Stability-Driven Motion Generation for Object-Guided Human-Human Co-Manipulation —-Supplementary Material—-

Jiahao Xu[1] Xiaohan Yuan[2] Xingchen Wu[1] Chongyang Xu[3] Kun Li[1] Buzhen Huang[1*]

[1]Tianjin University [2]National University of Singapore [3]Sichuan University

In this document, we provide the following supplementary contents:

- Affordance Training
- Manipulation strategy generation
- Interaction Prior
- Flow Matching Network
- User Study
- Simulation
- Result Analysis
- Limitation and Future Work

We also provide a demo video along with this document.

1. Affordance Training

To guide the generation of physically valid manipulation strategies, we train a network that predicts a dense affordance field on the object surface. Core4D provides sparse human–object contact points, which vary across demonstrations because identical object motions can be completed with different gripping styles. Following the dense supervision idea in LASO [2], we convert these sparse contacts into soft labels by assigning higher values to points near observed contact and lower values elsewhere. This produces a continuous graspability signal that reflects the regions most likely to support manipulation.

A lightweight encoder–decoder network takes the object point cloud as input and regresses a probability value α_k for each surface point. The model is trained using binary cross-entropy and Dice loss:

$$\mathcal{L}_{\text{aff-pred}} = \lambda_{\text{bce}}\,\text{BCE}(\alpha_k, \alpha_k^*) + \lambda_{\text{dice}}\,\text{Dice}(\alpha_k, \alpha_k^*), \quad (1)$$

which encourages accurate point-wise regression and stable spatial predictions. Since interactions in Core4D predominantly involve carrying and reorientation, we learn a single continuous affordance measure without defining multiple affordance categories.

The predicted affordance field is used in two ways during strategy generation. First, the per-point scores are concatenated with BPS features as conditioning for the diffusion model, providing a geometry-aware prior that highlights contact-suitable regions. Second, evaluating $\alpha(\hat{\mathbf{p}})$ at predicted anchors forms the availability term in the strategy loss (Eq. (8) in the main paper), which encourages the sampler to remain within high-affordance zones. Together, these components provide a stable prior that supports diverse yet physically plausible contact generation strategies.

2. Manipulation strategy generation

This section provides the architectural details of our manipulation strategy module, which predicts sparse 3D contact anchors used to guide the dual-human flow-matching generator. Unlike the temporal backbone, strategy generation is a static 3D prediction task. Therefore, we adopt a lightweight PointNet-based denoising network augmented with two attention blocks, rather than a full Transformer architecture, to ensure computational efficiency while preserving object-aware geometric reasoning.

Affordance-guided conditioning. Given the sampled object surface points $\{\mathbf{q}_k\}$ drawn from the object mesh $\mathcal{O}$, the affordance regressor described in the main paper produces a continuous graspability field α_k. Together with BPS descriptors computed around the object, these per-point features are embedded through a compact PointNet encoder:

$$\mathbf{z} = \text{PointNet}_{\text{cond}}\left(\{[\text{BPS}(p_k),\, \alpha_k]\}\right), \quad (2)$$

yielding a global latent code $\mathbf{z}$ that highlights regions suitable for manipulation. This latent code serves as the geometric prior for the diffusion model.

Diffusion-based anchor generation. We represent a manipulation strategy as M contact anchors $\mathcal{C} = \{\mathbf{p}_i\}_{i=1}^{M}$. During training, noisy anchors $\mathbf{c}_t$ are sampled at a diffusion timestep t, and encoded by an MLP with sinusoidal timestep embedding:

$$\mathbf{h}_i = \text{MLP}_{\text{enc}}\left([\mathbf{c}_{t,i}, \gamma(t)]\right). \quad (3)$$

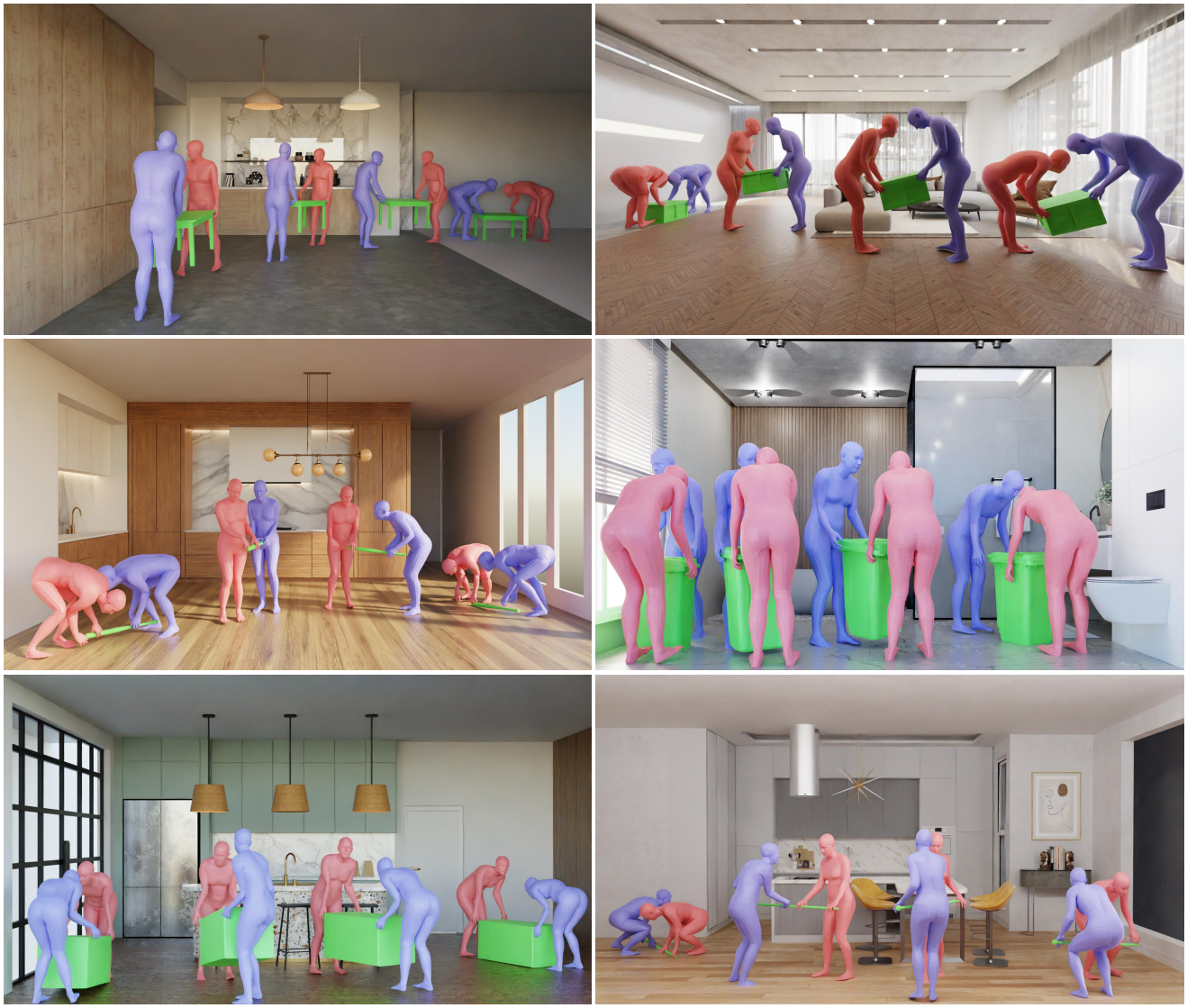

Figure 1. Results produced by our co-manipulation generation framework. Our method generates coordinated dual-human motions conditioned on the provided object trajectory that remain physically plausible across diverse scenes and object geometries. The two agents consistently maintain synchronized lifting and steering behaviors, exhibit natural interactions with substantially reduced interpenetration and floating. Fine-grained grasp readjustments and cooperative handling patterns can be observed throughout the sequences, highlighting the robustness and realism of the generated motions.

These anchor features undergo two lightweight attention refinements:

(1) Self-Attention. A multi-head self-attention block processes the set $\{\mathbf{h}_i\}$ to model inter-anchor structure such as symmetry, pairing, and spatial consistency. Since M is small (3-10 anchors), this operation is computationally negligible yet significantly stabilizes the denoising process.

(2) Cross-Attention. To inject object-awareness, the self-attended anchor features attend to the object latent $\mathbf{z}$ via a cross-attention module:

$$\tilde{\mathbf{h}}_i = \text{CrossAttn}(\mathbf{h}_i, \mathbf{z}), \tag{4}$$

allowing each anchor to selectively focus on high-affordance and geometrically valid object regions. This block replaces the heavier Transformer encoder–decoder structure while retaining the essential geometric reasoning capability.

Finally, a decoder MLP predicts the denoising residual:

$$\hat{\varepsilon}_i = \text{MLP}_{\text{dec}}(\tilde{\mathbf{h}}_i), \tag{5}$$

and the diffusion loss follows the standard noise-prediction objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\varepsilon} \left\| \varepsilon - \varepsilon_\theta(\mathbf{c}_t, t, \mathbf{z}) \right\|_2^2 . \tag{6}$$

Strategy supervision. The diffusion objective is jointly optimized with the strategy-consistency loss from Eq. (6)-(8) of the main paper:

$$\mathcal{L}_{\text{str}} = \mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{aff}}, \tag{7}$$

which enforces positional correctness, normal alignment, and high-affordance validity of the predicted anchors. This multi-term supervision significantly reduces implausible grasping choices and stabilizes downstream motion generation.

Sampling. During inference, DDIM sampling starts from Gaussian noise and iteratively denoises the anchor set through the attention-enhanced diffusion model. The resulting contact anchors $\hat{\mathcal{C}}$ are passed to the flow-matching module as manipulation strategies, enabling dual-human motions that respect both object geometry and affordance-driven contact priors.

3. Interaction Prior

Our adversarial regularization uses two discriminators that operate at complementary scales. The pose prior receives short motion clips from one character at a time. InterX sequences are labeled as positive examples because they contain clean, object-free bimanual motions, whereas Core4D single-character crops and the generator's own samples are labeled as negatives since they often reveal payload-induced artifacts. Each clip passes through stacked temporal convolutions and residual blocks to produce a realism score. We optimize the discriminator with a non-saturating binary cross-entropy objective. The generator is updated with the opposite gradient so that its outputs move toward the high-realism region, which improves the naturalness of individual motions and smooths joint transitions.

The interaction discriminator targets cooperative timing and improving interactive quality. Core4D ground-truth pairs serve as positives and generated dual-human rollouts serve as negatives. We concatenate both agents' pose features, add their relative root transforms, and process the sequence with a lightweight transformer followed by a linear classifier to obtain a coordination score. The non-saturating binary cross-entropy loss is used to train the interaction discriminator, rewarding genuine Core4D interactions and penalizing synthesized ones that exhibit misaligned forces or delayed grasps.By learning from these signals, the generator adapts its outputs to mirror the interpersonal rhythm and shared-load behavior observed in the dataset, raising the overall quality of cooperative manipulation. Since Inter-X does not provide object trajectories, it cannot be used for trajectory-conditioned quantitative evaluation. Therefore, we use Inter-X only during training to regularize motion realism via the adversarial interaction prior, while all quantitative evaluations are conducted on Core4D.

4. Flow Matching Network

The flow-matching backbone models two characters jointly across time. For each frame we concatenate a character's 24 joint rotations in 6D form, body-shape coefficients, and global translation into a 157-dimensional vector. A shared linear layer lifts these vectors to a 512-dimensional embedding, after which sinusoidal time encodings are added to preserve temporal order. The two characters streams are processed in parallel by eight stacked transformer blocks; each block performs self-attention along the temporal axis of one character and cross-attention to the other character's hidden states, allowing the model to exchange load-balancing cues while maintaining identical weights for both agents.

The final hidden activations are projected back to the original motion space with linear heads that predict the instantaneous flow for every frame and agent. Training uses the flow-matching objective that regresses these velocities to the finite-difference targets derived from paired clean and noisy trajectories. This architecture offers three advantages: temporal self-attention captures long-range dependencies within each character, symmetric weight sharing keeps the partners' responses consistent, and the cross-agent attention explicitly models coordination, yielding precise co-manipulation trajectories without excessive parameters. In practice, we set $\gamma = 0.8$ and $\eta = 0.2$, as a large γ reduces motion naturalness by over-constraining wrist positions, while a large η degrades contact accuracy. The contact and prior guidance terms are applied at all K integration steps. The simulation is applied only at the penultimate step, after which $\mathbf{x}_\tau$ is replaced by $\tilde{\mathbf{x}}_\tau$ as input to the final integration step to recover motion naturalness.

5. User Study

Because our task generates dual-human collaboration conditioned on an object trajectory, purely quantitative metrics cannot fully capture motion quality. In particular, we care whether the generated motion plausibly explains how two people cooperate to realise the given object dynamics. We therefore conduct a small-scale user study.

We randomly select 30 test trajectories and render 3-second videos for our method and three baselines ComMDM [4], OMOMO [1], and InterGen [3]. For each trajectory, we present our method and one baseline side-by-side in random order without method labels. Twelve graduate students and researchers with experience in vision participate in the study.

For each video pair, participants rate both methods on a 5-point Likert scale along three aspects: (i) how well the dual-human motion explains the prescribed object trajectory, (ii) the perceived quality of collaboration between the two humans, and (iii) the plausibility of hand-object con-

tact and overall motion naturalness. Participants can replay videos freely and complete the session independently.

Results. Table 1 reports the average scores across all participants. Our method consistently outperforms the baseline under all criteria. The largest improvements appear in trajectory explanation and cooperation quality, suggesting that affordance-informed manipulation strategies help participants better understand how the dual-human interaction realises the object motion. Contact plausibility also improves, indicating that our anchor-guided flow reduces hand-object inconsistencies and enhances perceived naturalness.

Table 1. **User study results.** Mean Likert scores (↑ higher is better).

Method	Traj. Expl.	Cooperation	Contact/Natural.
ComMDM	2.31 ± 0.62	2.44 ± 0.58	2.51 ± 0.55
OMOMO	2.58 ± 0.67	2.63 ± 0.61	2.72 ± 0.60
InterGen	2.74 ± 0.65	2.81 ± 0.63	2.77 ± 0.59
Ours	$\mathbf{3.89 \pm 0.63}$	$\mathbf{3.96 \pm 0.59}$	$\mathbf{3.82 \pm 0.60}$

Analysis. Across all 30 trajectories, participants prefer our method over the baselines in 83% of comparisons. These subjective results corroborate the quantitative evaluations in the main paper, confirming that affordance-guided strategy generation significantly improves interpretability, coordination quality, and overall physical plausibility in dual-human manipulation.

6. Simulation

To physically validate and refine the generated co-manipulation motions, we run a short-range stability-driven simulation at the penultimate flow-matching step. Before describing the refinement procedure, we first explain how the SMPL-X body is converted into a physics-ready humanoid model.

SMPL-X humanoid construction. Given the SMPL-X parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \gamma)$ decoded from $\mathbf{x}_\tau$, we construct a rigid-body humanoid following the SMPL-X kinematic tree. Using the official joint regressor, we compute the rest-pose joint locations in T-pose, from which we derive the bone lengths for each limb. For symmetric limbs (e.g., left/right arms and legs), we average their T-pose bone lengths to eliminate small mesh asymmetries. Each limb is represented as a capsule or box link aligned with the parent–child joint direction, and the link shapes are generated automatically from the computed bone lengths.

A fixed density is assigned to all links such that the total mass remains constant across different shape coefficients, preventing instability when varying $\boldsymbol{\beta}$. The root is modeled as a floating base with 6-DoF, while elbows, knees, and ankles use hinge joints with one-axis limits, and shoulders, hips, and wrists use three-axis ball joints with SMPL-X–consistent angular limits. Hands and feet are kept rigid for stability, following common practice in physics-based humanoid control. The resulting articulated body provides consistent mass distribution, joint structure, and link geometry for the simulation module.

Simulation refinement. With the humanoids instantiated, the decoded SMPL-X pose and translation initialize the two agents and the manipulated object in the physics engine. A proportional-derivative (PD) controller tracks the target configuration while respecting joint limits and hand-object contacts.

Starting from the nominal prediction $\mathbf{x}_\tau$, we sample corrective offsets $\Delta\mathbf{x}_\tau$ from a multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$, producing PD targets

$$\bar{\mathbf{x}}_\tau = \mathbf{x}_\tau + \Delta\mathbf{x}_\tau.$$

The distribution is initialized as standard normal and iteratively updated using CMA-ES. For each rollout, the PD controller applies torques toward $\bar{\mathbf{x}}_\tau$, and the resulting motion is evaluated with

$$\mathcal{L}_{\text{phys}} = \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{sta}}. \tag{8}$$

The similarity term

$$\mathcal{L}_{\text{sim}} = \left\|\tilde{\mathbf{x}}_\tau - \mathbf{x}_\tau\right\|_2^2 + \left\|\tilde{R}^o - R^o\right\|_2^2 + \left\|\tilde{\mathbf{d}}^o - \mathbf{d}^o\right\|_2^2 \tag{9}$$

keeps simulated humanoids and object aligned with the generator's proposal. The stability term

$$\mathcal{L}_{\text{sta}} = \frac{\left\|\vec{f}(t) - M_o\vec{a}\right\|_2^2}{\|M_o\vec{g}\|_2^2} + \frac{\left\|\vec{\tau}(t) - I_o\vec{\alpha}\right\|_2^2}{\|I_o\vec{\alpha}\|_2^2} + e^{-m(t)} \tag{10}$$

penalizes mismatches between simulated forces/torques and the accelerations implied by the object trajectory.

CMA-ES retains the rollout with the lowest $\mathcal{L}_{\text{phys}}$, overwrites $\mathbf{x}_\tau$ with the simulated state $\tilde{\mathbf{x}}_\tau$, and passes it to the last integration step. This refinement suppresses floating artifacts and maintains object stability as described in the main paper.

7. Result Analysis

The qualitative examples in Fig. 1, together with the quantitative metrics and user evaluations, demonstrate the advantages of each component. Compared with recent diffusion based baselines, our flow matching backbone accelerates inference while preserving coordinated dual agent

motion, enabling fast sampling without compromising synchronization or interaction consistency. The manipulation strategy generator leverages the affordance field and BPS based geometric descriptors to expand the diversity of feasible grasping strategies. The predicted contact anchors are further constrained by positional and normal consistency terms, which enforce geometrically valid surface contacts, reduce interpenetration, and support smoother regrasp adjustments around the object. These effects can be observed qualitatively in Fig. 1, where our method produces cleaner and more accurate hand–object interactions. The adversarial interaction prior enhances the naturalness of individual poses and interpersonal timing through discriminator driven refinement of generated trajectories, effectively suppressing jitter and unnatural pacing that often appear in baseline outputs. Finally, the stability driven simulation aligns hand object interactions with the underlying object dynamics and suppresses floating artifacts, resulting in more stable grasp maintaining motions over long sequences. Taken together, these components produce realistic and intention consistent co manipulation that follows the target trajectory more reliably across diverse scenes and object geometries than prior approaches.

Table 2. **Foot sliding comparison on Core4D-S1.**

Method	InterGen	Ours	Ours + $\mathcal{L}_{\text{foot}}$
Foot sliding ↓	0.71	0.68	0.60

8. Limitation and Future Work

While our framework delivers intention-aware co-manipulation, it inherits the biases of existing dual-human–object datasets (e.g., Core4D), whose curated interactions concentrate on collaborative carrying and support tasks, leaving richer co-manipulation skills underrepresented. Conditioning solely on preplanned object trajectories also makes the generator sensitive to trajectory noise or missing contact intent, hindering scenarios where partners improvise or switch grasp roles mid-task. Moreover, the stability surrogate must be recalibrated with simulator feedback whenever objects with new physical properties are introduced, delaying deployment across diverse assets. Future work could couple the model with online perception that corrects trajectories in real time, broaden data collection to cover richer collaborative skills, and develop stability estimators that generalize across object categories without per-asset retraining.

References

[1] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 3

[2] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 1

[3] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 3

[4] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3