SynAgent: Generalizable Cooperative Humanoid Manipulation via Solo-to-Cooperative Agent Synergy

Wei Yao*, Haohan Ma*, Hongwen Zhang, Yunlian Sun, Liangjun Xing, Zhile Yang, Yuanjun Guo, *Member, IEEE* Yebin Liu, *Member, IEEE* and Jinhui Tang, *Senior Member, IEEE*

***Abstract*—Controllable cooperative humanoid manipulation is a fundamental yet challenging problem for embodied intelligence, due to severe data scarcity, complexities in multi-agent coordination, and limited generalization across objects. In this paper, we present SynAgent, a unified framework that enables scalable and physically plausible cooperative manipulation by leveraging Solo-to-Cooperative Agent Synergy to transfer skills from single-agent human-object interaction to multi-agent human-object-human scenarios. To maintain semantic integrity during motion transfer, we introduce an interaction-preserving retargeting method based on an Interact Mesh constructed via Delaunay tetrahedralization, which faithfully maintains spatial relationships among humans and objects. Building upon this refined data, we propose a single-agent pretraining and adaptation paradigm that bootstraps synergistic collaborative behaviors from abundant single-human data through decentralized training and multi-agent PPO. Finally, we develop a trajectory-conditioned generative policy using a conditional VAE, trained via multi-teacher distillation from motion imitation priors to achieve stable and controllable object-level trajectory execution. Extensive experiments demonstrate that SynAgent significantly outperforms existing baselines in both cooperative imitation and trajectory-conditioned control, while generalizing across diverse object geometries. Codes and data will be available after publication. Project Page: https://yw0208.github.io/synagent/.**

***Index Terms*—Embodied Intelligence, articulated character control, multi-agent coordination, physics-based simulation, skill transfer.**

I. Introduction

EMBODIED intelligence has emerged as a key research frontier due to its potential to substantially improve productivity. However, most existing work focuses on single-robot locomotion [1], [2] and isolated object manipulation [3], [4]. In contrast, future large-scale deployments will require robots to operate synergistically in shared and dynamic environments. In such settings, two capabilities are crucial: **cooperative manipulation** for multi-robot collaboration and **precise control** for reliable motion execution. Developing these capabilities is essential for advancing from isolated robotic skills to synergistic multi-agent systems capable of complex real-world tasks.

Despite their importance, robust cooperative manipulation and precise control remain challenging to achieve. A key limitation lies in the scarcity of high-quality large-scale datasets that capture human–object–human interactions (HOHI). Most existing datasets focus on single-person motion [5], dual-human interaction [6], [7], and human–object interaction (HOI) [8]–[10], while available HOHI data [11] is limited in scale and often ill-suited for learning physically grounded control policies. In addition, cooperative manipulation is inherently much more complex than single-agent manipulation: the joint action space grows exponentially with the number of agents, leading to pronounced difficulties in optimization, convergence, and training stability. As a result, even methods that perform well in restricted settings often struggle to generalize to diverse interaction patterns, novel object geometries, and unseen coordination scenarios.

To overcome these hurdles, as shown in Figure. 1, we present **SynAgent**, a unified framework designed to bridge the chasm between single-agent motor proficiency and multi-agent collaborative synergy. Our first contribution addresses the fundamental data bottleneck through an **interaction-preserving retargeting** strategy. Direct motion transfer often suffers from severe skeletal discrepancies, so we utilize a differentiable SMPL-X proxy to bridge the morphology gap between human performers and humanoid agents. We then construct an *Interact Mesh* via Delaunay tetrahedralization [12], which effectively encapsulates the local spatial structure between agent joints and object vertices. By minimizing the Laplacian deformation energy of interact meshes during the retargeting process, our method explicitly maintains the semantic topological integrity and contact relationships.

Second, we introduce a **Solo-to-Cooperative Agent Synergy** paradigm. We observe that cooperative manipulation can be conceptually reformulated as a single-agent control problem subject to external force disturbances. This insight allows us to leverage abundant, high-quality single-human HOI data to bootstrap robust motion imitation priors. These individual skills are subsequently evolved into multi-agent synergy through a decentralized training scheme with shared weights, optimized via Multi-Agent Proximal Policy Op-

Wei Yao, Yunlian Sun are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: wei.yao@njust.edu.cn; yunlian.sun@njust.edu.cn.

Haohan Ma, Zhile Yang, Yuanjun Guo are with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China. E-mail: hh.ma2@siat.ac.cn; zl.yang@siat.ac.cn; yj.guo@siat.ac.cn.

Hongwen Zhang is with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China. Email: zhanghongwen@bnu.edu.cn.

Liangjun Xing, Yebin Liu are with the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: xlj24@mails.tsinghua.edu.cn; liuyebin@mail.tsinghua.edu.cn.

Jinhui Tang is with the College of Artificial Intelligence, Nanjing Forestry University, Nanjing 210023, China. E-mail: tangjh@njfu.edu.cn

*These authors contributed equally to this work (co-first authors).

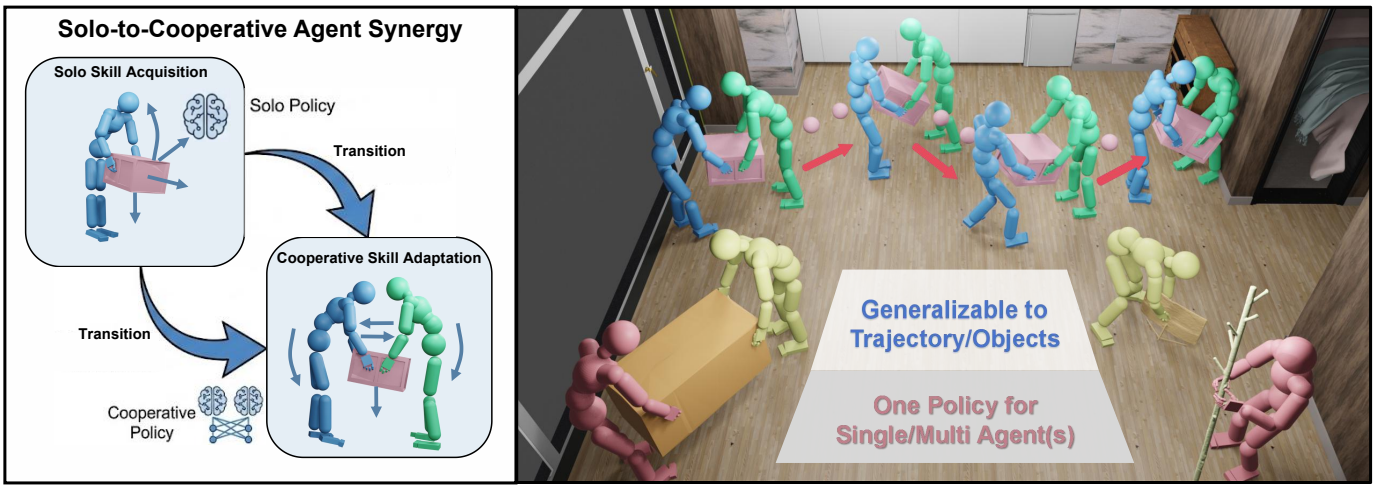


Fig. 1. **Features of SynAgent.** As the first model to address trajectory-following object manipulation with multiple humanoid agents, SynAgent generalizes across diverse object geometries and supports cooperative manipulation.

timization (MAPPO) [13] to foster emergent collaborative behaviors. Finally, to achieve precise execution, we develop a **trajectory-conditioned policy** instantiated as a conditional VAE (CVAE), which integrates agent states, object trajectories, and geometry-aware interaction graphs. To mitigate the ill-posed nature of trajectory-following, we employ a **Multi-Teacher Distillation** framework. By distilling the collective expertise of specialized imitation teachers into a unified student model through a progressive DAgger-style schedule, our framework produces stable, coherent, and highly generalizable multi-agent manipulation motions across diverse object geometries.

In summary, our primary contributions include:

- A novel interaction-preserving retargeting method utilizing an Interact Mesh representation to explicitly maintain the semantic integrity and spatial relationships of HOHI data during motion transfer.
- A **Solo-to-Cooperative Agent Synergy** paradigm that bootstraps scalable multi-agent cooperative behaviors from abundant single-human interaction priors via distilling and decentralized learning.
- A multi-teacher distillation framework that synthesizes diverse motion priors into a unified trajectory-conditioned policy, enabling robust generalization across varying object geometries and trajectories.

II. Related Work

A. Motion Imitation and Retargeting

Physics-based motion imitation learns control policies via reinforcement learning to track reference motions in simulation, enabling physically plausible behaviors, as demonstrated by prior work such as DeepMimic [14]and Mimickit [15], which commonly adopts policy optimization methods like PPO [16]. Its greatest value lies in building a bridge from the digital virtual world to the real physical world. By simply constructing motions, motion imitation can be used to drive agents to perform in physical environments. It successfully decomposes complex robotic tasks into two subtasks generation and imitation. Prior work has demonstrated compelling results in structured settings, such as text-to-motion generation [17], sports-related interactions [18]–[20] and object transportation [21], often by leveraging imitation learning and task-specific motor skills. However, HOHI imitation, especially for humanoid agents with hands, remains an unsolved problem, and we have made a preliminary attempt. Prior to this, [22] achieved handless HOHI motion imitation, but it was limited to imitation only. In this paper, we go further and use the knowledge learned from imitation to achieve more complex trajectory controlling.

Motion retargeting aims to adapt existing motion data to new embodiments or actors. Recent works [23], [24] have achieved impressive progress in this domain, but most of them are not directly designed for scalable multi-human interaction scenarios. For instance, Tairan et al. [25] primarily focus on retargeting motions to robotic, without addressing human–object interaction. Meanwhile, methods such as OmniRetarget [26] and SPIDER [27] extend retargeting to human-object and are capable of handling interactive motions. But they do not support coordinated retargeting across multiple interacting characters. In contrast, we use interaction-preserving retargeting to refine and enhance HOHI data, enabling effective optimization of multi-character interactions.

B. Physics-based Human-Object Interaction Generation

Motion generation [28], [29] is gradually becoming a popular research area. Physics-based human–object interaction (HOI) generation has gained increasing attention as a means to synthesize realistic and dynamically consistent interactions between humans and objects in simulation. Existing HOI datasets have provided a strong foundation for learning interaction behaviors, and extensive research has explored a wide spectrum of HOI generation tasks, ranging from hand–object interactions [30]–[36] and static HOI generation [37]–[42] to full-body and dynamic interaction synthesis [43]–[55].

However, compared to HOI, high-quality and large-scale datasets capturing HOHI interactions remain relatively

scarce [56], [57], like the newly emerged CORE4D [11]. CORE4D focuses exclusively on two-person interaction synthesis. So, its interactions are primarily constrained at a kinematic level, with limited physical fidelity. Scarce data is not the only problem, existing HOHI approaches still face notable limitations. Progress in multi-agent reinforcement learning and physics simulation has been made in the past [58]–[61]. For example, CooHOI [62] achieves collaborative multi-human interaction with objects by explicitly structuring cooperative behaviors, but it simplifies all objects into unified box-like proxies. Furthermore, CooHOI constructs an overly engineered reward system that is specifically tailored for box manipulation, which severely limits its generalizability. On the contrary, our SynAgent adopts a scalable data-driven approach. Using HOI data cleverly compensates for the shortage of HOHI data. Without meticulously tuned rewards, our method can generalize to different geometries and various cooperation modes.

III. Method

A. Interaction-Preserving Motion Retargeting

As shown below, the entire data processing workflow is divided into four steps. Overall, our retargeting method is a gradient descent-based optimization algorithm. Due to the non-differentiability of the forward kinematics of the simulated agent, we use shape fitting to construct a Bridge SMPL-X model, transforming the problem into retargeting between two SMPL-X models with different shapes. Then, based on the Interact mesh, we construct an optimization function Equation. 2 and obtain the retargeted motion sequence through gradient descent optimization. Finally, we apply additional smoothing to the data and filter out low-quality data.

Shape Fitting Due to skeletal discrepancies between MoCap actors and simulated humanoid agents, directly transferring MoCap data often results in incorrect interaction, as highlighted by the red circles in Figure 2. These artifacts significantly hinder training, since agents fail to manipulate objects even when closely following the reference motion. This is why we need motion retargeting. A central challenge is that the simulated agent's native forward kinematics pipeline is non-differentiable, which prevents gradients from propagating through losses defined on joint positions. To mitigate this issue, i.e., , we perform shape fitting via gradient-based optimization [63]. So, we adopt SMPL-X [64] as a differentiable proxy. By instantiating an SMPL-X template whose body shape matches the proportions of the target agent, we reformulate retargeting as a differentiable mesh-to-mesh alignment problem between two SMPL-X instances, bypassing the non-differentiability of the agent's kinematics. The shape fitting objective is defined as

$$\beta_b^* = \arg\min_{\beta_b} \left\| \mathcal{J}\left(\theta_0, \beta_b\right) - p^{\text{target}}\ \left(q_0\right) \right\|^2 \tag{1}$$

where $\beta_b \in \mathbb{R}^{10}$ denotes the SMPL-X shape parameters, $\mathcal{J}$ and p^{target} are the SMPL-X and agent forward kinematics that outputs joint positions, θ_0 and q_0 are joint rotation parameters of SMPL-X and agent in the T-Pose state. Through multiple iterations of optimization, we obtain an SMPL-X shape β_b^* that closely matches the agent, as shown in the *Bridge SMPL-X* in Figure 2.

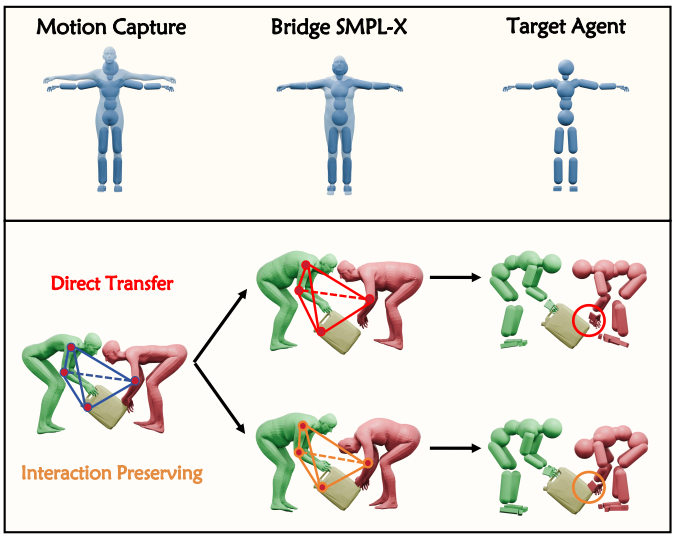


Fig. 2. **Interaction-Preserving Retargeting.** (1) The upper part shows the difference between the MoCap actor and the actual agent's skeletons. And we bridge this difference using shape fitting. (2) In the lower part, the red shows the result of directly retargeting motion capture joints onto the agent, leading to incorrect interaction relationships. The brown area shows how we constructed tetrahedrons to describe the interaction relationships, maintaining the tetrahedron's invariance to ensure the interaction relationships remained unchanged after retargeting.

Retargeting To preserve interaction semantics in complex human-object and human-object-human scenarios, we introduce an *Interact Mesh* constructed via Delaunay tetrahedralization. For each interaction frame, two joints from one human, one object mesh vertex, and one joint from the other human form a tetrahedron that captures the local interaction structure. We represent each tetrahedron $\{p_1, p_2, p_3, p_4\}$, using Laplacian coordinates $L(p_i) = \sum_{i=1}^{4} \{p_i - p_j\}_{j \neq i}$, i.e., a 4×3 matrix. By enforcing L's invariance during retargeting, we thereby preserve relative spatial relationships among agents and objects. The retargeting objective at time step t is formulated as

$$\begin{aligned} \hat{q}_t^* = \arg\min_{q_t} \sum_i \left\| L(p_{t,i}^{\text{source}}) - L\big(p_{t,i}^{\text{target}}(q_t)\big) \right\|^2 + \|q_t - q_{t-1}\|^2 \\ + \max(0, q_{\min} - q_t) + \max(0, q_t - q_{\max}) \\ + \max\big(0, v_{\min}\cdot dt - (q_t - q_{t-1})\big) \\ + \max\big(0, (q_t - q_{t-1}) - v_{\max}\cdot dt\big) \\ + \|p_t^F - p_{t-1}^F\|^2, \quad \forall\, v_{\text{horizontal}}^F < 0.01 \text{ m/s}, \end{aligned} \tag{2}$$

where the first term enforces Laplacian consistency of the Interact Mesh, and the remaining terms impose temporal smoothness, joint position and velocity limits, and a foot-sliding penalty. Together, these constraints yield retargeted motions that are physically plausible and faithful to the original interactions.

Smooth Although we incorporate smoothness terms in the retargeting objective, the optimized motion sequence may still exhibit jitter. To address this, we apply a post-processing step to the root translation and joint rotations. For the root trajectory

$\mathbf{t}$, we employ Sobolev norm regularization. By solving linear equations $(I + \alpha(D^2)^T D^2)\mathbf{t}^* = \mathbf{t}$, we get smoothed root trajectory $\mathbf{t}^*$, where α is regularization parameter and D^2 is second-order difference operator matrix. For joint rotations, we leverage the pre-trained SmoothNet [65] filter, a lightweight temporal network designed specifically for motion de-jittering. It processes the rotation sequence in a sliding-window manner, effectively suppressing jitter while preserving the essential kinematic posture. This decoupled approach ensures both positional and orientational smoothness in the final retargeted motion.

Filter Despite prior refinement, HOHI motion data remains noisy, and some failure cases cannot be corrected by retargeting alone, making training impractical. We therefore propose a *train-to-filter* strategy that iteratively cleans the dataset in a performance-driven manner. The policy is first trained on the full data, and the average inference episode length σ is recorded. Motion clips that consistently terminate early, i.e., with episode lengths below σ, are treated as low-quality and removed. The policy is then retrained on the filtered data, and this process repeats until σ converges. This automatic filtering allows the model to focus on reliable demonstrations, leading to improved training stability and final performance. Note that evaluation set used in subsequent experiments is fixed, so there is not using training to select good examples to improve metrics.

B. Single-Agent Pretraining for Cooperation

1) Reinforcement Learning Formulation: Our policy follows a single-agent paradigm: at each control step, it observes the state of one agent and outputs only that agent's action. During deployment, multiple agents are independently controlled by identical copies of this shared policy, enabling decentralized cooperation. We describe the formulation from four aspects.

State To capture motion dynamics, the state is represented by two consecutive observations $(o_t, o_{t+\Delta t})$. Each observation o consists of three components: an agent–object observation o^{ao}, an interaction graph ig, and a delta interaction graph $\Delta ig = ig - ig_{\text{ref}}$. The interaction graph encodes interaction cues as vectors from agent joints to their nearest object mesh vertices, explicitly providing geometric contact information. The agent–object observation o^{ao} is further divided into agent observations o^{a} and object observations o^{o}. For each agent, o^{a} includes joint positions $\mathbf{p}$, joint rotations $\mathbf{q}$, linear velocities $\dot{\mathbf{p}}$, angular velocities $\dot{\mathbf{q}}$, contact indicators $\mathbf{c}$, as well as their deviations from reference motion capture signals (e.g., $\Delta\mathbf{q} = \mathbf{q} - \mathbf{q}_{\text{ref}}$). The object observation o^{o} contains the object position $\mathbf{p}_o$, orientation $\mathbf{q}_o$, linear velocity $\dot{\mathbf{p}}_o$, angular velocity $\dot{\mathbf{q}}_o$, together with corresponding deviations from the reference object trajectory.

Reward The reward function is designed to encourage accurate motion imitation while maintaining physical realism during cooperative manipulation. It can be summarized as

$$R = e^{-\lambda_\Delta(\Delta\cdot\omega)} \cdot e^{-\lambda_c \sum \|\hat{c}-c\|\odot\hat{c}} \cdot e^{-\lambda_v \sum \|v\| - \lambda_f \max\|f\|}, \quad (3)$$

where the first term is an *imitation reward* that penalizes deviations Δ between simulated states and reference motions. Based on inferred contact labels, we divide the human–object distance into three regions: a contact-promotion zone within an adaptive threshold σ, a neutral buffer zone, and a contact-penalty zone that discourages penetrations. Contact zones are defined by the joint-to-mesh distance. Distances < 0.07 m fall into the contact zone (label 1, encouraging contact); 0.07 m to 0.2 m is a buffer zone (label 0, ignored); and > 0.2 m is the penalty zone (label -1, penalizing undesired contact). The contact term rewards alignment between simulated contacts c and reference contacts $\hat{c}$, while suppressing undesired contacts. In addition, an *energy efficiency* term penalizes high-frequency joint motion and excessive contact forces, encouraging smooth and physically plausible behavior. In Eq 3, λ and ω denote groups of hyper-parameters associated with different state components. Importantly, we use the same set of hyper-parameters across all objects, motions, and tasks, without task-specific tuning, highlighting the robustness and generality of our method.

Policy & Action We adopt a straightforward Multi-Layer Perceptron (MLP) to parameterize our policy within the standard actor-critic architecture. The network receives the concatenated observation vector $(o_t, o_{t+\Delta t})$ as defined above and output actions for one agent. As the agent we used is a whole-body humanoid with dexterous hands, the actions are defined as $a_t \in \mathbb{R}^{51\times 3}$, which are joint PD targets using the exponential map. Finally, these actions a_t will be converted into torques applied to agent joints.

2) Training Stages of Imitation Policies: Overall, our core idea is to learn fundamental skills from simple tasks and progressively transfer them to more complex scenarios involving two agents. Throughout this process, we keep the network architecture, hyper-parameters, and reward function unchanged.

Stage I In the first step of Stage I, we partition the single-human data into N subsets and train a separate policy on each subset, corresponding to the blue networks in Stage I of Figure 3. Subsequently, we perform three operations. First, we augment the single-human data by duplicating the original human. The duplicate is either placed identically to the original or randomly positioned nearby. On this augmented data, we proceed to train a multi-agent policy. For the two agents in the scene, we employ two networks with shared weights, initialized with the previously trained single-agent policy. A key design is the joint training of this policy using the Multi-Agent Proximal Policy Optimization (MAPPO) algorithm, whose core objective is given by

$$\pi_c = \arg\min_{\pi_c} \frac{1}{|\text{agent}|T} \sum_{agent\in\mathcal{S}} \sum_{t=0}^{T} \left(\pi_c(o_t, o_{t+\Delta t}) - \hat{R}_t\right)^2 \quad (4)$$

where π_c is the critic network, $|\text{agent}|$ is the number of agent in a simulated scene $\mathcal{S}$, and $\hat{R}_t$ is the real reward. This function aims to integrate information from all agents to jointly optimize the value function model, thereby helping the policy learn collaborative skills. At last, as illustrated at the bottom of Stage I in Figure 3, both agents are trained simultaneously in a shared environment. We disable collisions between agents but enable collisions between each agent and

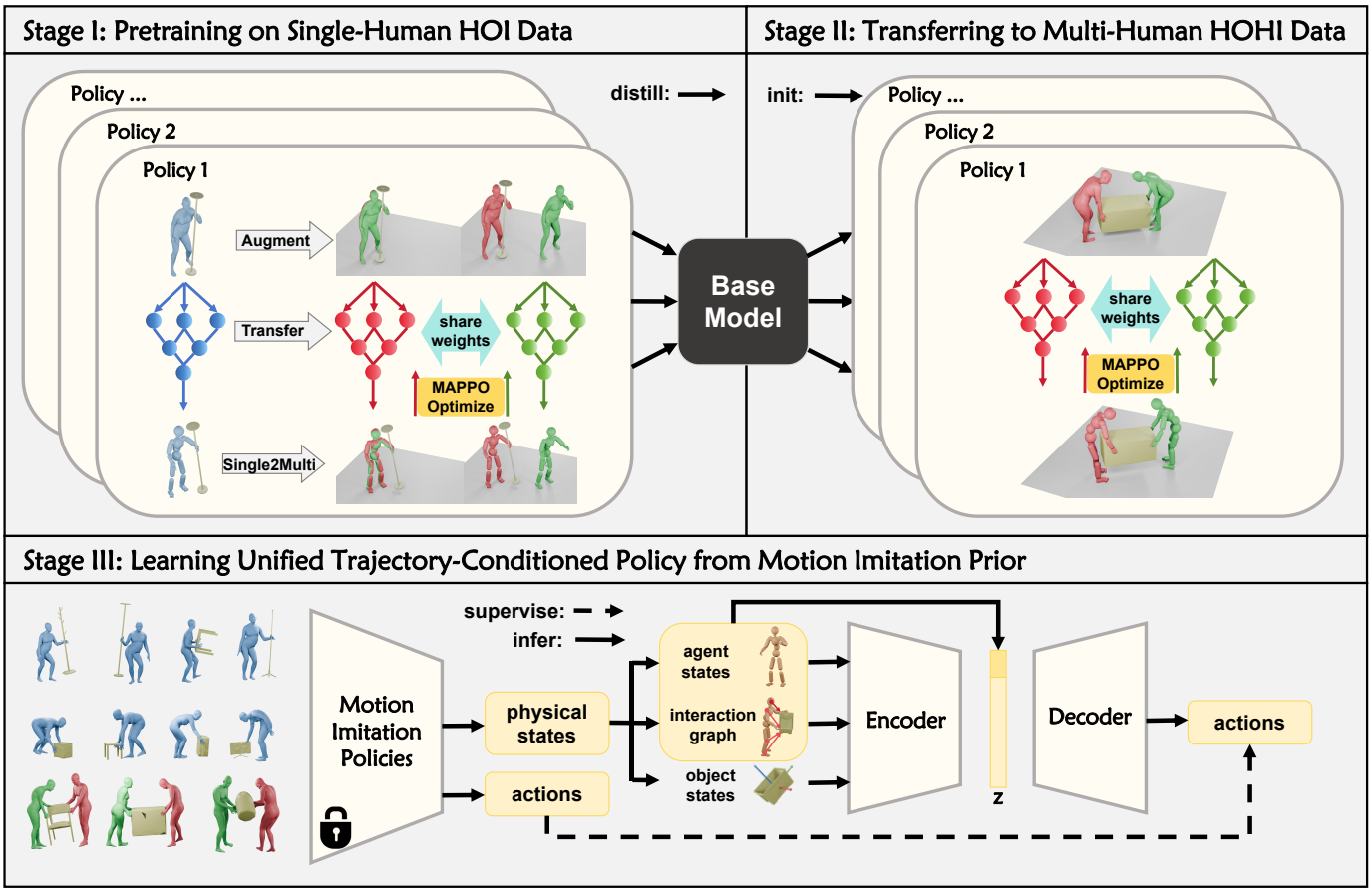


Fig. 3. **Overview of SynAgent Training Pipeline.** (1) Stage I pre-trains imitation policies $\{\pi_i^s\}_{i=0}^N$ on single-human HOI data, then adapts them to multi-agent scenarios with MAPPO algorithm. (2) After distilling $\{\pi_i^s\}_{i=0}^N$ into a unified Base Model, Stage II adapts the Base Model to multi-human HOHI data and get policies $\{\pi_i^m\}_{i=0}^M$. (3) Stage III learns a trajectory-conditioned cVAE policy. Motion imitation policies $\{\pi_i^s\}_{i=0}^N$ and $\{\pi_i^m\}_{i=0}^M$ are used to provide physical states as refined training data, and actions used to supervise learning, which stabilize and boost the model training.

the object. This design allows us to directly leverage single-human data to train multi-agent cooperative skills without forgetting the single-agent abilities, and the agents can perceive each other's presence through the object's dynamics. Note inter-agent collisions are disabled only during Stage I. Because we initialize multi-agent training using augmented single-person HOI data, the two agents naturally overlap in spatial coordinates. During subsequent training and inference, collision detection is enabled.

Stage II After obtaining a set of policies $\{\pi_i^s\}_{i=0}^N$ trained on single-human HOI data, we unify them by distilling all these policies into a shared *Base Model* with an identical network architecture. The distillation procedure follows a similar scheme as described later in Section III-C, and we therefore omit the details here. The purpose of this distillation step is to provide a well-initialized model for the subsequent training on multi-human HOHI data. Our experiments in Table III confirm that this initialization is crucial: training directly from scratch leads to a significant drop in performance, whereas leveraging knowledge priors distilled from single-human data substantially accelerates and stabilizes the learning of multi-agent cooperative skills. A set of policies $\{\pi_i^m\}_{i=0}^M$ are trained on M subsets of multi-human HOHI data. These $\{\pi_i^s\}_{i=0}^N$ and $\{\pi_i^m\}_{i=0}^M$ constitute the *Motion Imitation Policies* in Stage III, which will provide state-action pairs for supervision during the final training.

C. Generalizable Trajectory-Conditioned Policy

As shown in Stage III of Figure 3, our final model, **SynAgent**, is essentially a conditional Variational Autoencoder (VAE). Unlike the imitation policy described earlier, SynAgent removes all observations related to the reference state of the agent from its input. The encoder receives three streams of information: the current *agent states* o^a, the *interaction graph* ig, and the *object states* o^o —where the object states o^o still include the reference object state as a control signal. The condition for our final model is defined as a $T \times 3$ sequence of translations, representing the 3D position of the object across T timesteps. The encoder outputs a latent vector z, which is then concatenated with o^o, and passed to the decoder to generate the corresponding actions a. In the following, we introduce two core designs of our training algorithm that enable effective learning of this trajectory-conditioned generative policy, which are also illustrated in Algorithm 1.

Distill from Imitation to Generation Unlike motion imitation, trajectory-conditioned control is an ill-posed problem. There exist infinitely many possible action sequences that can move an object along a given trajectory, which often leads

Algorithm 1 Learning from Motion Imitation Prior

Require: Motion imitation policies π^{imit} composed of $\{\pi_i^s\}_{i=0}^N$ and $\{\pi_i^m\}_{i=0}^M$, SynAgent policy parameters ψ, SynAgent value function parameters ϕ, DAgger hyper-parameter ϵ and κ, horizon length H, max rounds T_{imit} of imitation training

for $t = 0, 1, 2, \ldots$ **do**
  for $h = 1$ **to** H **do**
    Sample a variable $u \sim \text{Uniform}(0, 1)$
    Collect s^e, a^e from imitation polices π^{imit}
    Obtain a from $\pi_\phi(a \mid s^e)$
    if $u \leq \max\left(1 - \max\left(\frac{t-\kappa}{\epsilon}, 0\right), 0\right)$ **then**
      Given s^e, execute a^e, observe s'^e
    else
      Given s^e, execute a, observe s'^e
    end if
    Store the transition (s^e, s'^e, a, a^e)
  end for
  Update ϕ according to Eq. 4
  Compute $J(\psi) = \|a - a^e\|$
  Compute $w = \min\left(\max\left(1 - \max\left(\frac{t-\kappa}{\epsilon}, 0\right), 0\right), 1\right)$
  if $t < T_{\text{imit}}$ **then**
    Compute imitation PPO objective: $L_{\text{imit}}(\psi)$
    Update ψ by $\nabla_\psi \left(wL_{\text{imit}}(\psi) + (1-w)J(\psi)\right)$
  else
    Compute trajectory PPO objective: $L_{\text{traj}}(\psi)$
    Update ψ by $\nabla_\psi \left(wL_{\text{traj}}(\psi) + (1-w)J(\psi)\right)$
  end if
end for

to training instability and convergence difficulties. To address this, we propose to learn action generation from motion imitation priors. As illustrated in Stage III of Figure 3, we first process the original reference data using our pre-trained imitation policies π^{imit} to produce physically consistent state-action pairs—*physical states* s^e and corresponding *actions* a^e. The physical states s^e enhance the raw motion capture data by ensuring full physical plausibility, while the actions a^e provide direct supervision for the generative model. This strong pairing effectively reduces the under-constrained nature of trajectory control, transforming an open-ended problem into a well-defined one with a unique solution per demonstration.

Progressive Skill Acquisition The complete algorithm flow is shown in Algorithm 1. To ensure stable training and enable a gradual transition from pure imitation of π^{imit} to autonomous action generation, we introduce three key hyper-parameters to control the learning schedule. The first two, ϵ and κ, are adopted from the DAgger algorithm. When $t <= \kappa$ only actions a^e are executed in simulation, since the untrained policy ψ would produce actions leading to early termination. This allows the policy to learn from meaningful states from the very beginning. In the subsequent $\kappa < t < \epsilon + \kappa$ episodes, we linearly anneal the probability of sampling actions from the π^{imit}, while also progressively decreasing the weight of the action imitation loss $J(\psi)$. After $\kappa+\epsilon$ episodes, the simulation acts a only and the $J(\psi)$ is removed. The third hyperparameter, Timit, governs the reward composition. As outlined in Algorithm 1, for the first T_{imit} epochs, we employ the full imitation reward used in Stages I and II to ground the policy. Beyond T_{imit}, we switch to a sparse reward that depends only on the object's trajectory tracking error. This shift encourages the policy to explore its own action sequences to achieve the goal, rather than slavishly mimicking the teacher's motions, thereby significantly improving generalization to novel trajectories and interaction scenarios.

IV. Experiments

A. Datasets

Training on OMOMO The OMOMO dataset serves as our primary source of single-human motion capture data, comprising approximately 10 hours of human-object interactions. To accommodate morphological variations, we partitioned the raw data into 17 distinct subsets based on the body shapes of the actors, yielding a total of 6,435 motion sequences. We initially trained separate imitation policies on these subsets and subsequently extended the training to dual-agent scenarios through data augmentation, resulting in a collection of expert policies denoted as $\{\pi_i^s\}_{i=0}^{16}$. A critical component of our pipeline is the refinement of this data using the trained policies. Since the original OMOMO dataset lacks dexterous hand articulation and contains physically inconsistent artifacts, we employed a physics-based re-tracking procedure to validate the motions. By retaining only the sequences where $\{\pi_i^s\}_{i=0}^{16}$ successfully completed the imitation, we curated a high-quality dataset of 2,888 physically plausible motions. Finally, this refined dataset was utilized to distill $\{\pi_i^s\}_{i=0}^{16}$ into a unified Base Model, providing a robust initialization for subsequent multi-agent learning.

Training on CORE4D In the second stage, we leverage the CORE4D dataset, a large-scale benchmark for human-object-human interaction (HOHI) that nominally comprises 11,000 collaboration sequences spanning 3,000 real and virtual object shapes. However, upon closer inspection, we observed that the synthetically generated portion of the dataset exhibits low physical fidelity and contains significant artifacts, rendering it unsuitable for training robust physics-based control policies. Consequently, we excluded these synthetic samples entirely and restricted our training set exclusively to the real-world motion capture segment. Following a data processing, we distilled the dataset down to 961 sequences, which serve as the initial training data for our multi-agent cooperative policies.

We initially organized the real-world motion capture data into 7 distinct subsets based on object categories to facilitate specialized training. However, as noted above, our HOHI imitation policies incorporate only 3 policies $\{\pi_i^m\}_{i=0}^{2}$. This reduction is necessitated by the critically low quality of the raw CORE4D data, which frequently exhibit severe artifacts such as abrupt trajectory discontinuities, skeletal distortions, and physically implausible contacts. These fundamental defects rendered training non-convergent for several categories, including boards and sticks, and could not be rectified through retargeting or post-optimization. Consequently, we were forced to discard the compromised subsets entirely.

B. Final Dataset

Ultimately, these sequences from both the HOI and HOHI datasets undergo refinement via our imitation policies to generate the physically consistent state-action pairs for training the final Trajectory-Conditioned Policy. By aggregating these resources, we construct a comprehensive training corpus consisting of 20 distinct subsets, comprising 17 derived from single-human data and 3 from multi-human interactions, yielding a total of 2,960 physically validated motion sequences. This diverse dataset encompasses 9 major object categories and 25 unique object geometries, the full variety of which is visualized in Figure 4. To promote reproducibility and facilitate further research in the community, we will release both the curated training dataset and our complete source code upon the acceptance of this paper.

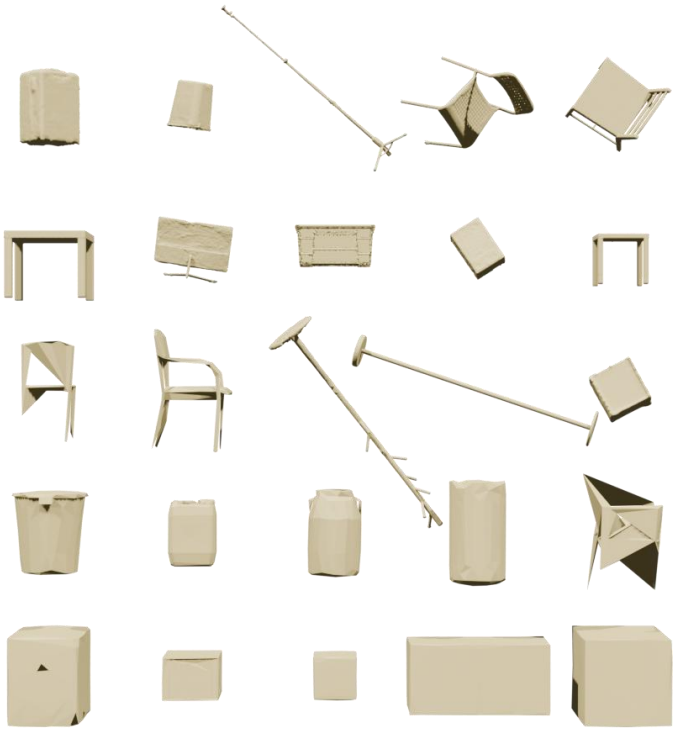


Fig. 4. **Overview of 25 Objects.** Our model can ultimately cover these 25 objects.

C. Implementation Details

Our experiments are conducted on the OMOMO and CORE4D datasets. OMOMO provides single-human HOI sequences, while CORE4D contains multi-human HOHI data. After automatic filtering to remove low-quality samples, we obtain 2,960 motion sequences covering 9 object categories and 25 distinct objects. Based on these data, we train 20 motion imitation models in total, including 17 on OMOMO and 3 on CORE4D. Each imitation model is trained on a single NVIDIA RTX 3090 GPU, while the final distillation stage for learning the generalizable trajectory-conditioned policy is performed on an NVIDIA A800 GPU. Our evaluation includes comparisons with existing methods and ablation studies that analyze the impact of individual components.

Comparison Since no prior methods are explicitly designed for HOHI imitation or trajectory-controlled cooperative humanoid manipulation, we adopt representative alternatives for comparison. For motion imitation, we use InterMimic [66], a state-of-the-art HOI imitation method, applied independently to each agent to generate actions. For trajectory-conditioned control, we compare against CooHOI [62], which supports goal-conditioned object manipulation but does not track trajectories. Both methods are evaluated on their ability to cooperatively move an object to a target position, enabling a fair comparison at the object-control level. Qualitative comparison is shown in Figure 5.

Ablation Studies We perform ablation studies to examine key design choices in our framework. These include comparing centralized and decentralized policy structures for dual-agent manipulation, evaluating different architectural variants of the VAE-based trajectory-conditioned policy, and assessing the transferability of knowledge learned from single-human data to multi-human cooperation. We also ablate the proposed data refinement strategies, including interaction-preserving retargeting and automatic filtering, to quantify their effects on training stability and performance.

Metrics We design task-specific evaluation metrics for imitation and control, respectively. For motion imitation, we report five metrics. The inference success rate (Succ. Rate) measures the probability that a policy can successfully complete an entire motion sequence when evaluated over N demonstrations, where higher values indicate better robustness. Episode length (Ep. Len.) reports the average duration for which imitation remains successful before termination, with longer episode lengths reflecting more stable tracking. The accumulated reward (Reward) evaluates imitation quality, where higher rewards correspond to closer adherence to the reference motion.

To directly quantify geometric accuracy, we further compute the mean human joint error E_h over only successful episodes, weighted by the inference success rate, such that lower values indicate more accurate human motion reproduction. Similarly, the object error E_o is defined as the weighted mean trajectory error between the manipulated object and the reference object motion, where smaller values correspond to more precise object-level imitation. For trajectory-conditioned control, we adopt two complementary metrics. The success rate (Succ. Rate) is defined by a minimum distance threshold, where a trial is considered successful if the final object position lies within this threshold of the target. In addition, we report the average distance between the final object position and the target position across all trials, with smaller distances indicating more accurate goal-directed manipulation.

D. Imitation and Control

Comparison of Imitation. Table I (first two rows) compares InterMimic and our method under the imitation setting. Both methods are trained on the same OMOMO dataset using identical data splits, where we train 17 imitation policies for our approach and 17 corresponding models for InterMimic. During evaluation, dual-agent scenarios are constructed using

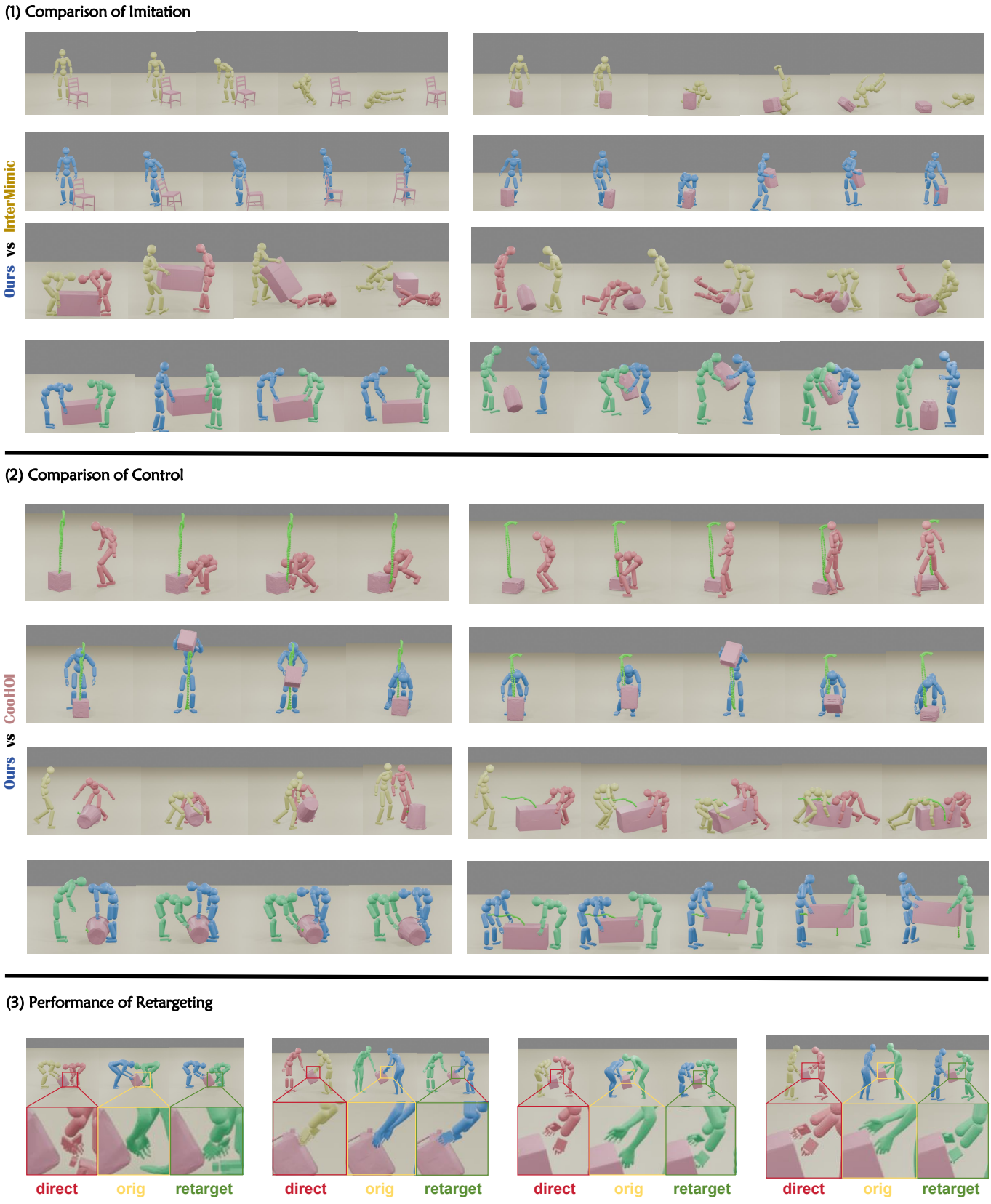


Fig. 5. **Qualitative Results.** In the comparison between Ours and existing comparable baselines, the blue and green agents are the test results from Ours. In *Comparison of Control*, the green ball represents the trajectory control signal. In *Performance of Retargeting*, "direct" indicates that MoCap data is directly transferred to the agent, "orig" represents the raw MoCap data, and "retarget" represents the effect of using our Interaction-preserving retargeting.

the single-human data augmentation strategy described in Sec. III-B2. As shown in the table, although InterMimic achieves strong performance in single-human HOI imitation, its performance degrades significantly when an additional agent is introduced as a source of physical interference, indicating limited adaptability to multi-human settings without targeted training. In contrast, our method substantially improves the imitation success rate and consistently outperforms InterMimic across all metrics. Notably, these gains are achieved without modifying the model architecture or tuning hyper-parameters, demonstrating that MAPPO-based joint training is critical for enabling robust multi-agent imitation and highlighting the superiority of our approach in cooperative scenarios.

TABLE I
COMPARISON OF IMITATION. THIS TABLE IS DIVIDED INTO TWO GROUPS BASED ON DIFFERENT BACKGROUND COLORS FOR COMPARISON WITHIN EACH GROUP. *Ours* ARE $\{\pi_i^s\}_{i=0}^N$, *Centralized* MEANS ONE POLICY OUTPUTS TWO AGENTS' ACTIONS, AND *Decentralized* MEANS TWO SHARE-WEIGHT POLICIES OUTPUT TWO AGENTS' ACTIONS SEPARATELY.

Method	Succ. Rate ↑	Ep. Len. ↑	Reward ↑	E_h ↓	E_o ↓
InterMimic	7.26%	51.9151	20.1757	1.7352	1.4329
Ours	**45.00%**	**78.7416**	**25.5555**	**0.2376**	**0.2387**
Centralized	12.61%	77.0029	17.9975	0.8170	0.8184
Decentralized (Ours)	**19.92%**	**84.0158**	**19.0685**	**0.5404**	**0.5568**

Comparison of Control. The closest existing baseline to our task is CooHOI, as both methods are able to cooperatively transport an object to a specified target position. As reported in Table II, except for box-shaped objects, our method consistently outperforms CooHOI across nearly all object categories. CooHOI is target-conditioned, so we provided it with the final 2D positions of test trajectories. And for fairness, only the distance on 2D plane is calculated when calculating the metrics. Moreover, CooHOI is an AMP-like method, meaning no training set bias existed. CooHOI's drastic performance degradation stems from its inherent design: it abstracts objects into simple 8-keypoint bounding boxes and lacks dexterous hand modeling. Its performance relies on an over-engineered reward function specifically tailored for picking up the box, which inevitably limits its generalization to other objects. Consequently, it suffers from severe generalization failures when faced with non-box-like (even non-square boxes), complex geometries.

Moreover, CooHOI does not model dexterous hands, which severely limits its ability to manipulate objects that require stable grasping, such as clothes stands, monitors, and tripods, where its success rate is nearly 0.00%. In contrast, our methods demonstrates substantially stronger generalization. The VAE-based policy enables fine-grained trajectory-conditioned control rather than pure goal-conditioned behavior, while the interaction graph provides a more expressive and geometry-aware object representation. Although incorporating dexterous hands significantly increases learning difficulty, our framework successfully overcomes this challenge, allowing a single model to handle a broader range of objects that require precise and coordinated manipulation.

E. Ablation and Analysis

Centralized or Decentralized After achieving strong single-human HOI imitation performance, we explore architectural designs that can better facilitate dual-agent cooperation. A straightforward approach is a centralized policy, where a single model observes both agents and directly outputs actions for both of them. In principle, such a design allows more explicit information sharing and could simplify coordination between agents. In practice, however, doubling the action space significantly increases the learning difficulty, leading to slower convergence and degraded performance. As a result in Table I, we observe inferior results compared to alternative designs. We therefore adopt a decentralized policy structure, where each agent is controlled by a shared-weight policy that outputs its own actions. This design not only yields better empirical performance but also offers superior scalability, as it naturally extends to scenarios involving varying numbers of agents.

Solo-to-Cooperative Skill Transition As a core contribution of our work, we need to explicitly validate its effectiveness in facilitating multi-agent skill learning. As shown by the comparison between *w/o init* and *Ours* in Table III, training the multi-human imitation policies $\{\pi_i^m\}_{i=0}^M$ without initializing from the Base Model distilled from single-human policies $\{\pi_i^s\}_{i=0}^N$ leads to a sharp performance degradation. Without a strong skill prior acquired from abundant single-human data, learning complex cooperative behaviors from scratch becomes extremely difficult, and the training process often fails to converge. These results demonstrate that transiting single-agent skills is crucial for stable and effective learning in multi-agent cooperative manipulation tasks.

Data Refinement Another major contribution of our work is a comprehensive data refinement pipeline for HOHI motion data, whose effectiveness is reflected in the model performance. As shown by the comparison between *w/o retarget* and *Ours* in Table III, interaction-preserving retargeting significantly improves learning by correcting motion artifacts caused by shape discrepancies between human performers and humanoid agents. However, retargeting alone is insufficient. The poor performance of the *w/o filter* variant highlights that MoCap data contains a substantial amount of erroneous and low-quality sequences that cannot be remedied through retargeting alone. Our automatic filtering strategy effectively removes such problematic samples, allowing the model to focus on reliable demonstrations. Together, retargeting and filtering address the fundamental data quality issues in HOHI learning, enabling stable training and substantially improved performance.

F. More Experiment Results

1) Motion Imitation: **Imitation on OMOMO** A comparative analysis of Table IV and Table V highlights a critical finding: proficiency in single-agent HOI imitation is insufficient for direct application to multi-agent collaborative tasks. As evidenced in Table IV, although the baseline method (InterMimic) achieves strong results on standard single-human OMOMO sequences, its performance degrades precipitously

TABLE II
COMPARISON OF CONTROL. OURS-1 ONLY USES AGENT STATES TO CONCATENATE WITH LATENT VECTORS z, WHILE OURS-2 ADDITIONALLY USE INTERACTION GRAPHS. NOTE THAT EACH MAJOR CATEGORY OF OBJECTS ALSO HAS VARIOUS SUBCATEGORIES. SEE THE SUPPLEMENTARY MATERIALS FOR DETAILED EXPERIMENTAL RESULTS.

Method		box	bucket	chair	clothes stand	table	floor lamp	monitor	suitcase	tripod	all
CooHOI	Dist. ↓	1.1801	1.382	1.4404	1.3825	1.2835	1.3837	1.1688	1.3088	1.3212	1.3074
	Succ. Rate ↑	14.14%	2.95%	4.26%	0.00%	10.66%	6.56%	1.64%	3.28%	0.00%	7.28%
Ours-1	Dist. ↓	**0.9863**	0.4077	0.4279	0.7205	0.7634	0.5193	**0.5106**	1.1204	0.9261	0.8192
	Succ. Rate ↑	**14.85%**	**52.98%**	28.83%	12.09%	**16.58%**	30.59%	18.75%	**6.84%**	10.93%	18.41%
Ours-2	Dist. ↓	0.9985	**0.3871**	**0.4099**	**0.6524**	**0.7488**	**0.4271**	0.6009	**1.1005**	**0.8179**	**0.7939**
	Succ. Rate ↑	11.88%	51.79%	**39.33%**	**16.39%**	12.94%	**37.23%**	**25.00%**	5.65%	**15.48%**	**19.16%**

TABLE III
ABLATION STUDIES. WE ANALYZE THE CONTRIBUTION OF EACH COMPONENT: INITIALIZATION WITH BASE MODEL (INIT), INTERACTION-PRESERVING RETARGETING (RETARGET), AND TRAIN-TO-FILTER STRATEGY (FILTER). ✓INDICATES THE COMPONENT IS USED.

Components			Metrics				
Init	Retarget	Filter	Succ. Rate ↑	Ep. Len. ↑	Reward ↑	E_h ↓	E_o ↓
–	✓	✓	3.33%	79.3216	16.3713	4.8528	4.4174
✓	–	✓	2.42%	98.4375	20.2552	6.6776	6.0785
✓	✓	–	1.43%	69.7496	13.7437	12.6062	7.9756
✓	✓	✓	**21.82%**	**108.4463**	**20.7419**	**0.7406**	**0.6741**

TABLE IV
EVALUATION RESULTS OF INTERMIMIC ON OMOMO. NOTE THAT THESE ARE EVALUATION RESULTS FOR OMOMO IN A TWO-AGENT SCENARIO AFTER IT HAS BEEN EXPANDED TO TWO-AGENT MODE.

	Percent.	Succ. Rate ↑	Ep. Len. ↑	Reward ↑	E_h ↓	E_o ↓
sub1	14.03%	1.40%	52.5902	9.5020	9.2286	7.83576
sub2	1.31%	10.34%	65.5994	22.4952	0.9107	0.9204
sub3	3.03%	0.50%	41.7012	7.2260	32.5400	22.1000
sub4	3.80%	0.40%	51.5622	9.8790	21.3250	17.2000
sub5	4.21%	8.24%	81.7084	18.1700	1.4256	0.4110
sub6	5.93%	3.31%	56.7799	11.7300	3.0333	1.3909
sub7	5.89%	11.03%	66.1516	15.7531	1.2155	1.0591
sub8	5.16%	9.06%	56.6106	14.5854	1.4330	1.5615
sub9	6.79%	0.00%	33.8088	3.8940	∞	∞
sub10	9.41%	15.06%	53.3284	7.4360	0.9364	0.8907
sub11	2.03%	37.78%	68.4531	21.4116	0.2622	0.0593
sub12	4.48%	2.02%	46.4172	10.3300	6.8750	5.7850
sub13	6.33%	8.81%	42.7553	6.8939	1.4352	1.4159
sub14	7.02%	4.73%	52.1470	8.9440	2.6340	2.3915
sub15	8.87%	16.50%	46.0051	7.6330	0.7909	0.6952
sub16	5.39%	5.04%	53.3982	12.9800	3.0120	2.2680
sub17	6.29%	5.28%	43.2625	6.3382	2.2377	3.0623
all	100%	7.26%	51.9151	20.1757	1.7352	1.4329

TABLE V
EVALUATION RESULTS OF OURS ON OMOMO. THE SAME DATA AS TABLE IV IS USED, AND THE RESULTS ARE COMPARED WITH THOSE IN TABLE 4. "OURS" REFERS TO IMITATION POLICIES $\{\pi_i^s\}_{i=0}^N$.

	Percent.	Succ. Rate ↑	Ep. Len. ↑	Reward ↑	E_h ↓	E_o ↓
sub1	14.03%	29.57%	82.4458	24.7642	0.3771	0.3923
sub2	1.31%	79.31%	116.6124	41.7244	0.1111	0.1265
sub3	3.03%	59.20%	84.7115	18.8653	0.2049	0.1995
sub4	3.80%	44.44%	83.0679	23.3746	0.2664	0.2459
sub5	4.21%	34.41%	93.8277	30.4317	0.3173	0.2647
sub6	5.93%	11.45%	76.9975	20.1347	0.8044	0.7205
sub7	5.89%	50.77%	93.6577	32.2836	0.2017	0.224
sub8	5.16%	60.82%	103.3444	34.2238	0.1771	0.1771
sub9	6.79%	40.00%	48.7679	9.0562	0.2483	0.2633
sub10	9.41%	60.58%	70.589	24.7642	0.2212	0.2136
sub11	2.03%	80.00%	95.9667	41.7244	0.1070	0.0616
sub12	4.48%	56.57%	87.9563	18.8653	0.1792	0.1944
sub13	6.33%	43.81%	68.8254	23.3746	0.2627	0.3186
sub14	7.02%	57.85%	95.0348	30.4317	0.1758	0.1801
sub15	8.87%	44.05%	59.5867	20.1347	0.2556	0.2427
sub16	5.39%	40.06%	79.4779	32.2836	0.2569	0.2683
sub17	6.29%	41.01%	65.4814	34.2238	0.2772	0.3299
all	100%	45.00%	78.7416	25.5555	0.2376	0.2387

TABLE VI
ABLATION STUDY OF TRAIN-TO-FILTER ON CORE4D. FROM TOP TO BOTTOM, THESE ARE: NO FILTER, FIRST FILTER AND SECOND FILTER (I.E., $\{\pi_i^m\}_{i=0}^M$).

	Succ. Rate ↑	Ep. Len. ↑	Reward ↑	E_h ↓	E_o ↓
box	4/272=1.47%	62.7061	12.8191	9.6054	2.7551
bucket	5/250=2.00%	78.3046	14.3882	8.9950	6.8500
chair	2/248=0.81%	68.8705	14.1121	19.5432	14.8395
all	11/770=1.43%	69.7496	13.7437	12.6062	7.9756
box	9/165=5.45%	109.9828	**22.3496**	2.7302	1.9137
bucket	28/250=11.20%	118.1502	20.5744	1.4232	1.2500
chair	6/162=3.70%	82.0555	16.9745	4.3027	4.7432
all	43/577=7.45%	102.7352	19.9511	2.0912	1.8778
box	29/102=28.43%	**110.9783**	21.8224	**0.5388**	**0.4618**
bucket	28/84=33.33%	**141.04**	**23.8459**	**0.4875**	**0.4347**
chair	15/144=10.42%	**87.6451**	**18.1666**	**1.6218**	**1.5854**
all	72/330=21.82%	**108.4463**	**20.7419**	**0.7406**	**0.6741**

when applied to our augmented dual-agent scenarios. This decline is primarily attributed to the inability of the policies to accommodate the dynamic perturbations introduced by a second interacting agent. In contrast, Table V demonstrates the superior robustness of our approach. By leveraging MAPPO-based joint training, our imitation policies effectively learn to mitigate external disturbances, resulting in substantially improved resistance to physical interference. Consequently, our method achieves comprehensive improvements across nearly all evaluated subsets, reaffirming the effectiveness of our framework in bridging the gap between isolated skill acquisition and resilient collaborative execution.

Imitation on CORE4D As demonstrated in Table VI, we validated the effectiveness of our data refinement pipeline by applying the Train-to-Filter strategy over two iterations. A consistent trend is observed across these iterations, where all performance metrics improve following each filtration pass. Crucially, the Succ. Rate metric reveals a dual improvement: not only does the percentage of successful episodes increase, but the number of successfully learned motions also rises. This phenomenon indicates that low-quality data acts as a source of interference during training. By purging these physically inconsistent samples, the model gains the capacity to master complex motions that were previously unlearnable, thereby underscoring the value of our filtering approach. Furthermore,

TABLE VII
ABLATION STUDY ON CORE4D. FROM TOP TO BOTTOM, THESE ARE: NO INITILIZATION WITH BASE MODEL, NO INTERACTION-PRESERVING RETARGETING AND OURS (I.E., $\{\pi_i^m\}_{i=0}^M$).

	Succ. Rate ↑	Ep. Len. ↑	Reward ↑	E_h ↓	E_o ↓
box	2/102=1.96%	84.2532	17.9991	7.8163	6.6989
bucket	7/84=8.33%	91.0018	18.2712	1.9507	1.7394
chair	2/144=1.39%	69.0171	14.1104	12.1582	11.8848
all	11/330=3.33%	79.3216	16.3713	4.8528	4.4174
box	2/102=1.96%	109.8439	**23.9613**	7.9336	2.5408
bucket	6/84=7.14%	105.8384	20.0035	2.6526	1.7450
chair	0/144=0.00%	86.0424	17.7771	∞	∞
all	8/330=2.42%	98.4375	20.2552	6.6776	6.0785
box	29/102=28.43%	**110.9783**	21.8224	**0.5388**	**0.4618**
bucket	28/84=33.33%	**141.04**	**23.8459**	**0.4875**	**0.4347**
chair	15/144=10.42%	**87.6451**	**18.1666**	**1.6218**	**1.5854**
all	72/330=21.82%	**108.4463**	**20.7419**	**0.7406**	**0.6741**

Table VII provides a more granular breakdown of our ablation studies, where the performance trends across subsets align with the aggregated results, confirming the consistency and robustness of our findings.

2) Motion Control: Complementing the aggregated results presented in the main paper, Table VIII provides a granular performance breakdown for each of the 25 individual object geometries. Consistent with the trends observed in Table II, our framework demonstrates superior control capabilities and robust geometric generalization, significantly outperforming the CooHOI baseline, even on the box-transport tasks for which CooHOI was specifically optimized. Furthermore, the detailed ablation results highlight the architectural advantage of explicitly encoding spatial constraints. specifically, the Ours-2 variant, which augments the latent vector concatenation with the Interaction Graph, yields higher performance metrics than the standard configuration. This evidence confirms that providing the decoder with precise spatial relationship cues is essential for enhancing the model's ability to execute stable and accurate trajectory-conditioned manipulation across diverse object shapes.

V. DISCUSSION

Despite its effectiveness, our work represents only an initial step toward cooperative humanoid manipulation and leaves substantial room for improvement. The primary limitation lies in the scarcity of high-quality HOHI data, which prevents our model from scaling to achieve performance leaps. As a result, learning more diverse and complex interaction behaviors beyond object transportation remains challenging. We emphasize that SynAgent is a pioneering "0-to-1" exploration of physics-based, dual-agent cooperative manipulation. Our approach achieves a 2-to-7-fold increase in imitation success and trajectory completion over existing methods. The primary bottleneck remains the extreme scarcity and poor quality of existing HOHI datasets, which makes even stable locomotion difficult. Our framework is specifically designed to scale efficiently as larger, higher-quality datasets become available in the future.

TABLE VIII
COMPARISON OF CONTROL ON OMOMO AND CORE4D. A MORE DETAILED VERSION OF TABLE II, CATEGORIZES OBJECTS IN A MORE SPECIFIC WAY.

Object	CooHOI		Ours-1		Ours-2	
	Dist.	Succ. Rate	Dist.	Succ. Rate	Dist.	Succ. Rate
box001	1.3959	22.95%	**0.355**	**60.00%**	0.3753	40.00%
box021	0.9676	16.39%	**0.2919**	**77.78%**	0.411	55.56%
box023	1.1619	1.64%	**0.1635**	**100.00%**	0.2019	**100.00%**
box024	1.1487	26.23%	0.387	57.14%	**0.2965**	**71.43%**
box025	0.968	37.70%	0.6197	**42.86%**	**0.3851**	**42.86%**
bucket007	1.2274	6.56%	0.3845	71.43%	**0.2817**	**85.71%**
bucket008	1.2214	3.28%	1.2021	0.00%	**0.2373**	**100.00%**
bucket009	**1.2589**	**1.64%**	1.9047	0.00%	1.3359	0.00%
bucket010	1.9481	1.64%	0.8664	0.00%	**0.5097**	**40.00%**
chair006	1.0154	6.56%	0.8462	10.53%	**0.6434**	**25.00%**
chair020	1.2553	1.64%	0.5582	0.00%	**0.2614**	**66.67%**
chair022	1.1807	3.28%	**0.1646**	66.67%	0.2591	**100.00%**
clothesstand	1.3825	0.00%	0.7205	12.09%	**0.6524**	**16.39%**
floorlamp	1.3837	6.56%	0.5193	30.59%	**0.4271**	**37.23%**
largebox	1.1982	4.92%	**1.0408**	**17.05%**	1.0594	11.64%
largetable	1.3339	18.03%	**0.7164**	**18.58%**	0.7253	13.60%
monitor	1.1688	1.64%	**0.5106**	18.75%	0.6009	**25.00%**
plasticbox	1.3409	1.64%	0.8744	5.19%	**0.8669**	**7.79%**
smallbox	1.2598	1.64%	**1.0218**	**12.04%**	1.0379	10.16%
smalltable	1.2330	3.28%	0.8275	**13.86%**	**0.7812**	12.05%
suitcase	1.3088	3.28%	1.1204	**6.84%**	**1.1005**	5.65%
trashcan	1.2545	1.64%	**0.358**	**55.63%**	0.3783	50.33%
tripod	1.3212	0.00%	0.9261	10.93%	**0.8179**	**15.48%**
whitechair	2.5922	0.00%	0.3607	43.51%	**0.3356**	**54.20%**
woodchair	1.1585	9.84%	0.4354	19.41%	0.4447	**28.07%**
all	1.3074	7.28%	0.8192	18.41%	**0.7939**	**19.16%**

VI. CONCLUSION

We presented SynAgent, a framework for generalizable cooperative humanoid manipulation that addresses the core challenges of data scarcity, coordination complexity, and trajectory control in multi-agent settings. By introducing an interaction-preserving retargeting strategy, we substantially improved the quality and usability of scarce HOHI data. Our single-agent pretraining paradigm demonstrates that cooperative manipulation can be effectively bootstrapped from abundant single-human experience, enabling robust decentralized coordination without specialized multi-agent supervision. Furthermore, by distilling motion imitation priors into a trajectory-conditioned generative policy, SynAgent achieves precise and stable object-level control while maintaining strong generalization across object geometries and agent counts. Together, these components establish a scalable pathway from individual manipulation skills to coordinated multi-humanoid behaviors, offering a practical foundation for future research in embodied intelligence, physics-based animation, and multi-agent robotic interaction.

REFERENCES

[1] Z. Luo, J. Cao, K. Kitani, W. Xu *et al.*, "Perpetual humanoid control for real-time simulated avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 895–10 904.

[2] Z. Luo, J. Cao, J. Merel, A. Winkler, J. Huang, K. Kitani, and W. Xu, "Universal humanoid motion representations for physics-based control," *arXiv preprint arXiv:2310.04582*, 2023.

[3] Z. Luo, J. Cao, S. Christen, A. Winkler, K. Kitani, and W. Xu, "Omnigrasp: Grasping diverse objects with simulated humanoids," *Advances in Neural Information Processing Systems*, vol. 37, pp. 2161–2184, 2024.

[4] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, "Maskedmimic: Unified physics-based character control through masked motion inpainting," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–21, 2024.

[5] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[6] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "Intergen: Diffusion-based multi-human motion generation under complex interactions," *International Journal of Computer Vision*, vol. 132, no. 9, pp. 3463–3483, 2024.

[7] L. Xu, X. Lv, Y. Yan, X. Jin, S. Wu, C. Xu, Y. Liu, Y. Zhou, F. Rao, X. Sheng *et al.*, "Inter-x: Towards versatile human-human interaction analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 22 260–22 271.

[8] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *European conference on computer vision*. Springer, 2020, pp. 581–600.

[9] N. Jiang, Z. He, Z. Wang, H. Li, Y. Chen, S. Huang, and Y. Zhu, "Autonomous character-scene interaction synthesis from text instruction," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[10] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang, "Scaling up dynamic human-scene interaction modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1737–1747.

[11] Y. Liu, C. Zhang, R. Xing, B. Tang, B. Yang, and L. Yi, "Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1769–1782.

[12] E. S. Ho, T. Komura, and C.-L. Tai, "Spatial relationship preserving character motion adaptation," in *ACM SIGGRAPH 2010 papers*, 2010, pp. 1–8.

[13] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in neural information processing systems*, vol. 35, pp. 24 611–24 624, 2022.

[14] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[15] X. B. Peng, "Mimickit: A reinforcement learning framework for motion imitation and control," *arXiv preprint arXiv:2510.13794*, 2025.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[17] W. Yao, Y. Sun, C. Liu, H. Zhang, and J. Tang, "Physiinter: Integrating physical mapping for high-fidelity human interaction generation," *arXiv preprint arXiv:2506.07456*, 2025.

[18] Z. Luo, J. Wang, K. Liu, H. Zhang, C. Tessler, J. Wang, Y. Yuan, J. Cao, Z. Lin, F. Wang *et al.*, "Smplolympics: Sports environments for physically simulated humanoids," *arXiv preprint arXiv:2407.00187*, 2024.

[19] Y. Wang, Q. Zhao, R. Yu, A. Zeng, J. Lin, Z. Luo, H. W. Tsui, J. Yu, X. Li, Q. Chen *et al.*, "Skillmimic: Learning reusable basketball skills from demonstrations," *arXiv e-prints*, pp. arXiv–2408, 2024.

[20] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner *et al.*, "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi8022, 2024.

[21] J. Bae, J. Won, D. Lim, C.-H. Min, and Y. M. Kim, "Pmp: Learning to physically interact with environments using part-wise motion priors," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–10.

[22] Y. Zhang, D. Gopinath, Y. Ye, J. Hodgins, G. Turk, and J. Won, "Simulation and retargeting of complex multi-character interactions," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[23] Y. Kim, H. Park, S. Bang, and S.-H. Lee, "Retargeting human-object interaction to virtual avatars," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 11, pp. 2405–2412, 2016.

[24] J.-Q. Zhang, M. Wang, F.-C. Zhang, and F.-L. Zhang, "Skinned motion retargeting with preservation of body part relationships," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[25] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8944–8951.

[26] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi, "Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction," *arXiv preprint arXiv:2509.26633*, 2025.

[27] C. Pan, C. Wang, H. Qi, Z. Liu, H. Bharadhwaj, A. Sharma, T. Wu, G. Shi, J. Malik, and F. Hogan, "Spider: Scalable physics-informed dexterous retargeting," *arXiv preprint arXiv:2511.09484*, 2025.

[28] S. Hou, H. Tao, H. Bao, and W. Xu, "A two-part transformer network for controllable motion synthesis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 8, pp. 5047–5062, 2023.

[29] X. Gao, Y. Yang, Z. Xie, S. Du, Z. Sun, and Y. Wu, "Guess: Gradually enriching synthesis for text-driven human motion generation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 12, pp. 7518–7530, 2024.

[30] J. Cha, J. Kim, J. S. Yoon, and S. Baek, "Text2hoi: Text-guided 3d motion generation for hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1577–1585.

[31] S. Christen, S. Hampali, F. Sener, E. Remelli, T. Hodan, E. Sauser, S. Ma, and B. Tekin, "Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

[32] Q. Li, J. Wang, C. C. Loy, and B. Dai, "Task-oriented human-object interactions generation with implicit neural representations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3035–3044.

[33] J. Ma, J. Xu, X. Chen, and H. Wang, "Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos," *arXiv preprint arXiv:2405.04370*, 2024.

[34] J. Tian, R. Ji, L. Yang, S. Ni, Y. Ma, L. Xu, J. Yu, Y. Shi, and J. Wang, "Gaze-guided hand-object interaction synthesis: Dataset and method," *arXiv preprint arXiv:2403.16169*, 2024.

[35] H. Zhang, S. Christen, Z. Fan, L. Zheng, J. Hwangbo, J. Song, and O. Hilliges, "Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 235–246.

[36] J. Zhang, Y. Zhang, L. An, M. Li, H. Zhang, Z. Hu, and Y. Liu, "Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[37] Z. Hou, B. Yu, and D. Tao, "Compositional 3d human-object neural animation," *arXiv preprint arXiv:2304.14070*, 2023.

[38] T. Kim, S. Saito, and H. Joo, "Ncho: Unsupervised learning for neural 3d composition of humans and objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 817–14 828.

[39] I. A. Petrov, R. Marin, J. Chibane, and G. Pons-Moll, "Object pop-up: Can we infer 3d objects and their poses from human interactions alone?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4726–4736.

[40] X. Xie, J. E. Lenssen, and G. Pons-Moll, "Intertrack: Tracking human object interaction without object templates," in *2025 International Conference on 3D Vision (3DV)*. IEEE, 2025, pp. 1427–1439.

[41] C. Yang, C. Kang, K. Kong, H. Oh, and S.-J. Kang, "Person in place: Generating associative skeleton-guidance maps for human-object interaction image editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8164–8175.

[42] Y. Yang, W. Zhai, H. Luo, Y. Cao, and Z.-J. Zha, "Lemon: Learning 3d human-object interaction relation from 2d images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 284–16 295.

[43] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "Imos: Intent-driven full-body motion synthesis for human-object interactions," in *Computer Graphics Forum*, vol. 42, no. 2. Wiley Online Library, 2023, pp. 1–12.

[44] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The kit bimanual manipulation dataset," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 499–506.

[45] N. Kulkarni, D. Rempe, K. Genova, A. Kundu, J. Johnson, D. Fouhey, and L. Guibas, "Nifty: Neural object interaction fields for guided human motion synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 947–957.

[46] J. Lee and H. Joo, "Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9663–9674.

[47] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–11, 2023.
[48] H. Razali and Y. Demiris, "Action-conditioned generation of bimanual object manipulation sequences," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2146–2154.
[49] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "Goal: Generating 4d whole-body motion for hand-object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 263–13 273.
[50] W. Wan, L. Yang, L. Liu, Z. Zhang, R. Jia, Y.-K. Choi, J. Pan, C. Theobalt, T. Komura, and W. Wang, "Learn to predict how humans manipulate large-sized objects from interactive motions," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4702–4709, 2022.
[51] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang, "Saga: Stochastic whole-body grasping with contact," in *European Conference on Computer Vision*. Springer, 2022, pp. 257–274.
[52] X. Xu, H. Joo, G. Mori, and M. Savva, "D3d-hoi: Dynamic 3d human-object interactions from videos," *arXiv preprint arXiv:2108.08420*, 2021.
[53] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll, "Couch: Towards controllable human-chair interactions," in *European Conference on Computer Vision*. Springer, 2022, pp. 518–535.
[54] K. Zhao, Y. Zhang, S. Wang, T. Beeler, and S. Tang, "Synthesizing diverse human motions in 3d indoor scenes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 14 738–14 749.
[55] W. Yao, Y. Sun, H. Zhang, Y. Liu, and J. Tang, "Hosig: Full-body human-object-scene interaction generation with hierarchical scene perception," *arXiv preprint arXiv:2506.01579*, 2025.
[56] Z. Hu, J. Xu, S. Schmitt, and A. Bulling, "Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses," *IEEE Transactions on Visualization and Computer Graphics*, 2024.
[57] I. Loi, E. I. Zacharaki, and K. Moustakas, "Machine learning approaches for 3d motion synthesis and musculoskeletal dynamics estimation: a survey," *IEEE transactions on visualization and computer graphics*, vol. 30, no. 8, pp. 5810–5829, 2023.
[58] J. Vaillant, K. Bouyarmane, and A. Kheddar, "Multi-character physical and behavioral interactions controller," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 6, pp. 1650–1662, 2016.
[59] K. Hu, B. Haworth, G. Berseth, V. Pavlovic, P. Faloutsos, and M. Kapadia, "Heterogeneous crowd simulation using parametric reinforcement learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 4, pp. 2036–2052, 2021.
[60] Z. Wang, B. Benes, A. H. Qureshi, and C. Mousas, "Evolution-based shape and behavior co-design of virtual agents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 12, pp. 7579–7591, 2024.
[61] Y.-Y. Tsai, W.-C. Lin, K. B. Cheng, J. Lee, and T.-Y. Lee, "Real-time physics-based 3d biped character animation using an inverted pendulum model," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 2, pp. 325–337, 2009.
[62] J. Gao, Z. Wang, Z. Xiao, J. Wang, T. Wang, J. Cao, X. Hu, S. Liu, J. Dai, and J. Pang, "Coohoi: Learning cooperative human-object interaction with manipulated object dynamics," *Advances in Neural Information Processing Systems*, vol. 37, pp. 79 741–79 763, 2024.
[63] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[64] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.
[65] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "Smoothnet: A plug-and-play network for refining human poses in videos," *ArXiv*, vol. abs/2112.13715, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245502027
[66] S. Xu, H. Y. Ling, Y.-X. Wang, and L.-Y. Gui, "Intermimic: Towards universal whole-body control for physics-based human-object interactions," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 266–12 277.

Wei Yao received the B.E. degree from the University of South China, Hengyang, China, in 2021. He is now a Ph.D. student in the School of Computer Science and Engineering at Nanjing University of Science and Technology, Nanjing, China. His research interests include computer vision, motion capture and embodied intelligence.

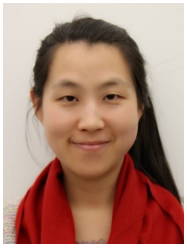
Haohan Ma is a Master's student at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences, Shenzhen, China. His research interests include robotics and motion generation, with a particular focus on multi-agent collaborative manipulation. He received his B.Eng. in Automation from South China University of Technology in 2023.

Hongwen Zhang received the B.E. degree from the South China University of Technology, Guangzhou, China, in 2015, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021, respectively. He has been working as a Post-Doctoral Researcher at Tsinghua University and is currently an Associate Professor at the School of Artificial Intelligence, Beijing Normal University. His research interests include computer vision, computer graphics, and their applications in 3D human modeling.

Yunlian Sun received the ME degree in computer science and technology from the Harbin Institute of Technology, China, in 2010 and the Ph.D. degree in ingegneria elettronica, informatica e delle telecomunicazioni from the University of Bologna, Italy, in 2014. After the Ph.D study, she worked as a postdoctoral researcher at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. She is currently an Associate Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research interests include biometrics, pattern recognition, and computer vision.

Liangjun Xing received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2024. He is currently pursuing the M.S. degree with the Department of Automation, Tsinghua University, Beijing, China, under the supervision of Prof. Yebin Liu. His current research interests include embodied AI, with a specific focus on mobile manipulation and dexterous manipulation for humanoid robots. His work involves reinforcement learning for whole-body control and high-level planning using vision-language-action (VLA) models.

Zhile Yang obtained his BSc and MSc from Shanghai University in 2010 and 2013 respectively, and received Ph.D. degree from Queen's University Belfast (QUB), UK. He is currently a professor in Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests focus on computational intelligence methods and their applications on smart grid and advanced manufacturing. he has led over 30 national and provincial research projects, published more than 230 SCI/EI-indexed papers in prestigious journals and conferences, with a total Google Scholar citation count exceeding 8,000 (H-Index=50). He holds 4 ESI highly cited papers, serves on 3 high-level SCI editorial boards in AI, organized 14 special issues as guest editor for top-tier international and domestic journals, and secured over 100 invention patents. His accolades include the Springer Nature First China New Development Award and the China Simulation Society Science and Technology Progress Prize. He has monetized over 30 intellectual property rights through cash and equity, incubated 4 high-tech AI enterprises, and received more than 10 international scientific awards.



Yuanjun Guo (Member, IEEE) received the B.Sc. degree in information engineering and the M.Sc. degree in optoelectronic engineering from Chongqing University, Chongqing, China, in 2008 and 2011, respectively, and the Ph.D. degree from the School of Electrical, Electronics and Computer Science, Queen's University Belfast (QUB), Belfast, U.K., in 2015. She is currently an Assistant Professor with the Shenzhen Institute of Science and Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include power big data analysis, artificial intelligence, fault diagnosis, and other applications in energy and power systems.

Yebin Liu (Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications in 2002, and the PhD degree from Automation Department, Tsinghua University in 2009. He is currently a full professor with Tsinghua University. He was a research fellow with the Max Planck Institute for Informatik, Germany, in 2010. His research interests include computer vision, computer graphics, and computational photography.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Nanjing Forestry University, Nanjing, China. He has authored more than 200 articles in toptier journals and conferences. His research interests include multimedia analysis and computer vision. Dr.Tang was a recipient of the Best Paper Awards in ACM MM 2007 and ACM MM Asia 2020, the Best Paper Runner-Up in ACM MM 2015. He has served as an Associate Editor for the IEEE TNNLS, IEEE TKDE, IEEE TMM, and IEEE TCSVT. He is a Fellow of IAPR.