PrimHOI: Compositional Human-Object Interaction via Reusable Primitives

Kai Jia[1,2*], Tengyu Liu[2*], Yixin Zhu[3], Mingtao Pei[1✉], Siyuan Huang[2✉]
[1] School of Computer Science & Technology, Beijing Institute of Technology
[2] National Key Laboratory of General Artificial Intelligence, BIGAI [3] School of Psychological and Cognitive Sciences, Peking University
[*] Equal contributors ✉peimt@bit.edu.cn, syhuang@bigai.ai Project Website: https://kairobo.github.io/PrimHOI/

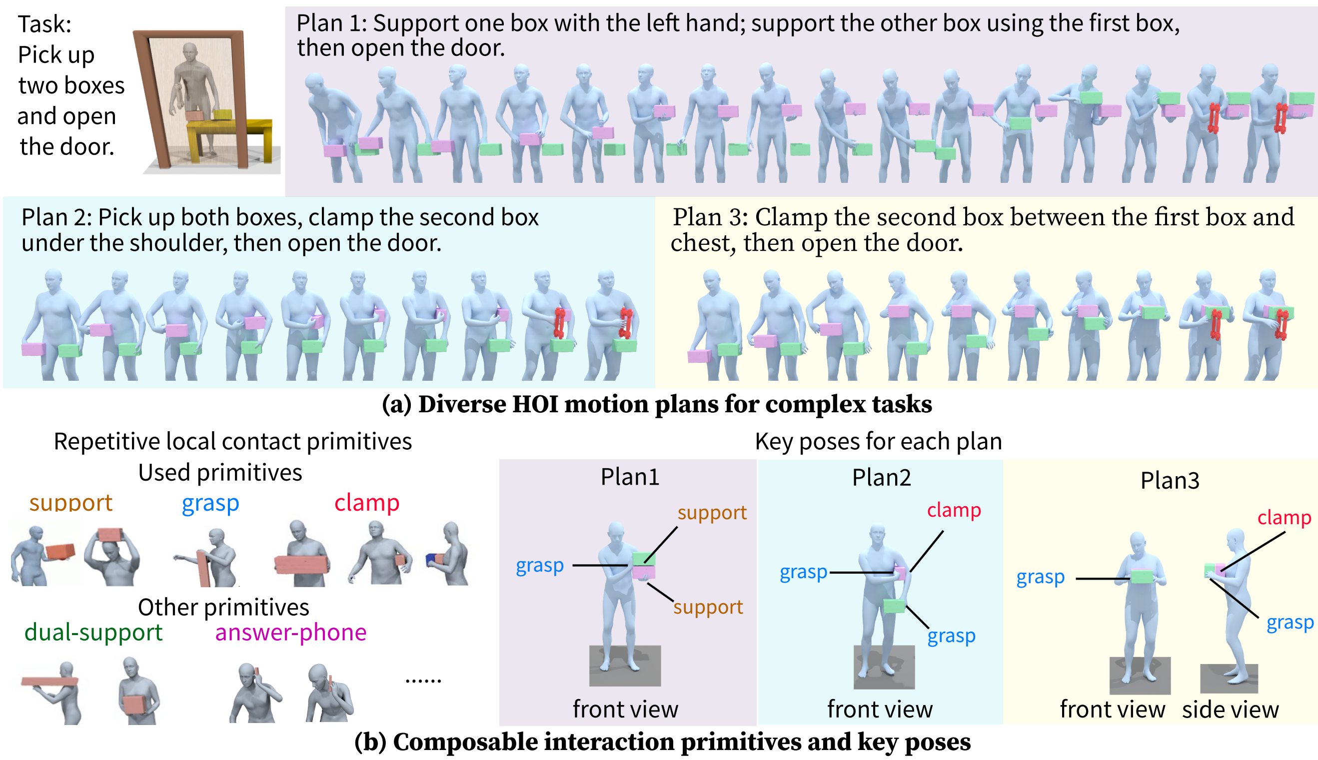


Figure 1. **Diverse HOI motions for complex tasks generated by PrimHOI**. Given an unseen high-level task description, our **PrimHOI** plans and generates diverse HOI motions that fulfill task requirements through spatial and temporal composition of generalizable interaction primitives. These primitives capture repetitive local contact patterns from everyday interactions, enabling systematic reuse across different scenarios. **PrimHOI** achieves zero-shot transfer to unseen HOI tasks without requiring task-specific training data.

Abstract

Synthesizing realistic Human-Object Interaction (HOI) motions is essential for creating believable digital characters and intelligent robots. Existing approaches rely on data-intensive learning models that struggle with the compositional structure of daily HOI motions, particularly for complex multi-object manipulation tasks. The exponential growth of possible interaction scenarios makes comprehensive data collection prohibitively expensive. The fundamental challenge is synthesizing unseen, complex HOI sequences without extensive task-specific training data. Here we show that ***PrimHOI*** *generates complex HOI motions through spatial and temporal composition of generalizable interaction primitives defined by relative geometry. Our approach demonstrates that repetitive local contact patterns—grasping, clamping, and supporting—serve as reusable building blocks for diverse interaction sequences. Unlike previous data-driven methods requiring end-to-end training for each task variant,* ***PrimHOI*** *achieves zero-shot transfer to unseen scenarios through hierarchical primitive planning. Experimental validation demonstrates substantial improvements in adaptability, diversity, and motion quality compared to existing approaches.*

1. Introduction

Synthesizing diverse, realistic HOI motions from simple instructions is essential for character animation [2, 8, 9, 12, 15, 16, 22, 28, 41, 42] and embodied AI applications [13, 32, 58, 59]. Current approaches map semantic descriptions to HOI motions [20, 26, 45, 53, 54], but struggle with the nuanced complexity of everyday interactions that require coordinated, interdependent object manipulation. Consider a seemingly simple task: picking up two boxes and opening a door. This requires one hand to be freed for door operation while the other manages both boxes, possibly with torso assistance. Such interactions demand both spatial composition—coordinating object positions and states—and temporal composition—sequencing actions over time, as shown in Fig. 1. Current methods struggle with these intricate motions as they face challenges in capturing inter-element dependencies, while the exponentially growing space of possible interactions makes comprehensive data collection prohibitively expensive. In contrast, humans excel at adapting prior skills to novel tasks through systematic generalization [18, 23, 27, 39, 51], reusing knowledge by recognizing similarities between familiar and new situations. This observation raises a fundamental question: how can we represent and reuse prior HOI knowledge as adaptable primitives for unseen tasks?

Recent studies have explored compositional motion generation through spatial composition of part-level motions [4, 17, 30] and temporal composition of motion segments [3, 10, 11, 24, 50]. However, these approaches focus primarily on spatial or temporal composition alone, leaving spatiotemporal compositional HOI generation largely unexplored. While UniUSI [54] and InterDreamer [54] have made initial attempts at compositional HOI generation, they are limited by either static object constraints or restrictive whole-body representations that prevent flexible object dynamics and precise local interaction control.

Motivated by these limitations, we propose a new approach based on the insight that repetitive geometric patterns emerge in localized regions of interaction [5, 41, 60]. Rather than relying on whole-body representations, we observe that simple interaction types like *support* or *clamp* can be reused across various body parts or objects while maintaining consistent geometric relationships (see Fig. 1). We formalize these consistent patterns as interaction primitives—reusable building blocks that capture essential geometric and semantic information of local interactions. This primitive-based representation enables decomposition of complex HOI tasks into learnable components that can be flexibly combined for unseen scenarios.

Building on this insight, we introduce **`PrimHOI`**, a hierarchical HOI generation framework that orchestrates interaction primitives to accomplish complex tasks from high-level descriptions. Our approach operates through three key stages: high-level planning that decomposes tasks into sequences of interaction primitives using our symbolic reasoning framework PDDL-HOI, key pose generation that instantiates these primitives into specific human-object configurations, and intermediate motion generation that creates smooth transitions between key poses. We represent planning problems as *subgoal graphs*—compositional symbolic structures where nodes represent manipulable objects and manipulators, while edges encode physical constraints based on interaction primitives. To generate action sequences, we develop PDDL-HOI by extending PDDLStream [14] and leverage Large Language Model (LLM)-based task translation to convert high-level descriptions into executable plans. For motion generation, we sample contact points using primitive contact models [25], optimize human poses with pose priors [33], and guide intermediate motion generation [52] using planned object trajectories.

Our contributions are as follows:

- We introduce interaction primitives—a generalizable representation of HOI patterns based on relative geometry between objects and body parts. This representation enables flexible reuse across different body parts and objects, allowing complex interactions to be decomposed into learnable, transferable components.
- We develop PDDL-HOI, a symbolic planning framework that leverages our primitive representation to enable systematic composition of interaction sequences. Combined with LLM-based task translation, this approach supports diverse and complex HOI scenarios through zero-shot generalization.
- We present a complete hierarchical synthesis pipeline that generates realistic HOI motions from high-level task descriptions. Our method demonstrates strong generalization capabilities, synthesizing novel multi-object interactions without requiring task-specific training data.

2. Related Work

Guided Human Motion Generation Generating human motion from limited guidance such as text [19, 20, 26, 28, 36, 45, 48, 55], object trajectories [25], and spatial constraints [21, 30, 40, 44, 46, 52] has broad applications in animation and robotics. Early approaches like TEMOS [36] employed cVAEs for text-to-motion mapping, while recent methods like MDM [43] leverage diffusion models for improved distribution modeling. For precise spatial control, OmniControl [52] adapts ControlNet [56] to provide guidance during diffusion, and ProgMoGen [30] achieves fine-grained control through latent optimization.

Extending these approaches to HOI motion generation introduces additional complexity due to coordinated human-object dynamics. IMoS [15] generates text-conditioned human motion and attaches objects to hands but lacks lower-body coordination. OMOMO [25] synthe-

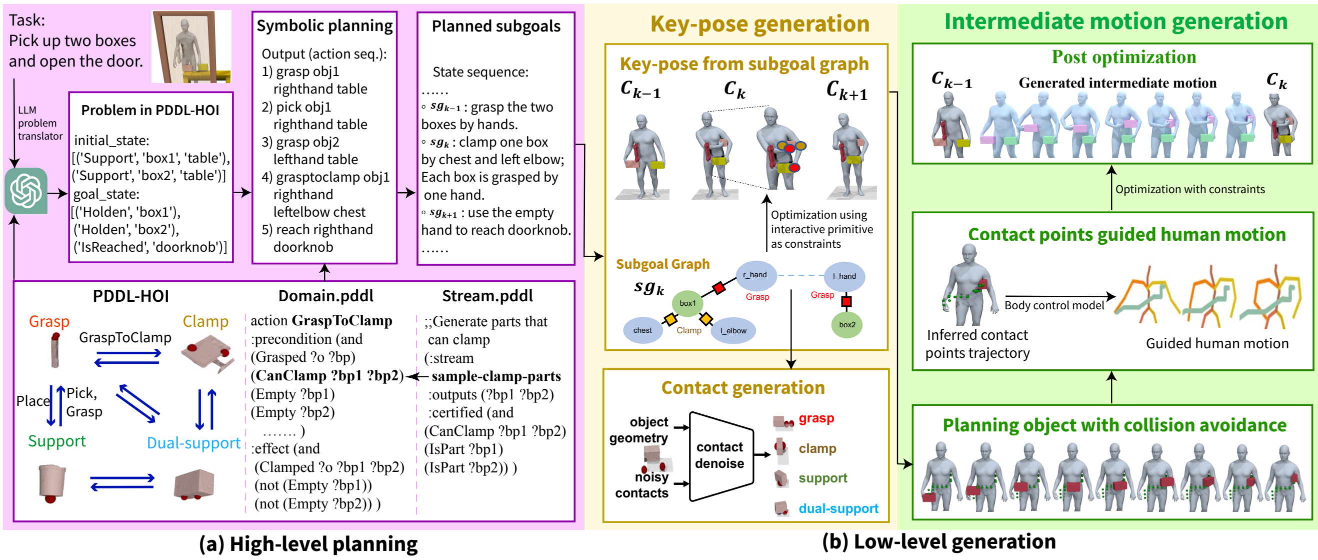


Figure 2. **Overview of PrimHOI. (a) High-Level planning:** Given a task description, an LLM translates it into a PDDL problem. Our PDDL-HOI defines actions (*e.g.*, GraspToClamp) with preconditions and effects, and generates valid body part combinations for interaction primitives. The symbolic planner produces an action sequence π_l with corresponding subgoals. **(b) Low-Level generation** includes two components. **Key pose generation:** For each subgoal, we sample contact points from interaction primitives (*e.g.*, grasp, clamp, support), then optimize human poses to satisfy these contact constraints, generating key poses C_k. **Intermediate motion generation:** We plan object trajectories between key poses and generate human motion guided by contact trajectories. A post-optimization step refines the motion to ensure smoothness, eliminate penetrations, and maintain consistency with subgoal constraints.

sizes human motion from given object trajectories, while CHOIS [26] extends this with text-based control. Recent works [12, 35] integrate affordance prediction to reduce explicit trajectory guidance. However, these data-driven approaches struggle with long-horizon, multi-object scenarios that require complex spatiotemporal reasoning beyond what can be captured in training data.

Compositional Human Motion Generation To address the limitations of end-to-end approaches, compositional methods enhance systematic generalization by decomposing complex motions into reusable components [6, 31, 38]. These approaches operate through two primary strategies: temporal composition, which sequences motion segments over time, and spatial composition, which coordinates concurrent body part movements.

Temporal composition methods focus on creating coherent motion sequences from discrete segments. TEACH [3] and Multi-Act [24] learn smooth transitions between motion primitives, while UniHSI [47, 50] employs LLM-based planning to generate contact point sequences for scene interaction. InterDreamer [54] extends this to HOI generation using LLM for high-level planning and text-to-action modules for low-level synthesis. Recent work by Wu *et al.* [49] combines LLM planning with scene parsing for temporal sequencing to ensuring physical plausibility.

Complementing temporal approaches, spatial composition methods coordinate simultaneous body part movements. SINC [4] uses GPT-3 to assign motion factors to different body parts but struggles with conflicting concurrent motions. CoMo [17] addresses this limitation by decomposing motions into distinct part-level codes, while ProgMoGen [30] breaks high-level tasks into atomic constraints for flexible motion editing. STMC [37] provides a unified framework combining both temporal and spatial composition through separate denoising and compositional re-denoising processes.

While these advances have significantly improved motion generation capabilities, most focus on either spatial or temporal composition in isolation, primarily for single-person scenarios. The challenge of spatiotemporal compositional HOI generation—where multiple objects must be manipulated through coordinated spatial and temporal reasoning—remains largely unexplored. Our work addresses this gap by introducing interaction primitives that enable systematic decomposition and flexible recombination of both spatial and temporal HOI components for complex multi-object scenarios.

3. The PrimHOI Framework

PrimHOI synthesizes complex Human-Object Interaction (HOI) motion sequences from high-level task descriptions. Given a natural language task T (*e.g.*, "pick up two boxes and open the door"), initial object layout L_0, and human pose $x^h_{t=0}$, our goal is to generate a complete motion sequence $x = \{x^h, x^O\}$ that accomplishes the specified task. Here, x^h represents the human motion in SMPLX format,

x^O denotes object trajectories, and $L_0 = \{x^o_{t=0}\}_{o \in O}$ specifies initial poses for the set of objects O.

Directly generating x from high-level descriptions poses significant challenges due to the inherent complexity of HOI motions. These tasks require coordinated handling of both spatial composition—managing multi-part interactions across different body regions—and temporal composition—sequencing multiple sub-tasks over extended horizons. To address this complexity, we decompose the motion into *subgoals* based on interaction primitives, where each primitive defines a local contact pattern (*e.g.*, support, grasp, clamp, dual-support) between body parts and objects.

We represent subgoals as graphs sg that describe interaction predicates between objects and body parts (see Fig. 2). Each element corresponds to an interaction primitive $P_i = \{o_m, f, \alpha\}$, where o_m is an object, f specifies the contact type (*e.g.*, grasped, clamped), and α represents the interacting body part or object. The set $A = \{\alpha\}$ encompasses all manipulator parts including body parts and objects O that can interact with other objects.

Following this subgoal-driven approach, we introduce an intermediate planning process to generate subgoals from the task description T. This expands our problem to jointly sampling motion x and plan π from $P(x, \pi | T, C_0)$, which we decompose as:

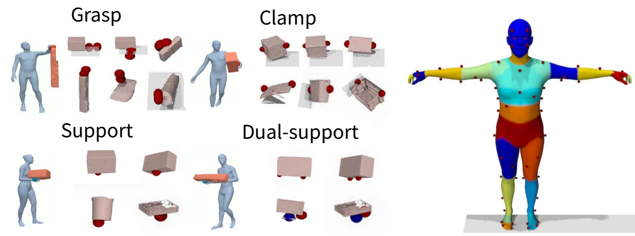
$$x, \pi \sim P(x, \pi | T, C_0) = P(x | \pi, C_0) P(\pi | T, C_0). \quad (1)$$

Our three-stage pipeline first generates a high-level plan $\pi = \{\text{sg}_k\}_{k=1}^K$ using PDDL-style planning with LLM, leveraging domain knowledge from PDDL-HOI to define the planning space. Subsequently, subgoals are translated into specific contact positions and keyframe poses $\{C_k\}_{k=1}^K$, where $C_k = \{L_k, x^h_{k,t=0}\}$ represents object layout and human pose at the beginning of segment k. Finally, intermediate motion generation bridges consecutive key poses, with $L_k = \{x^o_{k,t=0}\}, o \in O$ and t denoting the frame index within segment k.

3.1. Interaction Primitive Generation

Our approach relies on four manually classified interaction primitives that capture fundamental contact patterns in HOI motions, as illustrated in Fig. 3a: support, grasp, clamp, and dual support. These primitives serve as building blocks for representing complex manipulation behaviors through their spatial and temporal combinations.

For each interaction primitive P, we generate object contact points $\{p^o_i\}^P$ using a diffusion-based model $P(\{p^o_i\}^P | \mathbf{V})$, where $\mathbf{V} \in \mathbb{R}^{K \times 3}$ represents the object mesh vertices and i indexes individual contact points. This data-driven approach learns contextually appropriate contact locations from training data, ensuring generated contacts align with natural interaction patterns.



(a) Interaction primitives (b) Body contact points

Figure 3. **Contact representations used in `PrimHOI`.** (a) The four interaction primitives that serve as building blocks for complex manipulation behaviors: *support*, *grasp*, *clamp*, and *dual support*. Each primitive defines a specific contact pattern between body parts and objects, with contact points shown relative to object surfaces. *Grasp* includes two contact points (wrist and hand) to capture grasping direction, while *clamp* and *dual support* each involve two contact points, and *support* requires only one contact point. (b) Body contact points (red dots) are strategically selected from mocap markers and manual curation, with each body part shown in a different color to illustrate the discrete vocabulary of candidate contact locations.

On the body side, we define a discrete set of candidate contact points $\{p^h_i\}$ selected from mocap markers [57] and manual curation, as shown in Fig. 3b. While this vocabulary is finite, it provides sufficient expressiveness to cover the wide range of contact configurations encountered in common manipulation tasks, striking a balance between computational efficiency and representational power.

3.2. High-Level Planning

The high-level planning process transforms natural language task descriptions into structured sequences of interaction subgoals, as depicted in Fig. 2. We adapt the Planning Domain Definition Language (PDDL) [1] and its extension PDDLStream [14] to create PDDL-HOI, our specialized HOI planning language that integrates symbolic planning with constraint sampling.

Leveraging LLM capabilities [29, 34, 54], task descriptions are translated into PDDL problem formats where interaction primitives become predicates describing interaction states. For example, predicates `(Grasped box1 righthand)` and `(Clamped box1 chest left_elbow)` jointly describe a state where `box1` is simultaneously grasped and clamped. Actions represent state transitions that modify these predicates—the action `GraspToClamp` transitions an object to a clamped state, but only when preconditions are satisfied (*e.g.*, clamp parts are empty and the object is already grasped).

To generate diverse planning solutions, we incorporate PDDLStream's *streams* concept. By removing predicates that specify which body parts perform specific primitives, the planner dynamically samples valid body part assignments during planning, enabling varied manipulation strategies for the same task. This process produces multiple plan candidates $\{\pi_l\}_{l=1}^N$ from initial condition C_0, each representing different sequences of subgoal predicates that directly transfer to subgoal graphs. Additional details are provided in Appendix A.1.

3.3. Low-Level Generation

The low-level generation creates detailed motion sequences from abstract high-level plans through two main steps: generating key poses and producing intermediate motion connecting these poses. This process is formulated as:

$$P(x \mid \pi, C_0) = \sum_{\{C_k\}_{k=1}^K} P(x \mid \{C_k\}_{k=1}^K) \\ P(\{C_k\}_{k=1}^K \mid \{sg_k\}_{k=1}^K, C_0), \tag{2}$$

3.3.1. Key-pose Generation

We transform planned subgoal graphs $\{\text{sg}_k\}_{k=1}^K$ into specific key poses $\{C_k\}_{k=1}^K$ sequentially from initial pose C_0:

$$P(\{C_k\}_{k=1}^K \mid \{sg_k\}_{k=1}^K, C_0) = \prod_{k=0}^{K-1} P(C_{k+1} \mid sg_{k+1}, C_k), \tag{3}$$

For each key pose C_k, we consider contact point locations on objects, object placement, and natural body pose maintenance [33]. Contact points on object surfaces are sampled using the primitive contact model $P(\{p_i^o\}^P|\mathbf{V_o})$. When multiple primitives involve the same object, they are grouped into *interaction primitive groups*, and compatible contact configurations are selected to avoid conflicts.

Object poses $\{x_{k+1,t=0}^o\}$ are sampled from an object placement prior $P(s_o|\{p_i^h\} = \{p_i^o\}^P)$ that aligns body and object contact points, where $s_o = x_{k+1,t=0}^o$ for brevity. We use a Mixture of Gaussians for this prior, placing objects near frequently used body regions. The body pose $x_{k+1,t=0}^h$ is then optimized with body prior regularization to align with contact points while incorporating normal constraints for certain primitives:

$$P(C_{k+1} \mid sg_k, C_k) = \sum_{p_i^o, s_o} P(x_{k+1,t=0}^h \mid \{p_i^h\}, x_{k,t=0}^h) \\ \prod_{P_i \in sg_k} P(s_o \mid \{p_i^h\} = \{p_i^o\}^{P_i}) P(\{p_i^o\}^{P_i}|\mathbf{V_o}), \tag{4}$$

3.3.2. Intermediate Motion Generation

After obtaining consecutive key poses, we generate intermediate HOI motion segments to produce the complete sequence:

$$P(x \mid \{C_k\}_{k=1}^K) = \prod_{k=0}^{K-1} P(x^k \mid C_{k+1}, C_k), \tag{5}$$

where $x^k = \{x_O^k, x_h^k\}$ represents the motion segment between key poses C_k and C_{k+1}.

The generation process operates in two stages. First, object trajectories are planned using A* algorithm with SDF-based collision checking as $P(x_O^k|C_k, C_{k+1})$, ensuring smooth transitions and collision avoidance. Second, given the inferred contact point sequence $\{p_{i,t}^h\}_{t\in T_k}$ from object trajectories, human motion is generated using a spatial-guided diffusion model (OmniControl [52]) as $P(x_h^k|\{p_{i,t}^h\}_{t\in T_k}, C_k, C_{k+1})$. The complete formulation is:

$$P(x_O^k, x_h^k \mid C_{k+1}, C_k) = P(x_h^k|\{p_i^t\}_{t\in T_k}, C_k, C_{k+1}) \\ F(\{p_i^t\}_{t\in T_k}|x_O^k, C_k, C_{k+1}) P(x_O^k|C_k, C_{k+1}), \tag{6}$$

where $F(\{p_i^t\}|x_O^k, C_k, C_{k+1})$ infers body contact points by maintaining consistent contact positions relative to objects.

We refer readers to Appendix A.2 for additional details.

3.4. Post-refinement Process

While the initial generative HOI motion provides a plausible sequence, it may lack precise adherence to physical constraints and contact accuracy. To enhance realism and correctness, we apply a post-optimization process to refine the human motion [30, 52, 54]. This optimization maintains interaction primitive constraints while minimizing collisions and penetrations.

The optimization objective E_{opt} comprises six complementary terms: contact closeness (E_{contact}), contact normal alignment (E_{normal}), body-object collision penalty (E_{colli}), body self-penetration prevention (E_{pene}), temporal smoothness (E_{temp}), and body pose regularization (E_{prior}) [33]. The complete optimization objective is formulated as:

$$E_{\text{opt}} = \lambda_{\text{contact}} E_{\text{contact}} + \lambda_{\text{normal}} E_{\text{normal}} + \lambda_{\text{colli}} E_{\text{colli}} \\ + \lambda_{\text{pene}} E_{\text{pene}} + \lambda_{\text{temp}} E_{\text{temp}} + \lambda_{\text{prior}} E_{\text{prior}}, \tag{7}$$

where the λ terms control the relative importance of each constraint. Specific formulations of these loss terms are detailed in Appendix A.3.

4. Experiments

We evaluate **PrimHOI**'s ability to generate compositional HOI motions through systematic assessment of both high-level planning and low-level motion generation capabilities. Unlike prior text-to-motion approaches [35, 53], our focus centers on achieving generalization to novel task compositions using modular interaction primitives. Our evaluation encompasses quantitative metrics for high-level planning (Sec. 4.2) and low-level generation (Sec. 4.3), complemented by qualitative analysis (Sec. 4.4). Additional experimental details and results are provided in the supplementary material.

4.1. Implementation Details

We adapt PDDLStream [14] for symbolic planning in PDDL-HOI, enabling structured reasoning about interaction sequences. The diffusion-based contact generation model from OMOMO [25] is modified to predict individual contact points rather than temporal sequences, with normalization applied to enhance generalization across diverse object geometries. Contact data collection follows a multi-source approach: *clamp* primitives utilize data from OMOMO [7], *grasp* primitives draw from BEHAVE [25], while *support* and *dual support* primitives employ analytical functions.

For human motion generation guided by contact constraints, we retrain OmniControl [52] with enhanced local control capabilities, termed *LocalControl*. Since OmniControl does not directly accept contact point guidance, we train a regressor mapping SMPL-X keypoints to our selected contact points (Fig. 3b), enabling gradient and realism guidance integration. Body pose optimization incorporates DPoser [33] as a diffusion-based prior that accommodates incomplete keypoint targets. Complete implementation details are provided in Appendix A.

4.2. High-Level Planning Evaluation

To validate our structured planning approach, we compare PDDL-HOI against three baseline methods: *GPT-4o* (direct task-to-plan generation), *GPT-4o + Primitives* (incorporating interaction primitive definitions as prior knowledge), and *GPT-4o + PDDL-HOI* (our hybrid approach).

Evaluation Metrics We assess planning quality using three complementary metrics: **Success Rate** measures the proportion of plans that successfully complete the task, **Plan Efficiency** quantifies the mean number of actions in successful plans, and **Solution Diversity** counts the number of different plans among successful ones (excluding left-right symmetry). Human evaluators assessed these metrics across three tasks (Tab. 1).

Table 1. **High-level planning performance comparison across task complexity levels.** We evaluate each method on Task 1 and Task 2 (5 trials each) and Task 3 (10 trials). Our *GPT-4o + PDDL-HOI* approach demonstrates superior performance in success rate and solution diversity, while maintaining competitive plan efficiency across all complexity levels.

Task 1: Pick up two boxes from table

Method	Success Rate	Plan Efficiency	Solution Diversity
GPT-4o	4.0/5	5.7	1.6/5
GPT-4o + PDDL-HOI (ours)	5.0/5	**4.6**	2.0/5

Task 2: Carry long box passing the door

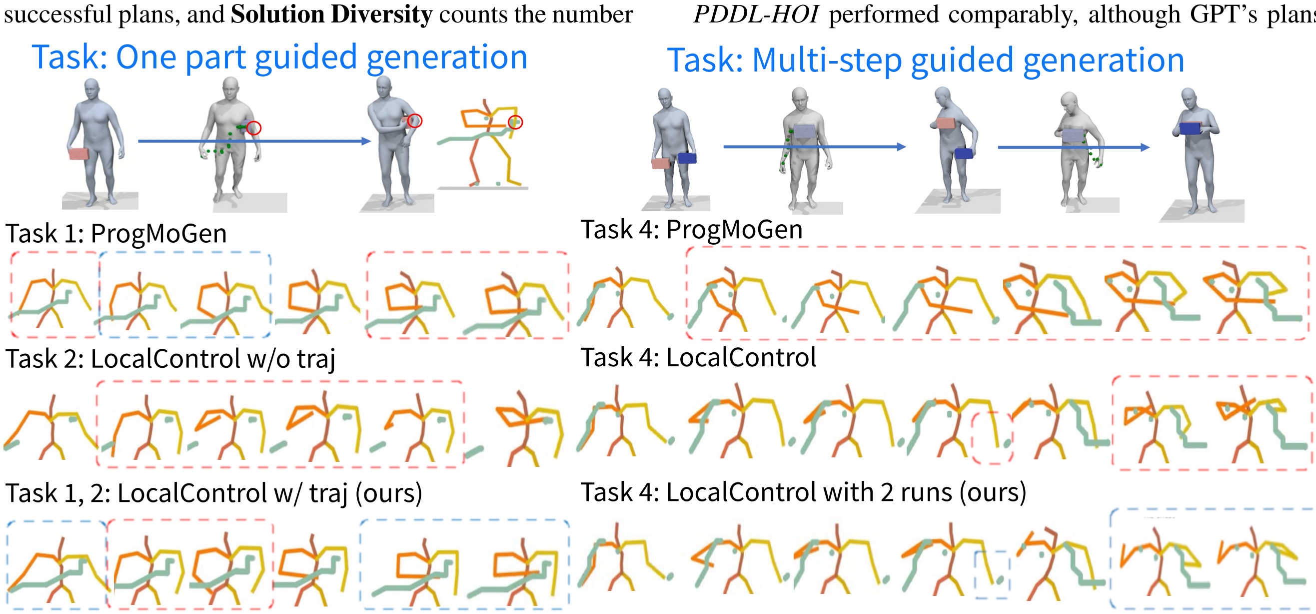
Method	Success Rate	Plan Efficiency	Solution Diversity
GPT-4o	5.0/5	**4.0**	1.2/5
GPT-4o + Primitives	5.0/5	4.2	1.8/5
GPT-4o + PDDL-HOI (ours)	5.0/5	**4.0**	**2.0/5**

Task 3: Pick up two boxes and open the door

Method	Success Rate	Plan Efficiency	Solution Diversity
GPT-4o	5.6/10	9.1	2.0/10
GPT-4o + Primitives	1.8/10	**5.0**	1.0/10
GPT-4o + PDDL-HOI (ours)	**10/10**	6.1	**2.8/10**

Task Design Three tasks include: Task 1 (one simple task), Task 2 (requiring flexibility to carry the box on the shoulder and hand for dual support), and Task 3 (requiring longer planning capability additionally).

Results Analysis In Task 1, *GPT-4o* and *GPT-4o + PDDL-HOI* performed comparably, although GPT's plans



(a) One part guided generation (b) Multi-step guided generation

Figure 4. **Evaluation of contact-guided motion generation capabilities.** We compare (a) one-part guided generation and (b) multi-step guided generation across different methods. Red/blue boxes highlight critical time frames that demonstrate our *LocalControl* method's superior performance in maintaining contact constraints and generating realistic motions.

Table 2. **Low-level motion generation performance across task configurations.** We compare *LocalControl* against baseline methods on four motion generation tasks. **C.Err.-se** denotes constraint error at start/end positions, **C.Err./g** evaluates trajectory/goal constraints. Results demonstrate the necessity of intermediate trajectory planning and multi-step generation for complex HOI motions.

Task 1: One part move with one contact trajectory guidance				
Method	Success	Max Acc.	C.Err.	Naturality
IK	6.2	0.062	0.43	6.5
ProgMoGen [30]	6.7	**0.020**	0.170	7.2
LocalControl (ours)	**7.3**	0.147	**0.079**	**8.3**

Task 2: Setting start and end of target positions for one part				
Method	Success	Max Acc.	C.Err.-se	Naturality
ProgMoGen [30] w/o Traj	2.6	**0.061**	0.274	4.4
LocalControl w/o Traj	6.6	0.146	**0.050**	4.9
LocalControl w/ Traj (ours)	**7.3**	0.147	0.079	**8.3**

Task 3: One part move and goal contact achieve				
Method	Success	Max Acc.	C.Err./g	Naturality
IK	6.3	0.077	0.136/0.097	6.5
ProgMoGen [30]	7.7	**0.021**	**0.084**/0.058	7.9
LocalControl (ours)	**8.4**	0.156	0.130/**0.045**	**8.5**

Task 4: Two-step motions				
Method	Success	Max Acc.	C.Err.	Naturality
ProgMoGen [30]	5.1	**0.023**	0.241	6.0
LocalControl x1	6.3	0.234	0.153	6.2
LocalControl x2 (ours)	**7.4**	0.198	**0.129**	**6.6**

sometimes produced redundant steps, whereas *GPT-4o + PDDL-HOI* provided clearer and more efficient plans. In Task 2, both *GPT-4o + Primitives* and *GPT-4o + PDDL-HOI* discovered additional solutions due to prior knowledge. In the more complex Task 3, *GPT-4o + Primitives* often failed due to misunderstandings of transition rules in interaction primitives, despite occasionally finding the most efficient solution (*e.g.*, 'clamp under shoulder'). *GPT-4o* generated tedious solutions involving unnecessary steps, such as placing boxes before opening the door. Our *GPT-4o + PDDL-HOI* achieved the highest success rate and diversity, benefiting from clearly defined state transition rules and diverse contact mode knowledge. More details of planning results and data statistics can be found in Appendix B.1.

4.3. Low-Level Generation Evaluation

Since there are no publicly available baselines for our designed compositional HOI tasks, we compare our method with existing guided motion generation methods that use interaction constraints but ignore specific object geometry [30, 52]. ProgMoGen [30] and an inverse kinematic method (IK) with human pose regularization [33] and temporal smoothness serve as comparison baselines.

Evaluation Metrics We use four metrics for evaluation: **Maximum Joint Acceleration** [30] measures the smoothness of joint movements; **Constraint Error [30]** assesses how well the generated motion follows the guidance constraints. The two additional metrics **Naturality** and **Success** are evaluated by humans ranging from 1.0 to 10.0 for the naturality of human motion (adherence to human kinematics) and the level of success in completing the guidance tasks respectively. **Success** considers whether the body parts move from the start to the end following the trajectory or maintain a static constrained point.

Table 3. **Performance comparison between *OmniControl* and *LocalControl* on distribution-based metrics.** We evaluate each method using its corresponding training data configuration. *LocalControl* achieves superior FID scores, particularly for dual-hand guidance tasks, demonstrating the benefits of focusing on local manipulation operations over global motion patterns.

Original HumanML3D				
Method	**Joints Guide**	FID ↓	R-precision (top-3) ↑	Diversity →
OmniControl	Pelvis	0.322	0.691	9.545
OmniControl	Left Wrist	0.304	0.680	9.436
OmniControl	Right Wrist	0.299	0.692	9.519
OmniControl	Right + Left Wrist	0.464	0.677	9.601

'No-Walk' HumanML3D				
Method (ours)	**Contact Points Guide**	FID ↓	R-precision (top-3) ↑	Diversity →
LocalControl	Chest Contact	0.263	0.603	8.859
LocalControl	Left Hand Contact	0.292	0.610	8.653
LocalControl	Right Hand Contact	0.231	0.606	8.585
LocalControl	Left + Right Hand	0.151	0.605	8.674

Experimental Results We evaluated four tasks to demonstrate the robustness of our pipeline design, illustrated in Fig. 4 and Tab. 2. Comparing *LocalControl* with ProgMoGen [30] across all tasks, we observe that while ProgMoGen achieves the best maximum acceleration (indicating smoother motion), our method outperforms in most other metrics. As shown in Fig. 4, ProgMoGen's performance is limited by the expressive power of the latent vector in its optimization process [30].

By comparing *LocalControl* with and without intermediate trajectory guidance in both quantitative and qualitative results of Task 2, we demonstrate the necessity of planning intermediate contact guidance. Without it, the intermediate motion can be random, potentially causing severe collisions between objects and humans. Finally, comparing single-run and multi-run approaches in Task 4, we find that generating the motion in two runs with the inferred intermediate key pose leads to more accurate and natural results, highlighting the importance of key pose inference to reduce error accumulation over long sequences.

Model Comparison Analysis To evaluate the performance of LocalControl compared with the original OmniControl [52], we provide results of FID, R-precision, and Diversity using different training data versions (Tab. 3). For the 'No-Walk' HumanML3D, we disable the root's translation and rotation variations. LocalControl's FID out-

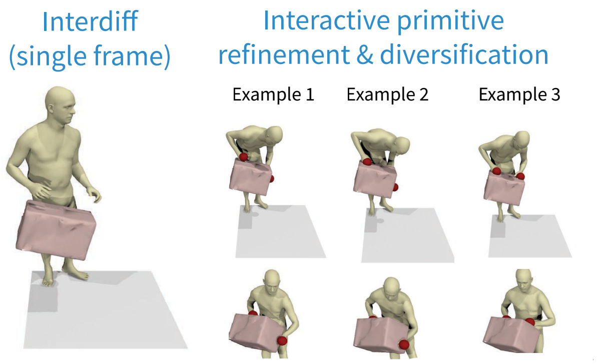


Figure 5. **Interactive primitive refinement and diversification.** Starting from a single frame generated by InterDiff [53], our interaction primitive model produces multiple refined solutions that exhibit improved physical realism and increased diversity. Each example demonstrates different plausible ways to complete the HOIs while maintaining contact constraints.

performs OmniControl (especially for dual-hand guidance) since there is less variation in the 'No-Walk' HumanML3D, allowing focus on learning local operations. For evaluating out-of-distribution motions such as multi-object interactions, distribution-based FID becomes unreliable for naturalness assessment, leading us to prioritize human evaluation for our multi-object cases. We include details of human evaluation and data statistics in Appendix B.2.

4.4. Ablations

Qualitative Results of Different Components To illustrate the generalization capabilities of our method, we present a complete motion sequence for the novel task "Pick up two boxes and open the door" in Fig. 6. Qualitative results for primitive contact generation and key pose generation are provided in Figs. 1 and 3 respectively. Finally, we demonstrate the benefits of refining poses using our learned local interaction model—interaction primitives. As shown in Fig. 5, applying our generative interaction primitive model to outputs from InterDiff [53] enhances physical realism and diversifies contact poses. In Appendix D.1, we present additional qualitative results, including two extra plans and generated motions for other objects.

Additional Ablations We conducted ablations on the interaction primitive model to evaluate the sampling procedure and normalization modifications, as detailed in Appendix C.1. Additionally, since the post-optimization step involves multiple terms, we provide a qualitative ablation study in Appendix C.2 to assess the effect of each term.

5. Conclusion

We presented **PrimHOI**, a novel framework for synthesizing complex daily-life HOI motions through symbolic planning and generalizable interaction primitives. By decomposing HOI generation into reusable submodules, our method demonstrates that symbolic planning can complement data-driven approaches to achieve systematic generalization across different spatial configurations, diverse objects, and temporal compositions. While this modular design enables zero-shot transfer to out-of-distribution multi-object scenarios, it also introduces challenges in recomposing submodules into seamless motion due to the separation of interdependent variables.

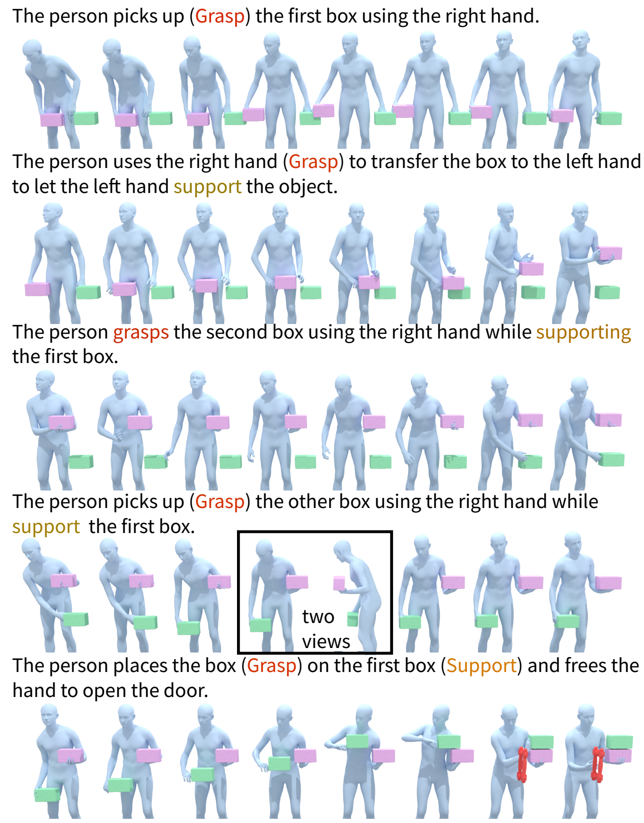


Figure 6. **Synthesized motion sequence for the "pick up two boxes and open door" task.** **PrimHOI** generates a complete motion sequence that demonstrates coordinated use of interaction primitives throughout the task execution. Highlighted text annotations indicate the specific interaction primitives (**Grasp** and **Support**) being employed at each step, showing how **PrimHOI** seamlessly transitions between different contact states to accomplish the complex multi-object manipulation task.

Capabilities and Limitations Our framework's flexible temporal and spatial composition enables strong generalization despite using only four interaction primitives (Fig. 2). Adding new primitives is straightforward, as demonstrated in Appendix A.4, which also discusses motion diversity. However, the inherent decomposition can introduce failures when interdependent variables are separated (Appendix D.3), and individual submodules have limitations that affect motion quality (Appendix D.2). We discuss potential improvements in Appendix E.

Acknowledgments This work is supported in part by the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, and the National Comprehensive Experimental Base for Governance of Intelligent Society.

References

[1] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. Pddl— the planning domain definition language. *Technical Report, Tech. Rep.*, 1998. 4

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, 2019. 2

[3] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *International Conference on 3D Vision (3DV)*, 2022. 2, 3

[4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Sinc: Spatial composition of 3d human motions for simultaneous action generation. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[5] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2023. 2

[6] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018. 3

[7] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, A1

[8] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 2

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[10] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[11] Jieming Cui, Tengyu Liu, Ziyu Meng, Jiale Yu, Ran Song, Wei Zhang, Yixin Zhu, and Siyuan Huang. Grove: A generalized reward for learning open-vocabulary physical skill. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[12] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[13] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, 4(37), 2019. 2

[14] Caelan Reed Garrett, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddlstream: Integrating symbolic planners and blackbox samplers via optimistic adaptive planning. In *Proceedings of the international conference on automated planning and scheduling*, 2020. 2, 4, 5

[15] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, 2023. 2

[16] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[17] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. *arXiv preprint arXiv:2403.13900*, 2024. 2, 3

[18] Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. Mewl: Few-shot multimodal word learning with referential uncertainty. In *Proceedings of International Conference on Machine Learning (ICML)*, 2023. 2

[19] Nan Jiang, Zimo He, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia Conference Papers*, 2024. 2

[20] Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *CVPR*, 2025. 2

[21] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2

[22] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[23] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 2

[24] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Multiact: Long-term 3d human motion generation from multiple action labels. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 3

[25] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2, 6, A1

[26] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[27] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[28] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2

[29] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 4

[30] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 7

[31] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 3

[32] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters (RA-L)*, 7(1):470–477, 2021. 2

[33] Junzhe Lu, Jing Lin, Hongkun Dou, Yulun Zhang, Yue Deng, and Haoqian Wang. Dposer: Diffusion model as robust 3d human pose prior. *arXiv preprint arXiv:2312.05541*, 2023. 2, 5, 6, 7, A1, A2, A3, A7

[34] OpenAI. ChatGPT. `https://chat.openai.com/`, 2023. 4

[35] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 3, 5

[36] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2

[37] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[38] Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. Modular deep learning. *arXiv preprint arXiv:2302.11529*, 2023. 3

[39] Ananya Rastogi. Learning about few-shot concept learning. *Nature Computational Science*, 2(11):698, 2022. 2

[40] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[41] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39 (4):54–1, 2020. 2

[42] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[43] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023. 2

[44] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 2

[45] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 35:14959–14971, 2022. 2

[46] Zhenzhi Wang, Jingbo Wang, Dahua Lin, and Bo Dai. Intercontrol: Generate human motion interactions by controlling every joint. *arXiv preprint arXiv:2311.15864*, 2023. 2

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022. 3

[48] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. *arXiv preprint arXiv:2403.11208*, 2024. 2

[49] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. *arXiv preprint arXiv:2406.17840*, 2024. 3

[50] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 2, 3

[51] Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Halma: Humanlike abstraction learning meets affordance in rapid problem solving. In *ICLR Workshop on Generalization beyond the training distribution in brains and machines*, 2021. 2

[52] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 2, 5, 6, 7, A1

[53] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2, 5, 8

[54] Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 37:52858–52890, 2024. 2, 3, 4, 5

[55] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 2

[57] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[58] Zihang Zhao, Yuyang Li, Wanlin Li, Zhenghao Qi, Lecheng Ruan, Yixin Zhu, and Kaspar Althoefer. Tac-man: Tactile-informed prior-free manipulation of articulated objects. *IEEE Transactions on Robotics (T-RO)*, 41:538–557, 2024. 2

[59] Zihang Zhao, Wanlin Li, Yuyang Li, Tengyu Liu, Boren Li, Meng Wang, Kai Du, Hangxin Liu, Yixin Zhu, Qining Wang, et al. Embedding high-resolution touch across robotic hands enables adaptive human-like grasping. *Nature Machine Intelligence*, 7(6), 2025. 2

[60] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2

`PrimHOI`: Compositional Human-Object Interaction via Reusable Primitives
Supplementary Material

A. Method and Implementation Details

A.1. High-Level Planning Details

Our PDDL-HOI planning domain consists of two files: `domain.pddl`, which specifies predicates and actions, and `stream.pddl`, which defines sampling streams for manipulating parts. Fig. A1 illustrates the construction of our planning framework.

Domain Definition In `domain.pddl`, we use *predicates* to describe static facts and dynamic states. Static predicates like `(IsObject box1)` represent unchanging truths, while dynamic predicates such as `(Grasped box1 righthand)` describe evolving states. The derivative predicates can be inferred from simple predicates: for example, `(Holden?o)` holds when the object `?o` is held by any interaction primitive.

Actions define state transitions through preconditions and effects. Fig. A1 shows the action `GraspToClamp`, which transitions an object to a clamped state only when preconditions are satisfied (*e.g.*, clamping parts are empty and the object is already grasped).

Stream Sampling The `stream.pddl` file declares sampling functions implemented elsewhere in the codebase. Streams enable dynamic sampling of manipulation parts by generating available predicates during planning rather than initially providing them.

A.2. Low-Level Generation Details

Primitive Contact Model Adapted from OMOMO [25], our contact generation uses a conditional diffusion model:

$$P(\{p_i^o\}^P|\mathbf{V_o}) = Q(\{p_i^o\}^P|\{p_i^{*,o}\}^P, \mathbf{V_o}), \tag{A1}$$

where Q represents the denoising process, $\{p_i^{*,o}\}^P$ are initial noisy contact points, and outputs $\{p_i^o\}^P$ are relative positions to the object center.

We collected interaction data from multiple sources: 378 video sequences from OMOMO for *Clamp* contacts (boxes, suitcases, monitors, trashcans, plastic containers) and 937 frames from BEHAVE [7] for *Grasp* data (boxes, trashbins, yoga mats, keyboards). For *Support* and *Dual Support*, we employ analytical functions that generate physically valid contact points with random rotational deviations up to 30° from horizontal.

Object scale normalization using the oriented bounding box radius significantly improves generalization across shape and pose variations.

Key Pose Generation Details The generation of key poses involves three sequential steps as shown in Fig. A2: generation of interaction primitive, placement of objects, and optimization of body poses.

The *object placement prior* $P(s_o \mid \{p_i^h\} = \{p_i^o\}^{P_i})$ uses Mixture of Gaussians distributions computed from BEHAVE clusters, positioning objects where interactions commonly occur relative to the human body. When multiple primitives are involved, placement follows priority order: Clamp/Support/Dual-Support $>$ Grasp.

Body pose optimization aligns contact points using DPoser [33] while maintaining pose plausibility. This system provides flexibility through valid placement and diversity through data clustering.

LocalControl Implementation Since OmniControl [52] performs poorly for stationary body movements with active limb manipulation, we retrained it focusing on local operations, creating *LocalControl*. For walking tasks, we retain the original OmniControl model.

During inference, we add static control signals to the feet to maintain body stability. Due to potential misalignment between guidance and generated positions in final frames, we employ inverse kinematics for "last mile" operations where collisions occur frequently (Fig. A9).

A.3. Optimization in Key Pose Generation and Post-Refinement

The post-optimization process maintains interaction primitive constraints while minimizing collisions and penetrations. The optimization objective comprises six complementary terms, with key pose generation using single-frame versions of these temporal formulations.

Contact Loss We minimize the Geman-McClure error function ρ (robust to outliers) between body and object contact points:

$$E_{\text{contact}} = \sum_{t=0}^{T-1} \sum_{P_i} \rho(\boldsymbol{p}_i^h - \boldsymbol{p}_i^o)^{P_i}, \tag{A2}$$

where P_i represents interaction primitives maintained during motion.

Normal Loss For *Support* and *Dual Support* primitives, we minimize the cosine distance between human and object surface normals:

$$E_{\text{normal}} = \sum_{t=0}^{T-1} \sum_{P_i} \text{cosine}(\boldsymbol{n}_i^h, \boldsymbol{n}_i^o)^{P_i}, \tag{A3}$$

where $\boldsymbol{n}_i^h$ and $\boldsymbol{n}_i^o$ are outward human and inward object surface normals, respectively.

domain.pddl

```
(define (domain object-manipulation)
(:predicates
;; Type Declarations
(IsPart ?p)
(IsObject ?o)
 ....
;; States
(Grasped ?o ?bp)
(Support?o ?bp)
(Clamped ?o ?bp1 ?bp2)
(DualSupport ?o ?bp1 ?bp2)
(IsReached ?o)
(Empty ?bp)
(Holden ?o)
(IsSupporten ?o)
(IsDualSupport ?o)
(IsGrasped ?o)
(IsClamped ?o)
;; Derived predicate
(:derived (Holden ?o)
  (or
    (IsGrasped ?o)
    (IsSupporten ?o)
    (IsDualSupport ?o)
    (IsClamped ?o))
)
....
```

```
(:action GraspToClamp
  :parameters (?o ?bp ?bp1 ?bp2)
  :precondition (and
    (IsObject ?o)
    (IsPart ?bp)
    (IsPart ?bp1)
    (IsPart ?bp2)
    (not (SameSide ?bp ?bp1))
    (not (SameSide ?bp ?bp2))
    (Grasped ?o ?bp)
    (not (IsClamped ?o)) ;; no multi-clamp
    (CanClamp ?bp1 ?bp2)
    (or
      (Empty ?bp1)
      (IsObjTool ?bp1) ;; allow use object
    )
    (not (= ?o ?bp1))
    (Empty ?bp2)

  )
  :effect (and
    (Clamped ?o ?bp1 ?bp2)
    (not (Empty ?bp1))
    (not (Empty ?bp2))
    (IsClose ?bp ?bp1)
    (increase (total-cost) 1)))
....
```

stream.pddl

```
(define (stream object-manipulation)

  ;; Stream to generate parts that can grasp and are empty
  (:stream sample-grasp-part
    :outputs (?bp)
    :certified (and (CanGrasp ?bp) (IsPart ?bp))
  )

  ;; Stream to generate parts that can pan-hold
  (:stream sample-support-part
    :outputs (?bp)
    :certified (and (CanSupport ?bp) (IsPart ?bp))
  )

  ;; Stream to generate pairs of parts that can clamp
  (:stream sample-clamp-parts
    :outputs (?bp1 ?bp2)
    :certified (and (CanClamp ?bp1 ?bp2) (IsPart ?bp1) (IsPart ?bp2))
  )

   ;; Stream to generate pairs of parts that can clamp
  (:stream sample-dualsupport-parts
    :outputs (?bp1 ?bp2)
    :certified (and (CanDualSupport ?bp1 ?bp2) (IsPart ?bp1) (IsPart ?bp2))
  )
 ....
)
```

Figure A1. **PDDL-HOI consists of two complementary files.** The `domain.pddl` file defines predicates, actions, and state transitions, while `stream.pddl` specifies sampling functions for dynamic body part selection. The example shows the `GraspToClamp` action, which transitions objects from grasped to clamped states when preconditions are met. This modular design enables flexible primitive combinations during planning.

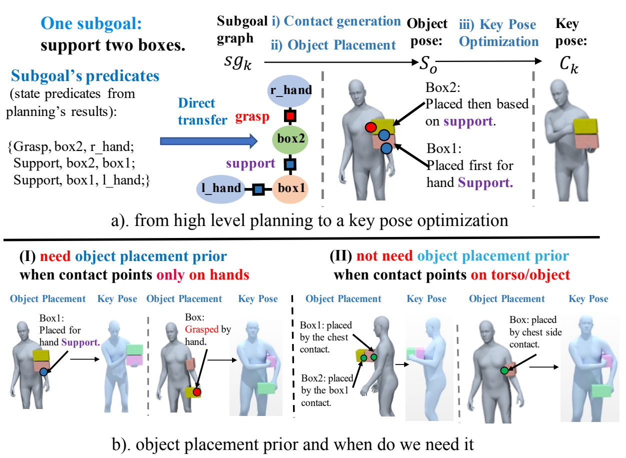


Figure A2. **Key pose generation follows a three-stage pipeline.** Starting from planned subgoal sg_k, we first generate primitive contact points (red and blue dots), then position objects using learned interaction location priors, and finally optimize body pose to satisfy contact constraints. Hand-only interactions bypass multi-primitive coordination by relying solely on object placement priors.

Collision Penalties Object collision loss prevents body-object interpenetration using signed distance fields:

$$E_{\text{colli}} = \sum_{t=0}^{T-1} \sum_{o \in O} \min(\mathbf{sdf}_o(\boldsymbol{v}_h), 0), \tag{A4}$$

where $\boldsymbol{v}_h$ represents the vertices of the human body.

Self-penetration loss prevents limb-torso intersections:

$$E_{\text{pene}} = \sum_{t=0}^{T-1} \sum_{l_i} \min(\mathbf{sdf}_{\text{torso}}(\boldsymbol{v}^{l_i}), 0), \tag{A5}$$

where l_i denotes the limbs and $\boldsymbol{v}^{l_i}$ represents the vertices of the arm.

Temporal and Prior Regularization Temporal smoothness is enforced through vertex consistency between adjacent frames:

$$E_{\text{temporal}} = \sum_{t=0}^{T-1} \rho(\boldsymbol{v}_{t+1}^h - \boldsymbol{v}_t^h). \tag{A6}$$

Body pose regularization employs DPoser [33] diffusion-based loss:

$$E_{\text{prior}} = \sum_{t=0}^{T-1} L_{\text{DPoser}}(\Theta_t), \tag{A7}$$

where Θ_t represents body pose parameters in frame t.

A.4. Task Range and Extension Capabilities

Generalization Scope Despite using only four interaction primitives, **PrimHOI** demonstrates a great generalization in diverse HOI tasks. Any task within the planning scope of these primitives can be successfully synthesized. Once an object type is learned within a primitive, our

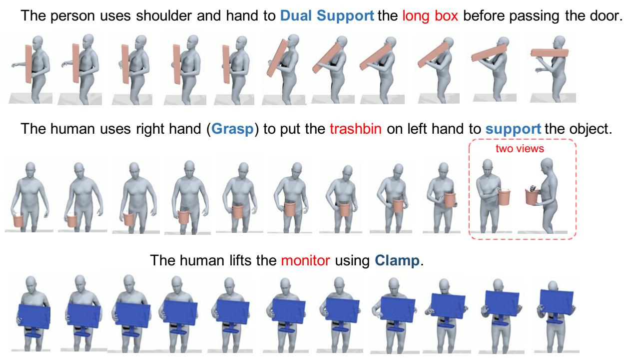


Figure A3. **Our framework demonstrates robust generalization across diverse object categories beyond training distributions.** Generated HOI motions span various object types not encountered during training, validating **`PrimHOI`**'s ability to transfer learned interaction patterns to novel geometric and functional contexts. The alternative view of the second motion reveals accurate contact normal computation for the Support primitive, confirming that **`PrimHOI`** maintains precise surface alignment even when generalizing to unseen object shapes and interaction scenarios.

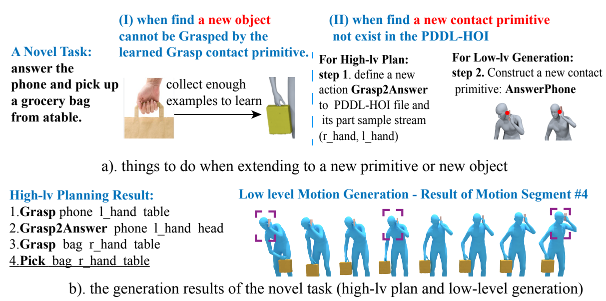


Figure A4. **Our framework enables systematic extension to novel interaction scenarios through structured primitive integration.** The three-step extension methodology facilitates incorporation of new tasks requiring previously undefined interaction primitives, while complete planning and generation results for the "pickup grocery bag while answering phone" task illustrate successful execution of the fourth planned action. This systematic extensibility demonstrates our framework's capacity to accommodate previously unseen interaction scenarios without requiring fundamental architectural modifications, establishing a scalable foundation for expanding human-object interaction capabilities.

method generates planning sequences for interactions with that object (Fig. A3).

Temporal composition enables sequences like "pick first, then place" or "clamp first, then place." In contrast, spatial composition allows flexible body part and object combinations for *Clamp*, *Support*, and *Dual Support* interactions.

Extension to Unseen Tasks Extending **`PrimHOI`** to new tasks like "picking up a grocery bag while answering a phone" follows a straightforward process (Fig. A4):

1. **Domain Update**: Add new action definitions and streams to PDDL-HOI (required only for new primitives, not new objects).
2. **Primitive Training**: Train new interaction primitives using example interactions and learn object placement priors.
3. **Pipeline Execution**: Generate task plans and low-level motion sequences.

For the grocery bag example, we manually select the upper grasp points for the grocery bag and introduce the `Grasp2Answer` action for the new *AnswerPhone* primitive, defining phone-ear contact configurations.

Motion Diversity Motion diversity arises from multiple sources: (1) variability in generated interaction primitive, (2) Gaussian mixture sampling for object placement, and (3) stochastic diffusion-guided human motion. The supplementary video demonstrates the diversity in object placements and primitive contact variations in different scenarios.

B. Experiment Details

Our evaluation employs five human raters in all tasks. For high-level planning, the raters collaboratively discuss and reach consensus on task success, step efficiency, and plan diversity using objective reasoning (Fig. A10). For low-level evaluation, each rater independently scores task success and motion naturalness (1.0-10.0 scale) using paired comparison interfaces with three shuffled examples per sheet.

B.1. High-Level Planning Evaluation Details

Fig. A10 presents detailed prompts and planning results for the three methods evaluated in all tasks. Statistical analysis with T-tests that compare other methods with ours is provided in Tab. A1. Each task was evaluated through multiple runs: Tasks 1 and 2 (5 trials each), Task 3 (10 trials), with five total runs per result. Failure cases were excluded from cost calculations.

Our GPT-4o + PDDL-HOI method shows superior performance in success rate and solution diversity while maintaining competitive plan efficiency at all complexity levels. In particular, in the complex Task 3, our method achieved a success rate of 100% compared to GPT-4o 56% and GPT-4o + Primitives 18%.

B.2. Low-Level Evaluation Details

The low-level evaluation uses guided intermediate motions from the three plans shown in the main paper. Although limited in number, these motion segments distinguish sufficiently between methods through quantitative and qualitative analysis.

Baseline Implementations **Inverse Kinematics (IK):** Employs DPoser [33] body pose prior with Contact Loss (Eq. (A2)) and Temporal Loss (Eq. (A10)).

Table A1. **Statistical comparison of high-level planning methods across three tasks.** Each task was evaluated over five runs with varying trial counts (Tasks 1-2: 5 trials per run; Task 3: 10 trials per run). T-tests compare baseline methods against our GPT-4o + PDDL-HOI approach, excluding failure cases from efficiency calculations. Our method achieves statistically significant improvements in success rate and solution diversity.

Task 1: Pick up two boxes from table			
Method	**Success Rate**	**T-statistic**	**P-value**
GPT-4o	3/5, 4/5, 4/5, 4/5, 5/5	-3.16	1.33e-2
GPT-4o + PDDL-HOI (ours)	5/5, 5/5, 5/5, 5/5, 5/5	–	–
Method	**Plan Efficiency**	**T-statistic**	**P-value**
GPT-4o	5.3, 5.5, 5.8, 5.8, 6.0	7.73	5.59e-05
GPT-4o + PDDL-HOI (ours)	4.4, 4.4, 4.6, 4.6, 4.8	–	–
Method	**Solution Diversity**	**T-statistic**	**P-value**
GPT-4o	1, 1, 2, 2, 2	-1.63	1.41e-1
GPT-4o + PDDL-HOI (ours)	2, 2, 2, 2, 2	–	–
Task 2: Carry long box passing the door			
Method	**Success Rate**	**T-statistic**	**P-value**
GPT-4o	5/5, 5/5, 5/5, 5/5, 5/5	–	–
GPT-4o + Primitives	5/5, 5/5, 5/5, 5/5, 5/5	–	–
GPT-4o + PDDL-HOI (ours)	5/5, 5/5, 5/5, 5/5, 5/5	–	–
Method	**Plan Efficiency**	**T-statistic**	**P-value**
GPT-4o	4.0, 4.0, 4.0, 4.0, 4.0	–	–
GPT-4o + Primitives	4.0, 4.2, 4.2, 4.2, 4.4	3.16	1.33e-2
GPT-4o + PDDL-HOI (ours)	4.0, 4.0, 4.0, 4.0, 4.0	–	–
Method	**Solution Diversity**	**T-statistic**	**P-value**
GPT-4o	1, 1, 1, 1, 2	-4.0	3.95e-3
GPT-4o + Primitives	1, 2, 2, 2, 2	-1.00	3.47e-1
GPT-4o + PDDL-HOI (ours)	2, 2, 2, 2, 2	–	–
Task 3: Pick up two boxes and open the door			
Method	**Success Rate**	**T-statistic**	**P-value**
GPT-4o	5/10, 5/10, 5/10, 6/10, 7/10	-11.0	4.15e-6
GPT-4o + Primitives	1/10, 2/10, 2/10, 2/10, 2/10	41.0	1.38e-10
GPT-4o + PDDL-HOI (ours)	10/10, 10/10, 10/10, 10/10, 10/10	–	–
Method	**Plan Efficiency**	**T-statistic**	**P-value**
GPT-4o	8.8, 8.8, 9.0, 9.3, 9.6	11.83	2.39e-6
GPT-4o + Primitives	5.0, 5.0, 5.0, 5.0, 5.0	-5.80	4.04e-4
GPT-4o + PDDL-HOI (ours)	5.4, 6.1, 6.3, 6.4, 6.5	–	–
Method	**Solution Diversity**	**T-statistic**	**P-value**
GPT-4o	2, 2, 2, 2, 2	-4.00	3.95e-3
GPT-4o + Primitives	1, 1, 1, 1, 1	-9.0	1.85e-5
GPT-4o + PDDL-HOI (ours)	2, 3, 3, 3, 3	–	–

ProgMoGen: For non-walking tasks, uses "stands" motion prompts with foot constraints to prevent locomotion.

Statistical results with T-test analysis are detailed in Tab. A2, demonstrating LocalControl's superior performance across most metrics, while ProgMoGen achieves better motion smoothness (lower maximum acceleration).

B.3. Additional Evaluation Metrics

We introduce F-best, the measurement frequency of selection, as the best method among candidates. Five participants selected the best from three examples in four tasks. The results in Tab. A3 show that our method was chosen as the best in 17 of 20 choices, confirming the superiority in the evaluation of human preferences.

Table A2. **Human evaluation demonstrates LocalControl's superior performance in low-level motion generation.** Five raters scored task success and motion naturalness (1-10 scale) across four motion synthesis tasks. T-tests compare baseline methods against our LocalControl approach, showing statistically significant improvements in both metrics across most tasks.

Task 1: One part move with contact trajectory guidance			
Method	**Success Score**	**T-statistic**	**P-value**
IK	6.0, 6.0, 6.5, 6.7, 6.0	-3.26	1.16e-2
ProgMoGen	7.0, 6.0, 6.4, 6.6, 6.7	-3.13	1.40e-2
LocalControl (ours)	8.0, 7.0, 7.4, 7.0, 7.2	–	–
Method	**Naturalness Score**	**T-statistic**	**P-value**
IK	6.0, 7.0, 6.3, 6.7, 6.5	-7.01	1.11e-4
ProgMoGen	8.0, 7.0, 6.7, 7.5, 7.2	-3.43	8.90e-3
LocalControl (ours)	8.0, 9.0, 8.3, 8.0, 8.1	–	–
Task 2: Start and end position targeting			
Method	**Success Score**	**T-statistic**	**P-value**
ProgMoGen w/o Trajectory	2.0, 3.0, 2.4, 3.3, 2.5	-15.87	2.49e-7
LocalControl w/o Trajectory	7.0, 6.0, 7.0, 6.5, 6.3	-2.81	2.27e-2
LocalControl w/ Trajectory (ours)	8.0, 7.0, 7.4, 7.0, 7.2	–	–
Method	**Naturalness Score**	**T-statistic**	**P-value**
ProgMoGen w/o Trajectory	4.0, 5.0, 5.2, 4.5, 3.5	-10.49	5.93e-6
LocalControl w/o Trajectory	6.0, 4.0, 5.3, 4.8, 4.5	-8.60	2.59e-5
LocalControl w/ Trajectory (ours)	8.0, 9.0, 8.3, 8.0, 8.1	–	–
Task 3: One part move with goal contact achievement			
Method	**Success Score**	**T-statistic**	**P-value**
IK	6.3, 6.0, 7.0, 6.4, 6.0	-7.29	2.63e-5
ProgMoGen	8.0, 8.5, 7.4, 7.0, 7.8	-1.45	1.80e-1
LocalControl (ours)	8.4, 9.0, 8.0, 7.8, 7.6	–	–
Method	**Naturalness Score**	**T-statistic**	**P-value**
IK	6.0, 7.0, 6.3, 6.6, 6.5	-8.29	3.38e-5
ProgMoGen	8.0, 8.0, 7.4, 7.5, 8.5	-2.09	7.01e-2
LocalControl (ours)	8.0, 8.9, 8.6, 8.5, 8.1	–	–
Task 4: Two-step sequential motions			
Method	**Success Score**	**T-statistic**	**P-value**
ProgMoGen	5.0, 5.3, 5.4, 5.1, 4.8	-11.06	3.98e-6
LocalControl x1	7.0, 6.0, 6.4, 6.3, 6.0	-4.28	2.70e-3
LocalControl x2 (ours)	8.0, 7.0, 7.5, 7.1, 7.6	–	–
Method	**Naturalness Score**	**T-statistic**	**P-value**
ProgMoGen	6.0, 5.7, 6.2, 5.9, 6.0	-2.82	3.14e-2
LocalControl x1	6.0, 5.6, 7.0, 6.5, 6.1	8.28e-1	4.34e-1
LocalControl x2 (ours)	6.0, 7.0, 6.6, 6.5, 6.3	–	–

C. Ablations

C.1. Contact Primitive Model Ablation

We evaluated different configurations by comparing denoising steps (100, 200, 1000) and object scale normalization (Tab. A4). Evaluation uses **Clamp Success** and **Grasp Success** rates assessed by human evaluators based on physical stability in four types of objects: box, monitor, plastic container, and trashcan.

Key findings:

- **Grasp model:** Normalization does not improve perfor-

Table A3. **Human preference evaluation confirms `PrimHOI`' superiority.** F-best measures how frequently each method was selected as the best among candidates by five evaluators across four tasks. Our LocalControl variants achieve 17 out of 20 best selections, demonstrating clear human preference for **`PrimHOI`**.

Task 1 Methods	F-best ↑	Task 2 Methods	F-best ↑
IK	0	ProgMoGen w/o Trajectory	0
ProgMoGen	0	LocalControl w/o Trajectory	5
LocalControl (ours)	**5**	LocalControl w/ Trajectory	–
Task 3 Methods	**F-best ↑**	**Task 4 Methods**	**F-best ↑**
IK	0	ProgMoGen	0
ProgMoGen	1	LocalControl x1	2
LocalControl (ours)	**4**	LocalControl x2 (ours)	**3**

Table A4. **Ablation study reveals optimal contact primitive model configuration.** We compare different denoising step counts and normalization strategies on Clamp and Grasp primitive success rates across multiple object types. The 200-step configuration with normalization provides the best efficiency-accuracy trade-off, achieving 92% success for Clamp while maintaining reasonable Grasp performance.

Configuration	Clamp Success	Grasp Success
1000 steps w/o normalization	0.46	**0.79**
100 steps w/o normalization	–	0.61
200 steps w/o normalization (our Grasp)	0.54	**0.81**
1000 steps w/ normalization	**0.93**	–
100 steps w/ normalization	0.77	–
200 steps w/ normalization (our Clamp)	**0.92**	0.57

mance, possibly disrupting fixed hand-to-wrist distance constraints.

- **Clamp model:** Normalization significantly improves success rate (from 0.54 to 0.92).
- **Denoising steps:** 200 and 1000 steps perform well; 100 steps show deterioration.

Based on these results, we selected 200 denoising steps with normalization to achieve an optimal efficiency-accuracy balance for Clamp and without normalization for Grasp (Fig. A5).

C.2. Post-Optimization Terms Ablation

We evaluated four specific loss terms beyond the essential contact and prior losses (Fig. A6):

- **Normal Loss:** Improves Support contact quality:

$$E_{\text{normal}} = \sum_{t=0}^{T-1} \sum_{P_i} cosine(\boldsymbol{n}_i^h - \boldsymbol{n}_i^o)^{P_i}. \quad \text{(A8)}$$

- **Self-Penetration & Object Collision Loss:** Minimize human-object penetrations using SDF metrics:

$$E_{\text{pene}} = \sum_{t=0}^{T-1} \sum_{p_i} \min(\mathbf{sdf}_{body}(\boldsymbol{v}^{p_i}), 0). \quad \text{(A9)}$$

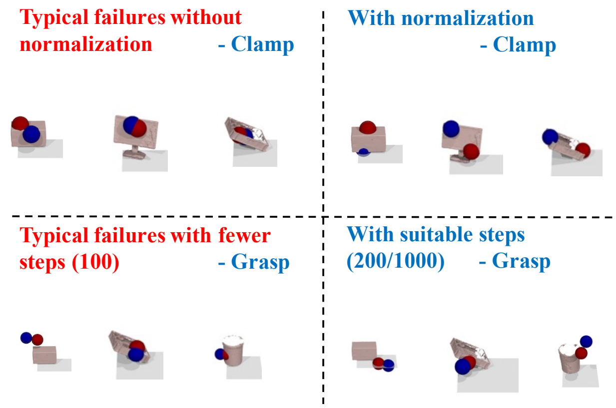


Figure A5. **Normalization enables better cross-dataset generalization for contact primitive models.** The Clamp model shows dramatically improved success when normalization is applied during cross-dataset evaluation (BEHAVE objects after OMOMO training). Similarly, increased denoising steps benefit Grasp primitive generation, with 200 and 1000 steps substantially outperforming 100 steps. Red and blue texts indicate successful and failed contact generation, respectively.

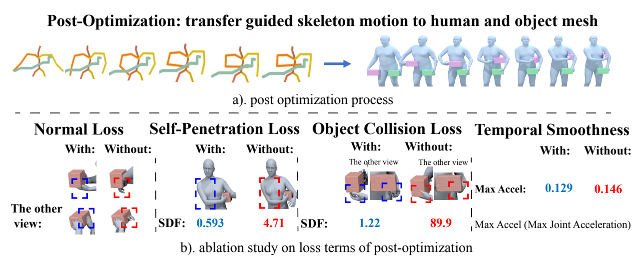


Figure A6. **Individual optimization terms address distinct motion quality challenges.** Contact guidance trajectories (green) demonstrate the post-optimization process, while ablation results reveal each term's specific contribution. Self-Penetration and Object Collision losses leverage SDF evaluation to eliminate body-body and human-object intersections, respectively. The Object Collision example shows successful prevention of hand-object collision before Support contact establishment, illustrating how each term targets essential aspects of realistic HOI generation.

where p_i denotes the part to avoid collision, either one object or one body part, and $\boldsymbol{v}^{p_i}$ represents the vertices of the part.

- **Temporal Loss:** Enhances motion smoothness across three motion sequences:

$$E_{\text{temporal}} = \sum_{t=0}^{T-1} \rho(\boldsymbol{v}_{t+1}^h - \boldsymbol{v}_t^h), \quad \text{(A10)}$$

Qualitative examples demonstrate each term's effectiveness in addressing specific motion quality issues, with SDF-based evaluation confirming reduced penetration artifacts.

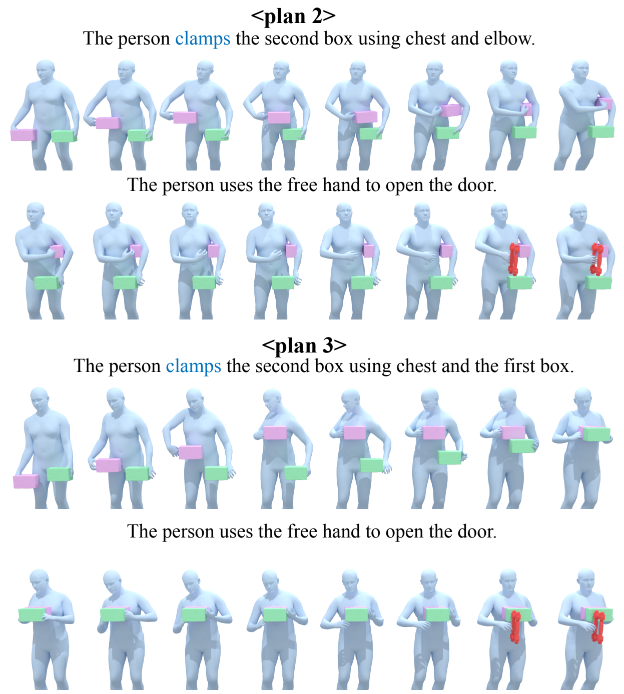


Figure A7. **Solution diversity emerges naturally from our structured planning framework.** Two alternative solutions for the “pick up two boxes and open the door” task demonstrate `PrimHOI`’s capability to generate multiple valid planning strategies for identical high-level objectives. Each solution employs distinct primitive combinations and sequencing approaches, illustrating how our PDDL-HOI framework enables flexible strategy exploration while maintaining task completion guarantees.

D. Qualitative Results and Failure Analysis

D.1. Additional Qualitative Results

Beyond the solution presented in the main paper, Fig. A7 shows two additional solutions for the task “pick up two boxes and open the door,” demonstrating `PrimHOI`’s planning diversity. Fig. A3 presents generated HOI motions for various objects, illustrating generalization capabilities in different object categories.

The supplementary video further demonstrates motion diversity arising from variations in object placement and diverse generated interaction primitives. These examples highlight the compositional flexibility achieved through our interaction primitive framework.

D.2. Failure Analysis

We identify three primary failure modes in our method (Fig. A8):

Penetration During Key Pose Generation Despite collision loss penalties, joint optimization with contact constraints can still produce body-object penetrations (Fig. A8(i)). This occurs when contact constraints override collision avoidance, requiring stronger pose priors, emphasizing collision-free configurations.

Incorrect Grasping Poses Relying solely on contact points for grasp constraints occasionally produces unrealistic grasps (Fig. A8(ii)). Although normal loss could improve accuracy, problematic edge normals on objects complicate this approach. A more sophisticated grasping pose model that incorporates geometric reasoning would address this limitation.

Interpolation Collisions Post-optimization of only keyframes followed by linear interpolation, can cause intermediate collisions with objects (Fig. A8(iii)). This occurs because the interpolation ignores the position of objects during transitions. Local motion models with collision avoidance or complete sequence optimization could mitigate this problem.

Table A5. **High-level planning components achieve perfect reliability across complex tasks.** Individual component evaluation using the “pick two boxes and open door” task over 5 runs demonstrates that both goal constraint translation and PDDL planning maintain consistent performance, establishing a robust foundation for the overall pipeline.

High-Level Steps	Goal Constraints Translation (GPT-4o)	PDDL Planning (PDDL-HOI)
Success Rate	5/5	100%

D.3. Multi-Stage Pipeline Failures

Our modular design enables zero-shot generalization but introduces potential failures by separating interdependent variables. However, this structure facilitates the detection of isolated failures and targeted corrections.

The primary issue involves the contradictions between high-level plans and detailed human-object layouts (Fig. A8(iv)). This can be resolved by identifying and resampling plans based on large SDF penalty terms during key pose generation or post-optimization.

Success Rate Analysis Tabs. A5 and A6 provide detailed step-wise success rates for the “pick two boxes and open door” task. High-level planning achieves 100% success in goal constraint translation (GPT-4o) and PDDL planning (semantic validity guaranteed).

The low-level generation shows an overall success rate of 88.4%, with individual components performing as follows:

- Primitive contact generation: 92% (Clamp), 81% (Grasp)
- Key pose generation: 96.8% (92/95)
- Object motion planning: 100% (92/92)
- Contact-guided motion: 100% (92/92)
- Post-optimization: 91.3% (84/92)

Some failures result from optimization randomness, which multiple sampling attempts and improved pose priors could mitigate.

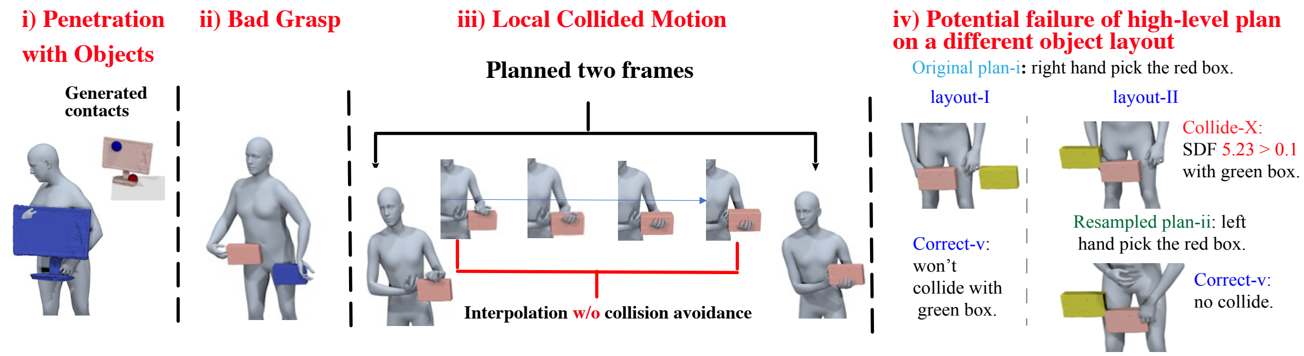


Figure A8. **Systematic failure analysis identifies four distinct limitation categories in our pipeline.** Body-object penetration occurs during key pose generation despite collision loss constraints, while incorrect grasp poses result from contact-point-only optimization without full hand orientation consideration. Interpolation-induced collisions emerge between optimized keyframes, and high-level plan contradictions arise when detailed human-object spatial layouts conflict with abstract planning assumptions. Each failure mode provides targeted directions for addressing specific pipeline limitations in future development.

Table A6. **Component-wise analysis reveals robust low-level generation pipeline performance.** Detailed evaluation of each pipeline stage using the "pick two boxes and open door" task shows consistently high success rates across most components. The 88.4% overall success rate demonstrates effective multi-stage coordination, while failures in the key-pose generation and post-optimization stem primarily from optimization randomness rather than systematic issues.

Low-Level Steps	Primitive Contact Gen.	Key Pose Generation	Object Motion Planning	Contact-Guided Human Motion	Post Optimization	Overall Success
Success Rate	92% (Clamp) 81% (Grasp)	96.8% (92/95)	100% (92/92)	100% (92/92)	91.3% (84/92)	88.4% (84/95)

E. Discussion and Limitations

While our framework demonstrates effective complex HOI motion generation through compositional primitives and hierarchical planning, several limitations warrant discussion.

E.1. Motion Naturalness

Unnaturalness in generated motions stems from challenges in joint human-object motion optimization. Our modular design enables zero-shot generalization, but creates difficulties in seamlessly reassembling components.

The unnaturalness arises from three sequential processes:

Object Motion Planning A* search with SDF-based collision avoidance produces geometrically valid but unnatural object trajectories. Despite reduced step sizes for smoother motion, the lack of real-world movement priors creates artificial motion patterns.

Contact-Guided Human Motion Our learned motion prior partially addresses object unnaturalness by adjusting trajectories based on human contact patterns (Fig. A9). However, limited training data for certain interactions (*e.g.*, "clamp under shoulder") prevents natural "last-mile" transitions and acceleration profiles.

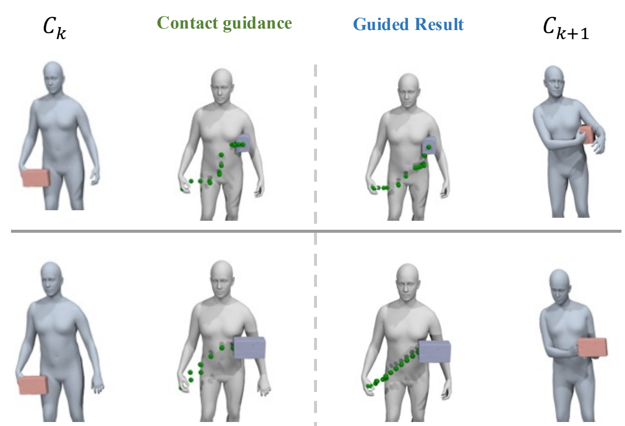


Figure A9. **LocalControl generates smooth intermediate motion between key poses.** Using planned contact points as guidance, our model produces natural trajectories that connect poses C_k and C_{k+1}. The comparison shows that guided motion significantly outperforms raw contact interpolation, demonstrating the importance of controllable motion models for realistic HOI synthesis.

Post-Optimization Final optimization incorporates human pose priors [33], contact constraints, and temporal smoothness. Unnaturalness persists due to unnatural last-mile regions in guided motion and limitations of static pose priors that do not jointly optimize temporal dynamics and contacts.

A joint human-object motion prior incorporating both kinematic constraints and physical interactions would address these issues. However, existing datasets (BEHAVE, OMOMO, HumanML3D) lack sufficient SMPL-X formatted training data for such models.

E.2. Technical Limitations

Global Motion Control Our guided motion model focuses on local operations and does not adequately handle locomotion or significant root movement, which prevents its extension to walking sequences. Enhanced model flexibility is crucial for dynamic whole-body coordination.

Grasp Accuracy Contact-point-only constraints lead to inaccurate grasp poses. Although normal constraints could improve accuracy, normal issues on the surface of the object, especially at edges, complicate implementation.

Computational Efficiency The pipeline requires significant computation time (approximately 40 seconds per motion segment) due to optimization and intermediate motion generation. Pre-computing contact primitives partially mitigates this, but post-optimization remains computationally intensive.

Future improvements include: developing joint human-object motion priors, incorporating physics-based motion models (PULSE, PHC, AMP), enhancing guided motion model precision, and optimizing computational efficiency through better initialization and fewer optimization iterations.

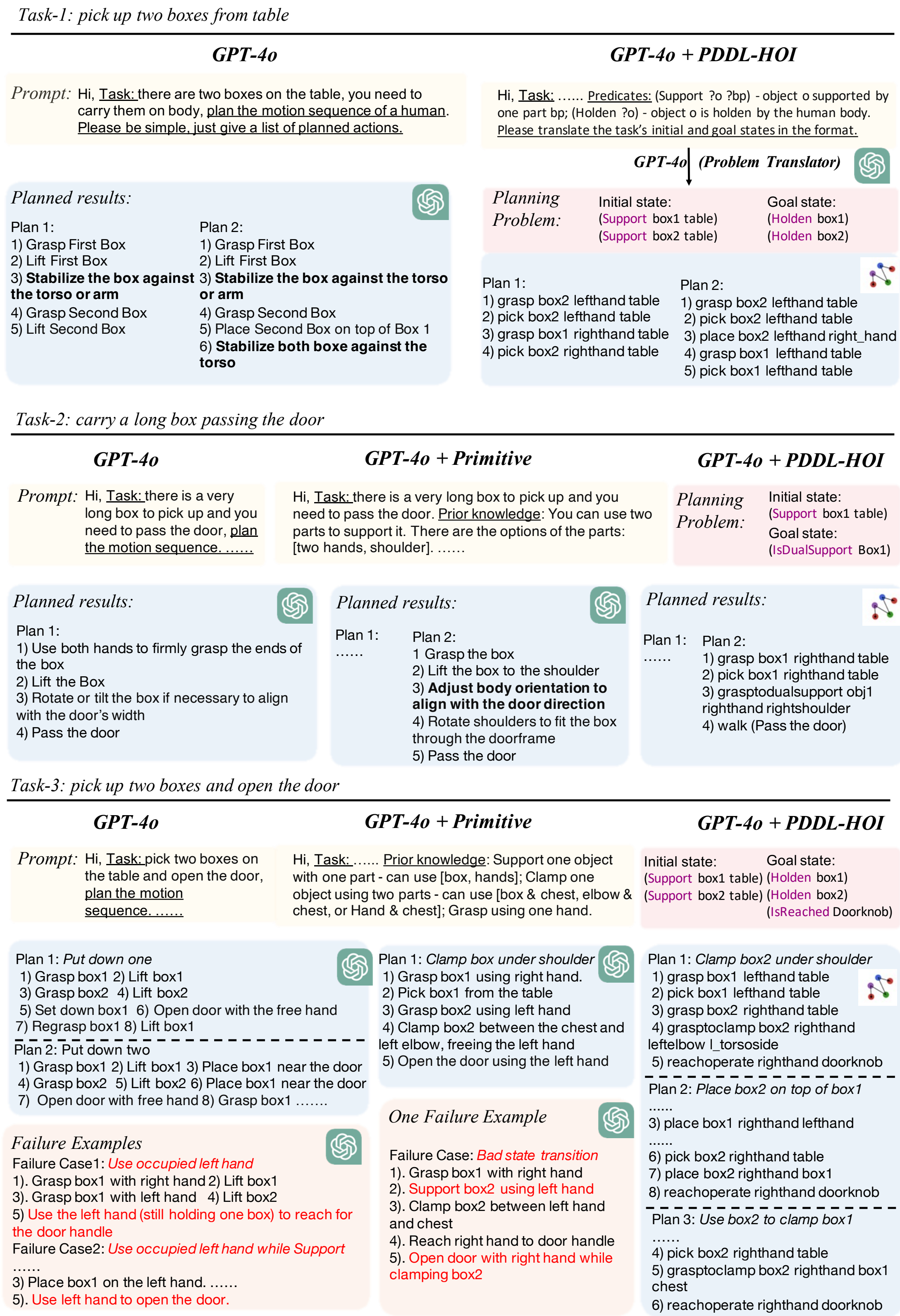


Figure A10. **Comprehensive planning evaluation demonstrates systematic superiority of our structured approach.** Complete prompts and planning outputs across three methods and all evaluation tasks reveal distinct performance patterns, where bold text indicates redundant steps and red text highlights planning failures. GPT-4o problem translations for Tasks 2-3 are omitted for brevity while maintaining result completeness. The systematic comparison establishes that our GPT-4o + PDDL-HOI method achieves superior precision and solution diversity compared to baseline approaches, validating the effectiveness of structured domain knowledge integration.