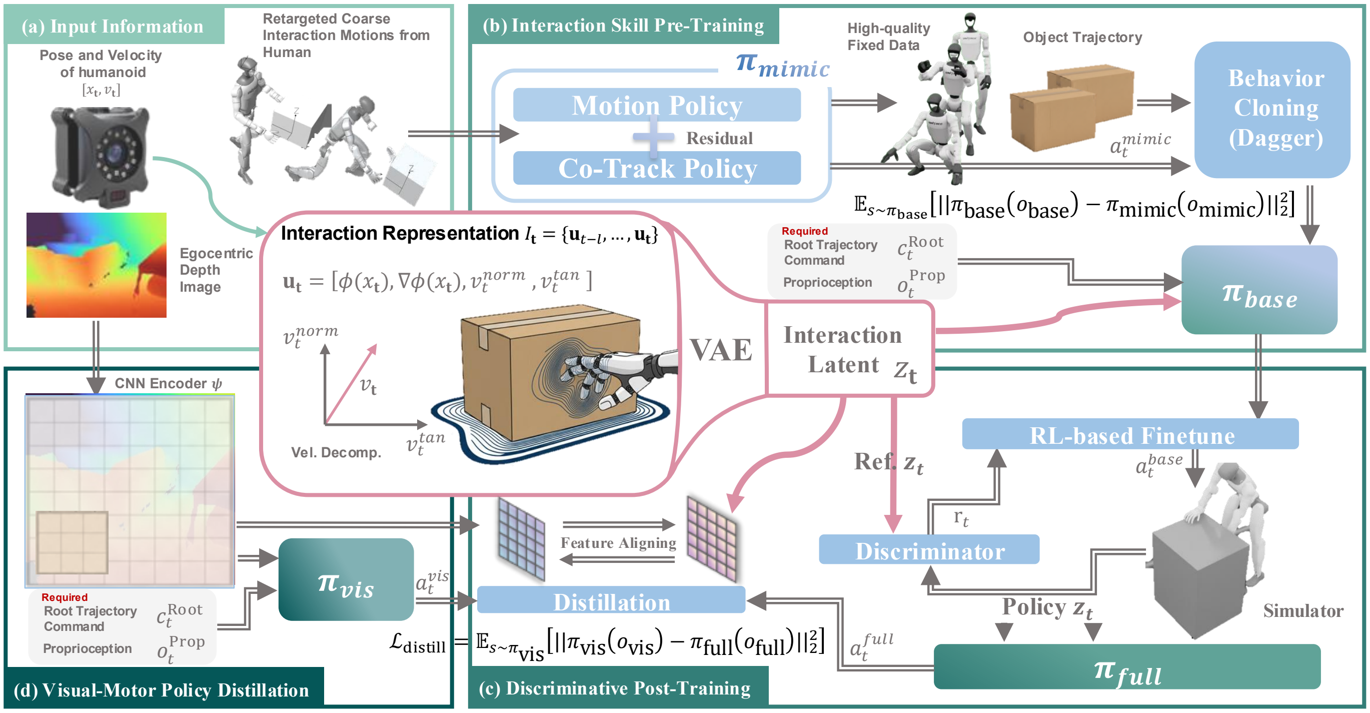

(a) Input Information
Pose and Velocity of humanoid
$[x_t, v_t]$
Retargeted Coarse Interaction Motions from Human
Egocentric Depth Image
(b) Interaction Skill Pre-Training
π_{mimic}
Motion Policy
Residual
Co-Track Policy
High-quality Fixed Data
Object Trajectory
a_t^{mimic}
Behavior Cloning (Dagger)
$\mathbb{E}_{s\sim\pi_{base}}[||\pi_{base}(o_{base}) - \pi_{mimic}(o_{mimic})||_2^2]$
Interaction Representation $I_t = \{\mathbf{u}_{t-l}, \ldots, \mathbf{u}_t\}$
$\mathbf{u}_t = [\phi(x_t), \nabla\phi(x_t), v_t^{norm}, v_t^{tan}]$
v_t^{norm}
v_t
v_t^{tan}
Vel. Decomp.
VAE
Interaction Latent z_t
Required
Root Trajectory Command c_t^{Root}
Proprioception o_t^{Prop}
π_{base}
CNN Encoder ψ
Ref. z_t
RL-based Finetune
a_t^{base}
r_t
Discriminator
Feature Aligning
Distillation
Policy z_t
Simulator
π_{vis}
a_t^{vis}
Required
Root Trajectory Command c_t^{Root}
Proprioception o_t^{Prop}
$\mathcal{L}_{distill} = \mathbb{E}_{s\sim\pi_{vis}}[||\pi_{vis}(o_{vis}) - \pi_{full}(o_{full})||_2^2]$
a_t^{full}
π_{full}
(d) Visual-Motor Policy Distillation
(c) Discriminative Post-Training

Fig. 3: **Training pipeline of LESSMIMIC.** (a) Object observations from either MoCap or an egocentric depth camera yield per-link DF features $\mathbf{u}_t$, which are velocity-decomposed and encoded into an interaction latent representation z_t via a VAE. (b) During interaction skill pre-training, a teacher policy π_{mimic} tracks retargeted human motions to generate physically valid data, from which π_{base} is trained via behavior cloning without access to reference motions, but takes root trajectory command c_t^{root}, humanoid proprioception o^{prop} and the DF latent z_t. (c) During discriminative post-training, π_{base} is fine-tuned with RL guided by AIP, a discriminator that regularizes interaction validity in the geometric domain across randomized object geometries, yielding π_{full}. (d) During visual-motor distillation, π_{full} is distilled into π_{vis} via DAgger-style supervision, replacing MoCap inputs with egocentric depth features for portable real-world deployment.

Sec. III-A, then detail the three-stage training pipeline: pre-training (Sec. III-B), discriminative post-training (Sec. III-C), and visual-motor distillation (Sec. III-D).

A. DF-based Interaction Representation

Methods conditioned on absolute motion trajectories or object-specific representations entangle interaction logic with particular spatial layouts, producing brittle behavior under geometric variation. **LESSMIMIC** instead constructs a geometry-aware state space in which interaction is described through local DF relationships rather than global coordinates. This formulation enables the policy to reason about contact and relative motion invariantly to object shape and scale, capturing the essential structure of an interaction rather than memorizing absolute trajectories.

Formally, we denote the DF as $\Phi : \mathbb{R}^3 \rightarrow \mathbb{R}$, serving as a dynamic local reference frame anchored to the object's geometry. At time t, for a humanoid link (*e.g.*, a hand or pelvis) at position $\mathbf{x}_t \in \mathbb{R}^3$ with linear velocity $\mathbf{v}_t \in \mathbb{R}^3$, the DF yields the distance to the object surface $\Phi(\mathbf{x}_t)$ and the local surface orientation $\nabla\Phi(\mathbf{x}_t)$—the gradient at the projection of $\mathbf{x}_t$ onto the zero-level set of Φ—supplying geometry-aligned cues for contact-aware interaction control.

Position alone is insufficient to characterize interaction dynamics, as it captures neither kinetic state nor directional intent relative to the object surface. We therefore decompose $\mathbf{v}_t$ into two orthogonal components using $\nabla\Phi(\mathbf{x}_t)$ as the local surface normal: motion along the normal direction, corresponding to approach or force application, and motion within the tangent plane, corresponding to sliding or surface traversal. Formally:

$$\mathbf{v}_t^{\text{norm}} = (\mathbf{v}_t \cdot \nabla\Phi(\mathbf{x}_t))\nabla\Phi(\mathbf{x}_t), \quad \mathbf{v}_t^{\text{tan}} = \mathbf{v}_t - \mathbf{v}_t^{\text{norm}}, \tag{1}$$

where $\mathbf{v}_t^{\text{norm}}$ captures the interaction intensity relative to the surface and $\mathbf{v}_t^{\text{tan}}$ captures the flow across the surface geometry, together projecting the global velocity into the local coordinate system defined by the object's surface.

To capture the temporal evolution of the interaction, we collect the per-link tuple $\mathbf{u}_t = [\Phi(\mathbf{x}_t), \nabla\Phi(\mathbf{x}_t), \mathbf{v}_t^{\text{norm}}, \mathbf{v}_t^{\text{tan}}]$ at each time step over a set of task-relevant links, and define the *interaction representation* I_t as a trajectory of these local geometric features over a temporal window of length l:

$$I_t = \{\mathbf{u}_{t-l+1}, \ldots, \mathbf{u}_t\}. \tag{2}$$

Since I_t is defined entirely through the robot's relation to the DF, it is invariant to the object's global pose and scale: interactions with objects of different sizes, shapes, or placements exhibit similar geometric structure, allowing the policy to learn the underlying geometry of interaction behaviors rather than memorizing absolute trajectories, and thereby supporting generalization across object geometries. An example of the DF distance and gradient evolution during interaction is shown in

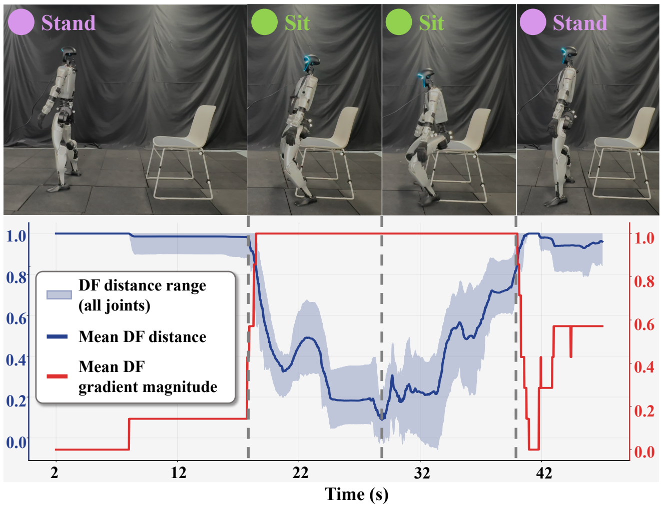


Fig. 4: **An example of the DF signals during a sitting interaction.** The blue curve shows the mean DF distance between the humanoid and the chair across all joints, with the shaded region indicating the full range; the red curve shows the mean DF gradient magnitude. As the humanoid approaches and makes contact with the chair, the distance decreases while the gradient magnitude increases, reflecting the intensifying geometric coupling. Vertical dashed lines indicate transitions between interaction phases (Stand, Sit, Sit, Stand).

Fig. 4. Prior to policy training, I_t is encoded by a VAE into a smooth latent z_t, improving robustness to sensor noise and facilitating convergence during training.

B. Interaction Skill Pre-Training

Learning the DF-based interaction representation requires data that is both semantically meaningful and physically feasible. Purely retargeted human MoCap data frequently violates physical constraints in simulation, while collecting high-quality interaction data at scale is costly. To balance feasibility and scalability, we generate training data in simulation via a mimic policy π_{mimic} that tracks retargeted reference motions under full physics. Following [62], π_{mimic} augments motion tracking with a residual module to compensate for dynamic mismatches during object interaction, producing physically valid state-action trajectories for downstream training. Further implementation details are provided in Sec. A-A.

We pre-train the target policy π_{base} via behavior cloning on data generated by π_{mimic}, with the goal of initializing the policy under the inference-time observation setting. Unlike π_{mimic}, which has access to privileged reference motions, π_{base} operates solely on $o_{\text{base}} = [o_{\text{prop}}, c_t^{root}, z_t]$, comprising proprioception o_{prop}, a sparse root-trajectory command c_t^{root}, and the DF interaction latent z_t. To mitigate covariate shift between training and rollout distributions, we apply DAgger [42] by rolling out π_{base} and querying π_{mimic} for corrective actions. Training minimizes the Mean Squared Error (MSE) loss between the actions of the base and teacher policies:

$$\mathcal{L}_{\text{BC}} = \mathbb{E}_{s\sim\pi_{\text{base}}}\left[\|\pi_{\text{base}}(o_{\text{base}}) - \pi_{\text{mimic}}(o_{\text{mimic}})\|_2^2\right], \quad (3)$$

where o_{mimic} includes privileged full-body reference motions unavailable at inference. This stage yields a stable initialization that grounds the DF interaction representation in whole-body control, preparing the policy for geometry-aware generalization in the subsequent post-training stage.

C. Discriminative Post-Training

Pre-training yields a stable initialization, but π_{base} is trained on a fixed set of objects paired with reference motions, which encourages memorization of specific kinematic trajectories rather than learning the underlying geometric rules of interaction. To promote genuine geometric generalization, we fine-tune the policy using RL in a procedurally augmented environment where object geometries are randomized across scale, shape, and surface properties.

In this setting, motion-tracking rewards and hand-crafted shaping terms are deliberately excluded: reference motions are unavailable under procedural object variation, and task-specific reward terms would require redefinition across geometries. Instead, we introduce Adversarial Interaction Priors (AIP) as a geometry-aware supervision signal. Inspired by Adversarial Motion Priors (AMP) [39], which regularizes motion naturalness, AIP instead regularizes *interaction validity* in the geometric domain. The key insight is that while the absolute joint configurations required for interaction vary with object geometry, the local geometric relationship between the robot and the object surface—captured by the interaction latent z_t—remains consistent across geometries and can serve as a transferable supervision signal.

We train a discriminator D to distinguish interaction latents generated by the policy on novel objects from those stored in a reference interaction buffer $\mathcal{B}_{\text{ref}}$, using a least-squares GAN objective:

$$\mathcal{L}_D = \mathbb{E}_{z\sim\mathcal{B}_{\text{ref}}}\left[(D(z)-1)^2\right] + \mathbb{E}_{z\sim\pi}\left[(D(z)+1)^2\right]. \quad (4)$$

The policy is simultaneously trained to maximize a composite reward $r_t = r_{\text{task}} + \lambda_i r_{\text{interact}} + \lambda_s r_{\text{style}}$, where r_{task} penalizes deviation from the target root command, r_{interact} is derived from the discriminator output, and r_{style} is a standard AMP loss that regularizes motion naturalness:

$$r_{\text{task}}(x_t, c_t) = -\left\|\mathbf{x}_t^{\text{root}} - \mathbf{c}_t^{\text{root}}\right\|_2, \quad (5)$$

$$r_{\text{interact}}(z_t) = \max(0, 1 - 0.25(D(z_t) - 1)^2), \quad (6)$$

$$r_{\text{style}}(z_t) = \max(0, 1 - 0.25(D_{\text{AMP}}(s_t) - 1)^2), \quad (7)$$

By conditioning the discriminator on z_t rather than the full robot state, AIP encourages the policy to reproduce the *geometric signature* of valid interactions without constraining it to a specific kinematic template, allowing the robot to synthesize novel poses for unseen geometries while maintaining stable and physically plausible contact. The resulting policy is denoted π_{full}. More details of the post-training process are provided in Sec. A-B

D. Visual-Motor Policy Distillation

While **LessMimic** achieves strong performance across interaction tasks, it relies on global object information from

TABLE II: **Comparison of baselines, ablations, and LESSMIMIC under object-scale variation in simulation.** Four interaction tasks (PickUp, SitStand, Push, Carry) are evaluated across object scales from 0.4× to 1.6×, with the training scale at 1.0×. Reference-based baselines track fixed demonstration motions; reference-free baselines and ablated variants operate without motion references. Each ablation removes one component from the full system: *AIP* removes the adversarial interaction prior, *Syn.* removes synthetic physicalization, *Rand.* disables geometry randomization, *RL* removes reinforcement learning fine-tuning, and *Trans.* replaces the Transformer with an MLP backbone. Results report mean ± std over 3 random seeds. R_{succ}: task success rate; R_{cont}: hand contact rate over total task duration. **Bold** and underline denote the best and second-best results per column. *: implemented by ourselves.

Method	**PickUp** R_{succ} (↑,%)					**SitStand** R_{succ} (↑,%)			**Push** R_{cont} (↑,%)			**Carry** R_{succ} (↑,%)
	0.4	**0.6**	**1.0**	**1.4**	**1.6**	**0.4**	**1.0**	**1.6**	**0.6**	**1.0**	**1.4**	
Reference-based baselines												
HDMI*	0.0 (±0.0)	99.7 (±0.5)	**100.0** (±0.0)	40.7 (±3.3)	1.7 (±1.2)	0.0 (±0.0)	99.0 (±0.0)	1.7 (±0.5)	0.3 (±0.0)	**97.3** (±0.5)	**10.6** (±2.9)	27.4 (±0.3)
ResMimic*	0.0 (±0.0)	18.3 (±1.2)	**100.0** (±0.0)	**99.7** (±0.5)	63.0 (±10.2)	0.0 (±0.0)	**100.0** (±0.0)	93.7 (±0.5)	0.4 (±0.2)	93.1 (±1.2)	2.2 (±0.7)	32.9 (±1.5)
VisualMimic	0.0 (±0.0)	0.0 (±0.0)	**100.0** (±0.0)	0.0 (±0.0)	0.0 (±0.0)	–	–	–	–	20.6 (±9.0)	–	–
Reference-free baselines												
PhysHSI	23.1 (±2.4)	70.5 (±2.6)	**100.0** (±0.0)	47.9 (±2.8)	39.9 (±2.8)	61.3 (±2.7)	76.2 (±2.3)	71.8 (±2.4)	–	–	–	81.6 (±1.9)
Ablation study												
Ours - AIP	0.0 (±0.0)	0.7 (±0.5)	23.3 (±5.0)	57.3 (±1.2)	64.0 (±2.9)	**98.7** (±0.5)	98.3 (±0.5)	89.0 (±0.8)	**82.9** (±1.1)	72.5 (±1.9)	6.5 (±1.9)	0.0 (±0.0)
Ours - Syn.	34.0 (±6.7)	94.3 (±2.5)	99.3 (±0.9)	**99.7** (±0.5)	**99.7** (±0.5)	95.3 (±1.7)	30.0 (±2.2)	86.3 (±3.9)	0.0 (±0.0)	41.9 (±2.1)	3.5 (±1.6)	66.5 (±2.0)
Ours - Rand.	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	81.3 (±3.4)	**97.7** (±1.2)	0.0 (±0.0)	67.0 (±1.6)	2.1 (±1.0)	0.0 (±0.0)
Ours - RL	0.0 (±0.0)	2.0 (±0.8)	31.7 (±2.6)	37.0 (±2.4)	9.7 (±1.7)	1.7 (±1.2)	64.7 (±3.1)	15.0 (±2.8)	0.0 (±0.0)	67.3 (±0.7)	0.0 (±0.0)	5.3 (±1.5)
Ours - Trans.	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	98.0 (±0.8)	77.7 (±0.5)	95.7 (±2.1)	18.3 (±1.4)	85.7 (±2.4)	3.0 (±0.6)	0.0 (±0.0)
Ours (Mocap)	63.0 (±5.0)	**100.0** (±0.0)	**100.0** (±0.0)	88.0 (±1.6)	94.0 (±1.6)	79.0 (±0.0)	80.3 (±2.1)	61.7 (±1.2)	13.9 (±0.2)	51.3 (±1.6)	2.0 (±0.3)	**82.9** (±1.4)
Ours (Vision)	**63.7** (±3.1)	94.7 (±1.9)	91.0 (±3.3)	**99.7** (±0.5)	93.0 (±1.6)	–	–	–	1.2 (±0.1)	35.2 (±1.3)	1.9 (±0.8)	35.8 (±2.7)

a MoCap system, which is unavailable in most real-world deployments. To relax this assumption, we distill π_{full} into a vision-based policy π_{vis} that operates solely on egocentric depth observations, and evaluate it on the same tasks and metrics as the MoCap-based model to enable direct comparison.

The vision-based policy observes $o_{\text{vis}} = [o_{\text{prop}}, c_t, S_t]$, where S_t denotes a history of egocentric depth frames. It comprises a visual encoder E_ϕ followed by the same control head as π_{full}, where E_ϕ maps S_t to a compact latent z_t—effectively learning to recover the geometric cues previously provided explicitly by I_t.

We adopt a DAgger-style [42] distillation procedure. At each iteration, π_{vis} interacts with the environment to collect trajectories, and the frozen teacher π_{full} is queried at every encountered state to provide supervision. The student minimizes the MSE loss against the teacher's actions:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{s \sim \pi_{\text{vis}}} \left[\| \pi_{\text{vis}}(o_{\text{vis}}) - \pi_{\text{full}}(o_{\text{base}}) \|_2^2 \right]. \tag{8}$$

Throughout distillation, we apply extensive domain randomization to facilitate sim-to-real transfer, including perturbations of camera extrinsics, additive noise on depth observations, and randomization of physical properties. These perturbations encourage E_ϕ to learn geometry-relevant features robust to sensor noise and modeling discrepancies, enabling reliable execution of contact-rich interactions from onboard perception alone. Additional details are provided in Appendix B.

IV. EXPERIMENTS

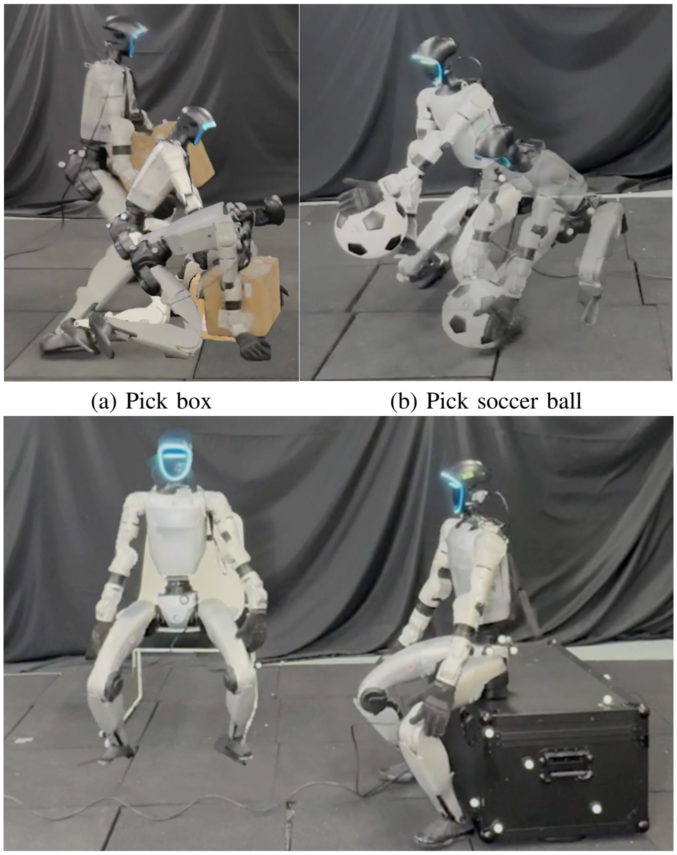
We evaluate **LESSMIMIC** across two complementary dimensions: its **generalization** to variations in object size and shape (Sec. IV-B), and its capability for **long-horizon composition** of diverse interaction skills (Sec. IV-C). Our analysis considers two primary policy variants: a MoCap-based model, π_{full}, and a closed-loop visual-motor model, π_{vis}. To ensure a rigorous assessment, **LESSMIMIC** is compared against representative reference-based and reference-free baselines under identical training and evaluation conditions. We further conduct ablation studies (Sec. IV-E) to isolate the specific contributions of individual system components. Finally, we demonstrate the robustness and transferability of our approach by validating all methods in both high-fidelity simulation and on a physical real-world humanoid platform.

A. Experimental Setup

Tasks: We evaluate on four representative interaction tasks that span a range of contact modes and object configurations. *PickUp* requires lifting a box of varying sizes from the ground using both hands and maintaining a stable grasp. *SitStand* evaluates seated interaction by requiring pelvis contact with chairs of different heights. *Push* focuses on bimanual contact-rich interaction in which the robot pushes a target object while continuously maintaining hand contact. *Carry* requires picking up a box and transporting it to a target location under continuous bimanual contact.

Evaluation Metrics: Task success is defined by geometric and contact-based criteria matched to each task. *PickUp* succeeds if the box is lifted above a height threshold and stably held. *SitStand* succeeds when stable pelvis contact with the chair is established within a valid root height range. *Carry* and multi-task success are measured by tracking the manipulated object along predefined trajectories, with deviations constrained within a fixed tolerance. For *Push*, success is measured by the hand–object contact rate, with body-dominated contact considered failure.

Baselines: We compare **LESSMIMIC** against two state-of-the-art reference-based methods—HDMI [51] and ResMimic [62]—which explicitly track motion demonstrations of object interaction, as well as VisualMimic [56], a reference-based variant that operates with reduced sensory input (depth observations only). PhysHSI [50] serves as the reference-



(a) Pick box

(b) Pick soccer ball

(c) Sit at different heights

Fig. 5: **Real-world generalization of LESSMIMIC.** (a) The policy successfully picks up a box, one of the training geometries. (b) The same policy generalizes to a soccer ball—a spherical object entirely unseen during training—demonstrating shape generalization beyond the training distribution. (c) The policy performs *SitStand* across two chair heights ($12\,\mathrm{cm}$ and $46\,\mathrm{cm}$), maintaining stable pelvis contact across diverse seat geometries.

free baseline. Additional details on motion sources, MoCap processing, controller design, and observation dimensions are provided in Appendix B.

B. Generalization on Object Sizes

Object shape and size variations can substantially affect humanoid interaction success: for large-amplitude motions (*e.g.*, reaching the ground) or long-horizon execution (*e.g.*, 40 task instances sequence), geometric discrepancies accumulate and compound across time steps. To evaluate generalization beyond the training distribution (scale $1.0\times$), we vary object scale from $0.4\times$ to $1.6\times$ across all tasks and, for *PickUp*, additionally vary object shape between box, spherical, and cylindrical geometries (Fig. 1 and Fig. 5). This setting directly tests whether policies adapt based on local geometric relations or rely on memorized motion patterns.

As shown in Tab. II, reference-based methods degrade predictably with scale deviation. HDMI [51] and ResMimic [62] achieve high success near the reference scale but deteriorate sharply at extreme sizes, and VisualMimic [56] exhibits the same trend despite its visual inputs. In contrast, LESSMIMIC demonstrates consistent robustness to geometric variation across all four tasks. For *PickUp*, success rates remain above 90% across most tested scales, degrading only gradually at the smallest object size ($15\,\mathrm{cm}^3$) where all methods struggle. Similar trends hold for *SitStand* and *Push*, where LESSMIMIC maintains stable success and contact rates across scales, outperforming the strongest baselines by 15–20% at larger object sizes. For *Carry*, LESSMIMIC achieves the highest success rate among both reference-based and reference-free methods under scale variation.

TABLE III: **Long-horizon skill composition with increasing task length in simulation.** Success rate (%) for completing sequences of N randomly ordered heterogeneous tasks under a single policy without environment resets. Each ablation removes one component from the full system; see Tab. II for ablation descriptions. All ablated variants collapse to zero beyond short sequences, whereas LESSMIMIC (Mocap) maintains non-zero success up to $N = 40$. Results report mean ± std over 3 random seeds. **Bold** denotes the best result per column.

Method	N=5	N=10	N=15	N=25	N=40
Ablation study					
Ours - AIP	5.2 (±0.2)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Ours - Syn.	22.1 (±0.8)	4.9 (±0.2)	1.0 (±0.3)	0.0 (±0.0)	0.0 (±0.0)
Ours - Rand.	1.9 (±0.1)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Ours - RL	3.2 (±0.2)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Ours - Trans.	1.7 (±0.1)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)
Our method					
Ours (Mocap)	**61.7** (±1.7)	**38.1** (±1.6)	**23.5** (±1.2)	**9.0** (±0.6)	**2.1** (±0.2)
Ours (Vision)	15.9 (±0.6)	2.5 (±0.1)	0.0 (±0.0)	0.0 (±0.0)	0.0 (±0.0)

C. Long-Horizon Skill Composition

Beyond single-skill evaluation, we assess long-horizon execution by randomly composing multiple interaction tasks into a single trajectory applied under a single policy without environment resets. Since reference-based baselines rely on predefined motions and reference-free baselines use task-specific policy, neither can be directly applied to randomly generated task compositions; we therefore compare ablated variants of LESSMIMIC to analyze the contribution of each design choice.

As shown in Tab. III, the full model achieves 62.1% success on 5-task trajectories and maintains 2.1% viability even at 40 sequentially composed tasks. All ablated variants, by contrast, collapse to zero success beyond short sequences, underscoring the necessity of each component. This long-horizon capability emerges from implicit skill transitions driven by the unified DF representation: rather than relying on explicit task sequencing, the policy continuously adapts to changing geometric contexts, enabling sustained execution of heterogeneous task sequences over extended horizons (see Fig. 6).

D. Distilling LESSMIMIC *to Visual Input*

As shown in Tab. II, the vision-based variant preserves the overall performance trends of π_{full} but with a consistent reduction in success rates across tasks. For *PickUp*, performance decreases from near-perfect levels to the 90% range

TABLE IV: **Real-world deployment of LessMimic on a physical humanoid platform.** Both the MoCap-based and vision-based variants are evaluated on *PickUp* ($22\,\mathrm{cm}^3$ and $60\,\mathrm{cm}^3$) and *SitStand* ($12\,\mathrm{cm}$ and $46\,\mathrm{cm}$ seat heights) under repeated executions. R_{succ} reports discrete task success rate; R_{acc} reports root trajectory tracking accuracy, measured via external MoCap for evaluation only. The vision-based variant is not evaluated on *SitStand* due to limited egocentric observability of back-side contacts.

Method	**PickUp** $22cm^3$ (↑,%)		**PickUp** $60cm^3$ (↑,%)		**SitStand** $12cm$ (↑,%)		**SitStand** $46cm$ (↑,%)	
	Real R_{succ}	**Root** R_{acc}	**Real** R_{succ}	**Root** R_{acc}	**Real** R_{succ}	**Root** R_{acc}	**Real** R_{succ}	**Root** R_{acc}
Ours (Mocap)	10 / 10	94.44	8 / 10	81.39	8 / 10	84.89	10 / 10	91.88
Ours (Vision)	8 / 10	89.15	7 / 10	75.24	–	–	–	–

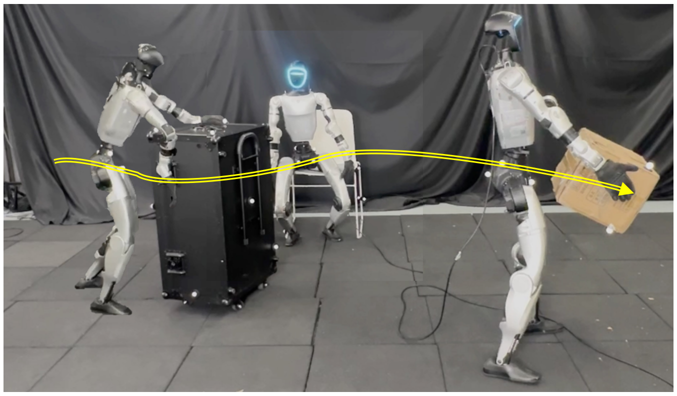


Fig. 6: **Long-horizon skill composition in the real world.** A single **LessMimic** policy executes a sequence of heterogeneous interaction skills without environment resets: pushing a cabinet to a target location, then picking up and carrying a box along the commanded trajectory (indicated by the yellow arrow). This demonstrates the policy's ability to implicitly transition between distinct interaction modes under a unified DF-based representation.

at larger scales and degrades further at smaller scales. Similar reductions are observed in *Push* and *Carry*, where success and contact rates remain below those of the MoCap-based model but are comparable to or exceed reference-free and vision-guided reference-based baselines. Notably, for *PickUp* under scale generalization, the vision-based policy achieves 63.7–99.7% success at scales where all baseline methods exhibit near-zero performance. *SitStand* is excluded from vision-based evaluation due to limited egocentric observability of back-side contacts. Across the remaining tasks, the moderate performance reduction is attributable to perceptual uncertainty introduced by the depth observations.

E. Ablation Study

To isolate the contribution of each component, we evaluate five ablated variants of the full system. As shown in Tab. II, removing the interaction prior (*Ours - AIP*) significantly reduces robustness to scale variation, confirming that AIP is the primary mechanism enforcing geometry-consistent interaction across object geometries. Removing synthetic physicalization (*Ours - Syn*), which replaces teacher-generated physically valid trajectories with raw retargeted MoCap data, most severely affects contact-rich tasks such as *Carry* while leaving others relatively intact, suggesting that data physical feasibility matters most when sustained contact is required. Disabling geometry randomization (*Ours - Rand*) causes severe overfitting with near-zero success outside training scales, confirming its necessity for scale-invariant generalization. Removing RL fine-tuning (*Ours - RL*) yields limited performance and poor generalization, demonstrating that behavior cloning alone is insufficient to bridge the gap to novel geometries. Finally, replacing the Transformer with an MLP backbone (*Ours - Trans*) severely degrades performance on *PickUp* and *Carry*, reflecting insufficient model capacity for capturing the temporal dependencies required in multi-skill interaction.

F. Real-World Deployment

We evaluate **LessMimic** on a physical humanoid platform to assess robustness beyond simulation. Experiments cover *PickUp* and *SitStand* across varying object sizes ($22\,\mathrm{cm}^3$ and $60\,\mathrm{cm}^3$) and chair heights ($12\,\mathrm{cm}$ and $46\,\mathrm{cm}$), with both the MoCap-based and vision-based variants evaluated under identical conditions using repeated executions.

As shown in Tab. IV, real-world performance is largely consistent with simulation trends in Tab. II. The MoCap-based model achieves 10/10 success on *PickUp* ($22\,\mathrm{cm}^3$) and *SitStand* ($46\,\mathrm{cm}$), with root tracking accuracy above 90% in both cases. The vision-based variant achieves 8/10 and 7/10 on the two *PickUp* conditions respectively, with slightly reduced tracking accuracy. For *SitStand*, the MoCap-based model remains robust across seat heights; the vision-based variant is not evaluated due to limited observability of back-side contacts (see project website).

V. Conclusion

We present **LessMimic**, a reference-free framework leveraging DFs as a unified representation for generalizable, long-horizon humanoid interaction. The key insight is that local DF geometry—surface distances, gradients, and velocity decompositions—provides a shape- and scale-invariant signal, enabling a single policy to master diverse skills. A three-stage pipeline of behavior cloning, AIP-guided RL, and visual-motor distillation grounds this into robust, transferable whole-body control. In simulation, **LessMimic** achieves 80–100% success on *PickUp* and *SitStand* across extreme object scales ($0.4\times$ to $1.6\times$). Notably, it executes 5-task trajectories with 62.1% success and remains viable up to 40 sequential tasks—a horizon where all baselines collapse. Real-world deployment confirms these capabilities transfer reliably across varying object geometries and seat heights. Future work will extend the DF-based representation to articulated and deformable objects, and improve robustness under partial observability.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (2025YFE0218200), the National Natural Science Foundation of China (62376009), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

[1] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Conference on Robot Learning (CoRL)*, 2025.

[2] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[3] Eugenio Chisari, Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Learning robotic manipulation policies from point clouds with conditional flow matching. In *Conference on Robot Learning (CoRL)*, 2024.

[4] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[5] Kourosh Darvish, Luigi Penco, Joao Ramos, Rafael Cisneros, Jerry Pratt, Eiichi Yoshida, Serena Ivaldi, and Daniele Pucci. Teleoperation of humanoid robots: A survey. *IEEE Transactions on Robotics (T-RO)*, 39(3):1706–1727, 2023.

[6] Alexander Dietrich, Christian Ott, and Alin Albu-Schäffer. An overview of null space projections for redundant, torque-controlled robots. *International Journal of Robotics Research (IJRR)*, 34(11):1385–1400, 2015.

[7] Yushi Du, Yixuan Li, Baoxiong Jia, Yutang Lin, Pei Zhou, Wei Liang, Yanchao Yang, and Siyuan Huang. Learning human-humanoid coordination for collaborative object carrying. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[8] Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. Demohlm: From one demonstration to generalizable humanoid loco-manipulation. *IEEE Robotics and Automation Letters (RA-L)*, pages 1–8, 2026.

[9] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024.

[10] Tuomas Haarnoja, Ben Moran, Guy Lever, Sandy H Huang, Dhruva Tirumala, Jan Humplik, Markus Wulfmeier, Saran Tunyasuvunakool, Noah Y Siegel, Roland Hafner, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics*, 9(89):eadi8022, 2024.

[11] Kris Hauser. Robust contact generation for robot simulation with unstructured meshes. In *International Symposium on Robotics Research*, 2016.

[12] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning (CoRL)*, 2024.

[13] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[14] Alexander Herzog, Nicholas Rotella, Sean Mason, Felix Grimminger, Stefan Schaal, and Ludovic Righetti. Momentum control with hierarchical inverse dynamics on a torque-controlled humanoid. *Autonomous Robots*, 40(3):473–491, 2016.

[15] Michael A Hopkins, Alexander Leonessa, Brian Y Lattimer, and Dennis W Hong. Optimization-based whole-body control of a series elastic humanoid robot. *International Journal of Humanoid Robotics*, 13(01):1550034, 2016.

[16] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*, 2023.

[17] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

[18] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[19] Xiaogang Jia, Qian Wang, Anrui Wang, Han A Wang, Balázs Gyenes, Emiliyan Gospodinov, Xinkai Jiang, Ge Li, Hongyi Zhou, Weiran Liao, et al. Pointmappolicy: Structured point cloud processing for multi-modal imitation learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[20] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. In *Robotics: Science and Systems (RSS)*, 2021.

[21] Oussama Khatib, Luis Sentis, Jaeheung Park, and James Warren. Whole-body dynamic behavior and control of human-like robots. *International Journal of Humanoid Robotics*, 1(01):29–43, 2004.

[22] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems (RSS)*, 2021.

[23] Jialong Li, Xuxin Cheng, Tianshu Huang, Shiqi Yang, Ri-Zhao Qiu, and Xiaolong Wang. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. In *Robotics: Science and Systems (RSS)*, 2025.

[24] Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, Alessandro Lazaric, Matteo Pirotta, and Guanya Shi. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2025.

[25] Yitang Li, Yuanhang Zhang, Wenli Xiao, Chaoyi Pan, Haoyang Weng, Guanqi He, Tairan He, and Guanya Shi. Hold my beer: Learning gentle humanoid locomotion and end-effector stabilization control. In *Conference on Robot Learning (CoRL)*, 2025.

[26] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *Conference on Robot Learning (CoRL)*, 2025.

[27] Qiayuan Liao, Takara E. Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C. Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

[28] Minghua Liu, Xuanlin Li, Zhan Ling, Yangyan Li, and Hao Su. Frame mining: a free lunch for learning robotic manipulation from 3d point clouds. In *Conference on Robot Learning (CoRL)*, 2022.

[29] Weiheng Liu, Yuxuan Wan, Jilong Wang, Yuxuan Kuang, Xuesong Shi, Haoran Li, Dongbin Zhao, Zhizheng Zhang, and He Wang. Fetchbot: Learning generalizable object fetching in cluttered scenes via zero-shot sim2real. In *Conference on Robot Learning (CoRL)*, 2025.

[30] Qingzhou Lu, Yao Feng, Baiyu Shi, Michael Piseno, Zhenan Bao, and C Karen Liu. Gentlehumanoid: Learning upper-body compliance for contact-rich human and object interaction. *arXiv preprint arXiv:2511.04679*, 2025.

[31] Gianni Lunardi, Thomas Corbères, Carlos Mastalli, Nicolas Mansard, Thomas Flayols, Steve Tonneau, and Andrea Del Prete. Reference-free model predictive control for quadrupedal locomotion. *IEEE Access*, 12:689–698, 2023.

[32] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Zi Wang, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi "Jim" Fan, and Yuke Zhu. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

[33] Federico L Moro and Luis Sentis. Whole-body control of humanoid robots. In *Humanoid Robotics: a Reference*, pages 1161–1183. Springer Dordrecht, 2019.

[34] Yoshihiko Nakamura, Hideo Hanafusa, and Tsuneo Yoshikawa. Task-priority based redundancy control of robot manipulators. *International Journal of Robotics Research (IJRR)*, 6(2):3–15, 1987.

[35] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79(1): 12–49, 1988.

[36] Jia Pan, Sachin Chitta, and Dinesh Manocha. Fcl: A general purpose library for collision and proximity queries. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[37] Xue Bin Peng. Mimickit: A reinforcement learning framework for motion imitation and control. *arXiv preprint arXiv:2510.13794*, 2025.

[38] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[39] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4):1–20, 2021.

[40] Michael Posa and Russ Tedrake. Direct trajectory optimization of rigid body dynamical systems through contact. In *Tenth Workshop on the Algorithmic Foundations of Robotics*, 2013.

[41] Carlos Quintero-Pena, Wil Thomason, Zachary Kingston, Anastasios Kyrillidis, and Lydia E Kavraki. Stochastic implicit neural signed distance functions for safe motion planning under sensing uncertainty. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[42] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[43] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[44] John Schulman, Yan Duan, Jonathan Ho, Alex Lee, Ibrahim Awwal, Henry Bradlow, Jia Pan, Sachin Patil, Ken Goldberg, and Pieter Abbeel. Motion planning with sequential convex optimization and convex collision checking. *International Journal of Robotics Research (IJRR)*, 33(9):1251–1270, 2014.

[45] Luis Sentis and Oussama Khatib. Synthesis of whole-body behaviors through hierarchical control of behavioral primitives. *International Journal of Humanoid Robotics*, 2(04):505–518, 2005.

[46] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[47] Minseok Song, JeongHo Ha, Bonggyeong Park, and Daehyung Park. Implicit neural-representation learning for elastic deformable-object manipulations. In *Robotics: Science and Systems (RSS)*, 2025.

[48] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH Conference Papers*, 2023.

[49] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[50] Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, Qifeng Chen, Jingbo Wang, and Jiangmiao Pang. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025.

[51] Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025.

[52] Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2020.

[53] Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, and Shankar Sastry. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.

[54] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[55] Wen Yang and Wanxin Jin. Contactsdf: Signed distance functions as multi-contact models for dexterous manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 2025.

[56] Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C. Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

[57] Yanjie Ze, Zixuan Chen, Joao Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist: Teleoperated whole-body imitation system. In *Conference on Robot Learning (CoRL)*, 2025.

[58] Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[59] Ryan S Zesch, Vismay Modi, Shinjiro Sueda, and David IW Levin. Neural collision fields for triangle primitives. In *SIGGRAPH Asia Conference Papers*, 2023.

[60] Xiang Zhang, Changhao Wang, Lingfeng Sun, Zheng Wu, Xinghao Zhu, and Masayoshi Tomizuka. Efficient sim-to-real transfer of contact-rich manipulation skills with online

admittance residual learning. In *Conference on Robot Learning (CoRL)*, 2023.

[61] Zhikai Zhang, Jun Guo, Chao Chen, Jilong Wang, Chenghuai Lin, Yunrui Lian, Han Xue, Zhenrong Wang, Maoqi Liu, Jiangran Lyu, Huaping Liu, He Wang, and Li Yi. Track any motions under any disturbances. *arXiv preprint arXiv:2509.13833*, 2025.

[62] Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

[63] Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Appendix A
The LessMimic Framework

This section provides additional implementation details for the three training stages introduced in Sec. III. We elaborate on network architectures, training procedures, and reward formulations to facilitate reproducibility and clarify key design choices. All training hyperparameters are summarized in Tab. A1.

A. Interaction Skill Pre-Training

During interaction skill pre-training, we learn a base policy π_{base} that maps DF-based interaction representations to whole-body actions. Rather than training on static offline data, the policy is trained via DAgger-style distillation using trajectories generated by rolling out π_{base} itself, allowing the training distribution to track the student's evolving behavior.

Teacher Policy: The effectiveness of behavior cloning depends directly on the quality of the demonstrations provided by the teacher. To provide the student policy with initial weights capable of physically valid interactions, the teacher must achieve a high success rate on the collected interaction dataset. We therefore employ ResMimic [62] as π_{mimic}, which augments standard motion tracking with a residual module specifically designed to compensate for dynamic mismatches that arise during contact-rich interaction. This choice ensures that the expert demonstrations are not only kinematically reasonable but also physically feasible under full simulation dynamics—a critical property for the downstream behavior cloning objective.

Network Architecture: We employ a Transformer-based architecture to jointly model the observation structure, interaction encoding, and policy learning objective. The Transformer's ability to capture long-range temporal dependencies across the interaction history makes it particularly well-suited for multi-skill whole-body control, where the relevant geometric context may span many time steps. Critically, this design aligns training-time supervision with inference-time observation settings while eliminating any dependency on reference motions during deployment.

TABLE A1: **Training hyperparameters for each stage of the LessMimic pipeline.** All three stages—interaction skill pre-training, discriminative post-training, and policy distillation—share the same policy architecture but differ in their optimization configurations. The batch size notation $M \times 8$ denotes M samples across 8 parallel environments.

Hyperparameter	Symbol	Value
Interaction Skill Pre-training		
Learning rate	η_{pre}	1×10^{-3}
Batch size	B_{pre}	4096×8
Number of training iterations	N_{pre}	$24,000$
Optimizer	Adam	–
Discriminative Post-Training		
Policy learning rate	η_{pol}	1×10^{-3}
Discriminator learning rate	η_{disc}	2×10^{-4}
Discount factor	γ	0.99
Reward weight	λ	1.5
Entropy coefficient	α_{ent}	5×10^{-4}
Number of environments	N_{env}	4096×8
Number of environment steps	N_{rl}	$240,000$
Policy Distillation		
Learning rate	η_{dist}	1×10^{-3}
Batch size	B_{dist}	2048×8
Number of distillation iterations	N_{dist}	$120,000$

Observations. The observation provided to π_{base} at time step t is defined as

$$o_{\text{base}} = \big[o_{\text{prop}},\ c_t^{\text{root}},\ z_t\big], \tag{A1}$$

where o_{prop} denotes humanoid proprioceptive states, including joint DoF positions and velocities; c_t^{root} is a sparse target root trajectory command specifying the desired root motion; and z_t is the DF-based interaction latent defined in Eq. (2) of the main paper. Importantly, this observation contains no reference motion or task-specific supervision signal, and exactly matches the observation structure used at inference time. This design choice is deliberate: by training π_{base} under the same observation constraints it will face at deployment, we avoid the distributional mismatch that would otherwise arise if the policy were trained with privileged information and then deployed without it.

Interaction Encoder. The interaction history $I_t = \{u_{t-l+1}, \ldots, u_t\}$ captures the recent geometric evolution between the humanoid and the object across a temporal window of length l. To process this sequence into a form suitable for policy conditioning, we utilize a VAE to encode each per-timestep feature u into a low-dimensional latent space. Specifically, the VAE encoder maps each u to a latent representation z using the reparameterization trick to maintain end-to-end differentiability during training. Both encoder and decoder are implemented as MLPs with ReLU activations. The resulting sequence of latents $\{z_{t-l+1}, \ldots, z_t\}$ is then concatenated to form a compact, structured geometric summary of the interaction history, which is passed as input to the policy network. The VAE bottleneck serves two complementary purposes: it smooths out sensor noise present in the raw DF measurements, and it provides a fixed-dimensional input to the policy regardless of the temporal window length l, simplifying the policy architecture.

Policy Architecture. The base policy π_{base} is implemented as a Transformer that takes as input the concatenation of proprioception o_{prop}, root command c_t^{root}, and interaction latent z_t, and outputs whole-body joint actions. All hidden layers use ReLU activation functions; the final output layer uses a linear activation to allow unconstrained action outputs. Notably, the same policy architecture is used across all three training stages—pre-training, post-training, and visual-motor distillation—with only the training objective, supervision signal, and environment configuration varying between stages. This architectural consistency simplifies the overall pipeline and ensures that the weights learned during pre-training can be directly carried over to subsequent stages without any architectural modification.

TABLE A2: **Reward components used in discriminative post-training.** The composite reward consists of task and style rewards that drive interaction quality, and regularization penalties that ensure motion stability and physical safety. Interaction Style operates on the DF-based interaction latent z_t via the AIP discriminator; Motion Style operates on the full robot state via the AMP discriminator to encourage natural gait and posture. Weights listed are baseline values and may be adjusted per task.

Reward Term	**Formulation**	**Weight**	**Description**
Task and Style Rewards			
Root Tracking	$-\|x_t^{\text{root}} - c_t^{\text{root}}\|^2$	1.0	Follow target root trajectory without motion references.
Interaction Style	Eq. (6)	2.0	Discriminator-based prior on distance-field representation.
Motion Style	$1 - 0.25(D(s) - 1)^2$	1.0	Discriminator-based motion prior for natural gait/posture.
Object Tracking	$\exp(-\|x_t^{\text{obj}} - \tilde{x}_t^{\text{obj}}\|^2/\sigma^2)$	1.0	Ensures object follows the desired manipulation trajectory.
Regularization and Penalties			
Action Reg.	$\|\Delta a_t\|^2$	5.0	Penalizes abrupt changes to improve control stability.
Termination	Constant	-10.0	Applied upon early termination (falls or loss of balance).
Joint Limit	n_{exceed}	-5.0	Penalty when joints exceed specified soft limits.

Teacher–Student Setup: The teacher–student framework is designed around a deliberate asymmetry in the information available to each policy. The teacher π_{mimic} follows a motion-tracking formulation with residual learning, granting it access to privileged information including full-body reference motions via o_{mimic}. This privileged access enables π_{mimic} to generate high-quality, physically valid interaction trajectories even for complex contact-rich scenarios. The student π_{base}, by contrast, operates solely on o_{base}, which excludes all reference motion information. This asymmetric information design is intentional: it forces the student to ground its behavior entirely in the DF-based interaction latent z_t rather than relying on reference cues, ensuring that the learned policy is fully compatible with reference-free inference at deployment. The teacher's role is thus not to be imitated directly at inference, but to provide a high-quality supervision signal during training that helps the student discover effective interaction strategies expressible within the reference-free observation space.

Training Objective: Interaction skill pre-training is performed using behavior cloning rather than reinforcement learning. No task-specific reward is introduced at this stage; the sole objective is to minimize the discrepancy between the actions predicted by π_{base} and those generated by π_{mimic} on the same states:

$$\mathcal{L}_{\text{BC}} = \mathbb{E}_{s \sim \pi_{\text{base}}} \left\| \pi_{\text{base}}(o_{\text{base}}) - \pi_{\text{mimic}}(o_{\text{mimic}}) \right\|_2^2 , \tag{A2}$$

where o_{mimic} includes privileged observations and reference motions available only to the teacher and not at inference time. To mitigate covariate shift between the supervised training distribution and the student's own rollout distribution, we adopt a DAgger-style procedure [42]: at each training iteration, π_{base} is rolled out in the environment to generate on-policy trajectories, and π_{mimic} is queried at each encountered state to provide corrective action labels. This on-policy data collection is crucial: it ensures that the student learns to recover from the kinds of states it will actually encounter during its own execution, rather than merely fitting the teacher's behavior on the teacher's state distribution—a distribution the student may never visit if it deviates even slightly from the expert trajectory.

B. Discriminative Post-Training

Discriminative post-training further refines π_{base} to improve robustness and generalization under novel object geometries and interaction conditions. This stage retains the same policy architecture and observation structure as interaction skill pre-training, but differs fundamentally in its training objective, supervision signal, and environment configuration. The key motivation is that behavior cloning on a fixed dataset, however high-quality, inevitably encourages the policy to memorize specific kinematic trajectories rather than learning the underlying geometric rules of interaction. Post-training addresses this by exposing the policy to procedurally varied geometries under a geometry-aware adversarial supervision signal that rewards valid interaction patterns regardless of the specific object shape or scale encountered.

Training Setup: Post-training is conducted in a procedurally augmented simulation environment in which object geometry, scale, and physical properties are randomized across episodes. This procedural variation is essential: it prevents the policy from overfitting to the specific object configurations seen during pre-training, and forces it to discover interaction strategies that generalize across the geometric distribution. In contrast to pre-training, no reference motions are available during this stage, and motion-tracking losses are deliberately excluded. Introducing motion-tracking rewards here would be counterproductive, as the reference motions were collected at a fixed object geometry and would not constitute valid interaction guidance for novel shapes and scales. The policy is instead optimized using reinforcement learning with a geometry-aware reward signal described below.

Adversarial Interaction Prior: To provide a transferable supervision signal without relying on task-specific shaping rewards, we introduce AIP, implemented via a discriminator $D(\cdot)$. The key design choice is that the discriminator operates solely on the DF-based interaction latent z_t rather than on the full robot state. This is crucial: by conditioning on the geometric interaction signature rather than absolute joint configurations, the discriminator captures what a valid interaction looks like in geometric terms—an approach-contact-release pattern expressed in local DF coordinates—without encoding any information about the specific kinematic template used to achieve it. The discriminator is trained to distinguish interaction latents generated by the policy on novel objects from those stored in a reference interaction buffer $\mathcal{B}_{\text{ref}}$ collected during pre-training, using a least-squares GAN objective:

$$\mathcal{L}_D = \mathbb{E}_{z \sim \mathcal{B}_{\text{ref}}} \left[(D(z) - 1)^2 \right] + \mathbb{E}_{z \sim \pi} \left[(D(z) + 1)^2 \right], \quad \text{(A3)}$$

where $\mathcal{B}_{\text{ref}}$ denotes interaction latent samples drawn from the reference buffer and $z \sim \pi$ denotes latents generated by the current policy during rollout. The least-squares formulation is preferred over the standard GAN cross-entropy objective for its more stable gradient behavior during adversarial training.

Policy Optimization: The policy is trained with reinforcement learning using a composite reward of the form

$$r_t = r_{\text{task}} + \lambda_i \, r_{\text{interact}} + \lambda_s \, r_{\text{style}}, \quad \text{(A4)}$$

where r_{task} encourages tracking of the root-level command, r_{interact} is derived from the AIP discriminator output, and r_{style} is derived from an AMP [39] discriminator that regularizes the naturalness of the robot's full-body motion. The separation between r_{interact} and r_{style} is intentional: r_{interact} operates on the geometric interaction signature z_t and rewards valid contact patterns, while r_{style} operates on the full robot state and penalizes unnatural or physically implausible motion. Together, they provide complementary supervision—one ensuring the interaction is geometrically consistent, the other ensuring the resulting motion looks natural. Detailed reward formulations and coefficient values are provided in Tab. A2.

Key Differences from Pre-Training: It is worth summarizing the three essential changes introduced in discriminative post-training relative to the pre-training stage, as they collectively define the post-training's role in the overall pipeline. First, reference motions are removed entirely from the training setup and replaced by the adversarial AIP supervision signal, which provides geometry-aware guidance without requiring any motion demonstrations. Second, the training environment is procedurally randomized across object geometries, scales, shapes, and physical properties, exposing the policy to a broad distribution of interaction conditions far beyond the pre-training dataset. Third, policy optimization is performed using reinforcement learning rather than behavior cloning, allowing the policy to discover novel interaction strategies not present in the original demonstrations. All other components—network architecture, observation definition, and the DF-based interaction representation—remain unchanged from pre-training, ensuring continuity across stages and allowing the post-training to build directly on the stable initialization provided by behavior cloning.

APPENDIX B
EXPERIMENTS

This section provides supplementary details for the experimental evaluation presented in Sec. IV, covering domain randomization strategies, reward design, baseline implementations, and evaluation metrics.

A. Domain Randomization

Extensive domain randomization is applied during both post-training and policy distillation to improve robustness and facilitate sim-to-real transfer. Randomization spans six complementary dimensions that collectively expose the policy to a broad distribution of interaction conditions.

Object Geometry: For each episode, object geometry is randomized by jointly sampling scale and shape. Scale factors are drawn independently along each spatial dimension, allowing both uniform scaling and mild anisotropic deformation to avoid overfitting to isotropic objects. Object shape is alternated between box-like and cylindrical primitives for applicable tasks. Object orientation and placement are further perturbed within bounded ranges to prevent the policy from exploiting canonical initial poses.

Physical Properties: Object mass, surface friction coefficients, and contact restitution are randomized independently at the start of each episode. This exposes the policy to variations in inertial response and contact dynamics that are difficult to model accurately in simulation, encouraging stable interaction under uncertain physical parameters—a property critical for real-world deployment where these parameters are never precisely known.

Initial Conditions: The initial poses of both the humanoid and the manipulated objects are perturbed with small translational and rotational offsets. Joint configurations are randomized around nominal standing poses while maintaining balance constraints. This prevents the policy from learning to exploit fixed initial configurations, which would cause it to fail when deployed in environments where precise initialization is unavailable.

Command Perturbation: Target root trajectory commands are injected with bounded stochastic noise in both position and heading direction. This encourages the policy to remain stable under imperfect command execution—an important property for long-horizon task composition, where small tracking errors can compound across successive interaction phases.

Actuation Noise: During post-training, zero-mean Gaussian noise is added to the policy's action outputs before they are applied to the low-level whole-body controller. This simulates actuation uncertainty arising from motor variability and controller latency, and has been found to improve control smoothness and stability during both training and real-world deployment.

Perceptual Randomization: For vision-based policy distillation, egocentric depth observations are augmented with camera pose jitter, depth quantization artifacts, random dropout, and additive sensor noise. Camera extrinsic parameters are additionally randomized across episodes to improve robustness to mounting inaccuracies that are common when deploying onboard sensors on a physical humanoid platform.

B. Reward Design

The reward function used during discriminative post-training is summarized in Tab. A2. The overall reward structure is organized around two complementary objectives: primary rewards that drive task completion and interaction quality, and auxiliary regularization terms that stabilize training and promote safe long-horizon execution.

The primary rewards consist of four components. Root trajectory tracking penalizes deviation of the humanoid's root position from the commanded trajectory, providing the primary task-level signal without requiring any motion reference. Interaction style, derived from the AIP discriminator operating on the DF-based interaction latent z_t, encourages the policy to reproduce geometrically consistent interaction patterns across object geometries. Motion style, derived from an AMP [39] discriminator operating on the full robot state, regularizes the policy toward natural and physically plausible humanoid motion. Object tracking provides an additional task-level signal by rewarding the policy when the manipulated object follows its desired trajectory within a tolerance σ.

The auxiliary regularization terms address practical concerns that arise during long-horizon execution and real-world deployment. Action regularization penalizes abrupt changes in control outputs to promote smooth, stable behavior and reduce mechanical wear on the robot. Early termination penalties discourage motions that lead to falls or loss of balance. Soft joint limit penalties prevent excessive joint excursions that could cause actuator overheating or mechanical stress during extended real-world operation—a concern that becomes increasingly important as task horizons grow. Together, these components form a balanced reward structure that enables robust interaction learning while maintaining motion naturalness and physical safety.

C. Baselines

All baseline methods are evaluated under the same simulator, control frequency, and low-level whole-body controller as **LessMimic**. All policies are trained and evaluated on the same task definitions, object configurations, and termination conditions, and no baseline has access to the DF-based interaction representation used by **LessMimic**. Reference-based methods explicitly condition policy execution on motion demonstrations or reference trajectories at inference time, while reference-free methods operate on simplified task-relevant observations. HDMI and ResMimic are each trained as a single policy that tracks task-specific reference motions across all tasks, while VisualMimic and PhysHSI are trained separately per task using task-specific planners or reward functions.

HDMI [51] is reproduced following the training and inference procedures described in the original paper. Motion references are generated from retargeted human motion data corresponding to each interaction task, and the policy tracks these references using full-body pose and velocity supervision. The reference motions are fixed at a nominal object scale; no adaptation mechanism is applied when object geometry deviates from the training distribution, which is precisely the generalization condition evaluated in our experiments.

ResMimic [62] is reproduced using its residual motion-object co-tracking formulation, in which a base motion tracking policy is augmented with a learned residual controller to compensate for contact dynamics. We train a single ResMimic policy across all tasks using task-specific reference motions. During evaluation, the policy tracks the same reference motions without modification under object scale variation.

VisualMimic [56] serves as a reference-based baseline that conditions the policy on egocentric visual observations in addition to motion references. Rather than training from scratch, we use the official pre-trained checkpoints provided by the authors to ensure a faithful evaluation. The policy tracks reference motions generated by a visual planner conditioned on egocentric observations, but remains subject to the architecture's fundamental limitation of not adapting explicitly to unseen object geometries.

PhysHSI [50] is a reference-free baseline that conditions the policy on handcrafted observation features encoding humanoid state and object-relative information. We use the original codebase and pre-trained checkpoints provided by the authors. While PhysHSI uses task-specific reward functions to facilitate interactions, it lacks a unified interaction representation, necessitating separate observation and reward designs for each task—a limitation that prevents it from being applied directly to the long-horizon multi-task setting evaluated in Sec. IV-C.

For all baselines, simulation parameters are matched to those used by **LessMimic**, and hyperparameters are tuned according to the original papers when available. When hyperparameters are unspecified, we select reasonable defaults to ensure stable training. All baseline results are averaged over three random seeds using identical evaluation protocols.

D. Evaluation Metrics

Success Rate: Task success is defined by geometric and contact-based criteria tailored to each task. For *PickUp*, the task is considered successful if the box is lifted above 0.3 m from the ground and held stably for at least 3 s. For *SitStand*, success requires the humanoid to establish stable contact with the chair with the root height falling within [0.3 m, 0.6 m], which filters out near-fall cases where the robot makes incidental low contact. For *Carry*, the robot must pick up the box, transport it along a given trajectory, and place it at the destination, with a maximum allowable deviation of 0.6 m at any point during execution. For long-horizon evaluation, the same trajectory-following criterion is applied across all tasks in the composed sequence, with a per-step deviation tolerance of 0.6 m.

Contact Rate: For the *Push* task, discrete success is not a well-defined criterion since the task requires maintaining continuous hand contact rather than reaching a terminal goal state. We therefore evaluate performance using the contact rate R_{cont}, defined as the proportion of time steps during which the humanoid's end-effectors maintain active contact with the target object. High contact rates indicate sustained and controlled interaction, while low rates suggest flickering or failed contact—both of which would result in the object not being pushed along the desired trajectory.

Root Tracking Accuracy: For real-world deployment evaluation, discrete task success alone is insufficient to characterize execution quality across repeated trials. We therefore report root trajectory tracking accuracy R_{acc}, which measures how consistently the humanoid follows the commanded root trajectory over the course of task execution. Specifically, we compute the Euclidean distance between the humanoid's root position and the commanded trajectory at each time step, and report the proportion of time steps for which this distance remains within $0.6\,\text{m}$. This metric provides a continuous measure of execution stability and is particularly informative for assessing whether performance differences between the MoCap-based and vision-based variants arise from high-level task failure or from degraded trajectory tracking quality.