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 [55] 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

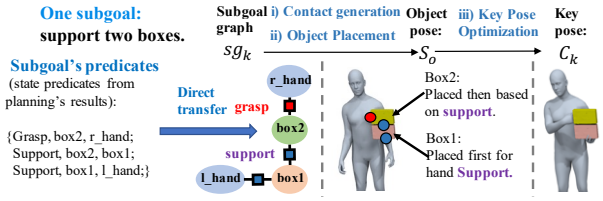
```
(define (domain object-manipulation)
(:predicates
;; Type Declarations
(IsPart ?p)
(IsObject ?o)
....
;; States
(Grasped ?o ?bp)
(Support?o ?bp)
(Clamped ?o ?bp1 ?bp2)
(DualSupport ?o ?bp1 ?bp2)
(IsReached ?o)
(Empty ?bp)
(Holden ?o)
(IsSupporten ?o)
(IsDualSupport ?o)
(IsGrasped ?o)
(IsClamped ?o)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

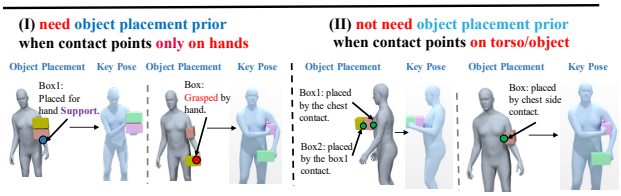
stream.pddl

```
(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.



a). from high level planning to a key pose optimization



b). object placement prior and when do we need it

Figure A2. **Key pose generation follows a three-stage pipeline.** (a) Starting from the 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. (b) We only need the object placement prior for the hand-only contact interaction.

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

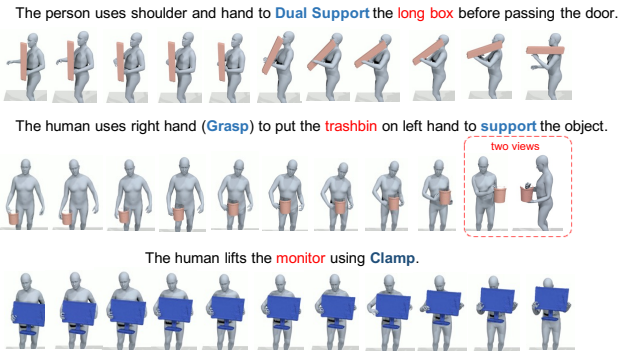


Figure A3. **Our framework demonstrates robust generalization across diverse object categories beyond training distributions.** Generated HOI motions span various object types not encountered during training, validating **PrimHOI**'s ability to transfer learned interaction patterns to novel geometric 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.

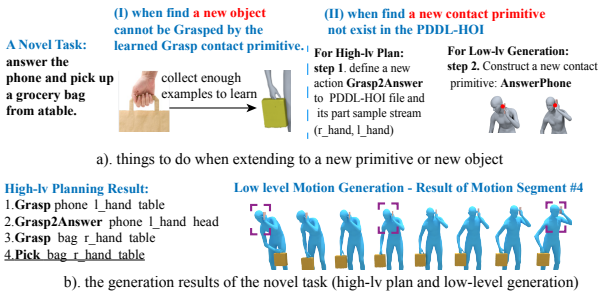


Figure A4. **Our framework enables systematic extension to novel interaction scenarios through structured primitive integration.** The three-step extension methodology facilitates incorporation of new tasks requiring previously undefined interaction primitives, while complete planning and generation 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. Interaction 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 interaction 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:

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

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

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

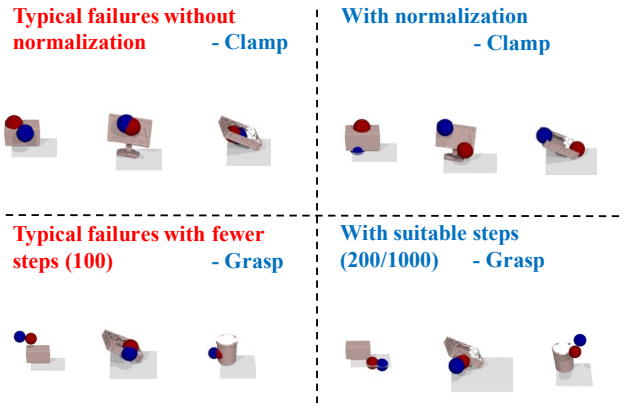


Figure A5. **Normalization enables better cross-dataset generalization for interaction primitive models.** The Clamp model shows dramatically improved success when normalization is applied during cross-dataset evaluation (BEHAVE objects after OMOMO training). Similarly, increased denoising steps benefit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

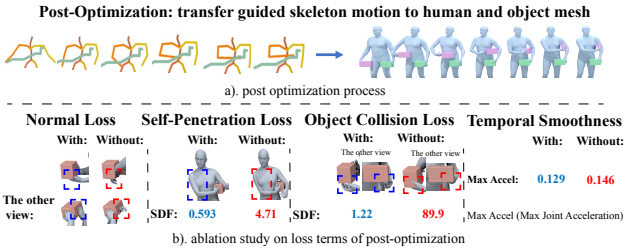


Figure A6. **Individual optimization terms address distinct motion quality challenges.** Contact guidance trajectories (green) demonstrate the post-optimization process, while ablation results reveal each term's specific contribution. Self-Penetration and Object Collision losses leverage SDF evaluation 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:

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

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

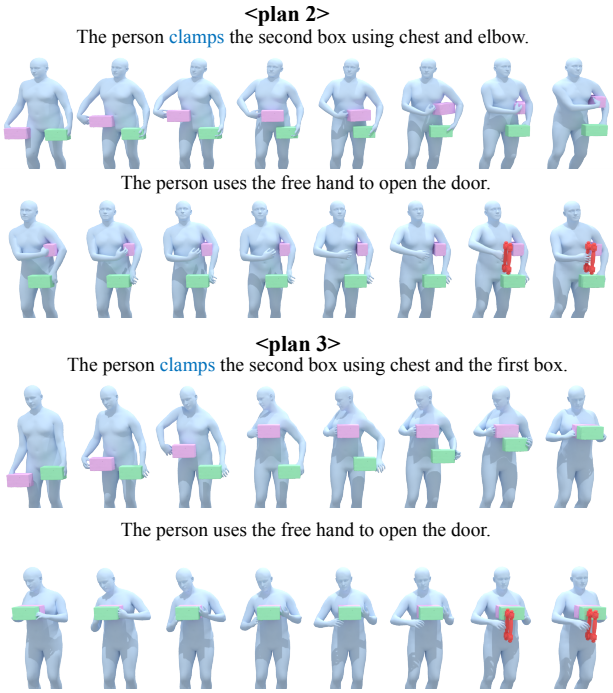


Figure A7. **Solution diversity emerges naturally from our structured planning framework.** Two alternative solutions for the "pick up two boxes and open the door" task demonstrate **PrimHOI**'s capability to generate multiple valid planning strategies for identical high-level objectives. Each solution 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.

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%

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.

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.

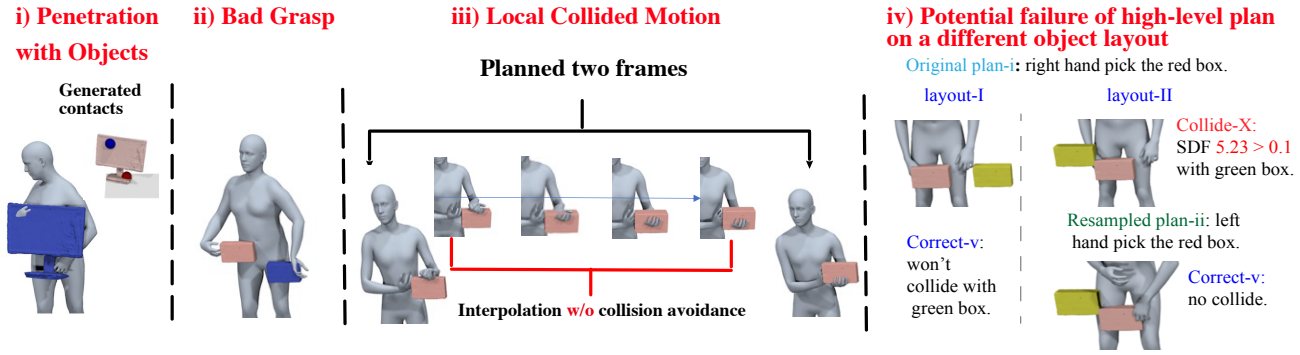


Figure A8. **Systematic failure analysis identifies four distinct limitation categories in our pipeline.** Body-object penetration occurs during key pose generation despite collision loss constraints, while incorrect grasp poses result from contact-point-only optimization without full hand orientation 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. The success rate for primitive contact generation is not included in the Overall Success calculation since they are pre-sampled.

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.

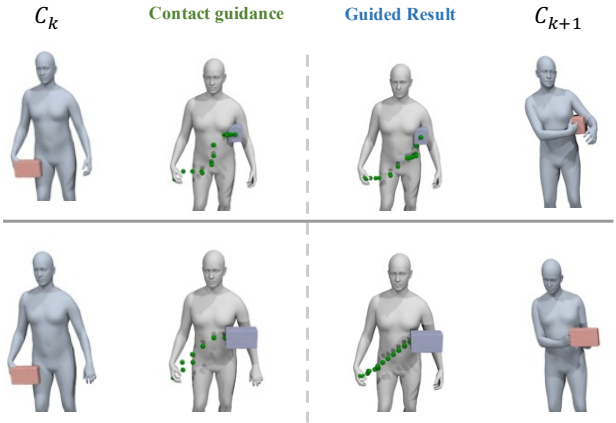


Figure A9. **LocalControl generates smooth intermediate motion between key poses.** Using planned contact points as guidance, our model produces natural trajectories that connect poses C_k and C_{k+1}. The comparison shows that guided motion significantly outperforms raw contact interpolation, demonstrating 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 interaction primitives partially mitigates this, but post-optimization remains computationally intensive.

Future improvements include: developing joint human-object motion priors, incorporating physics-based motion models (AMP [38], PHC [34], PULSE [35], etc), enhancing guided motion model precision, and optimizing computational efficiency through better initialization and fewer optimization iterations.

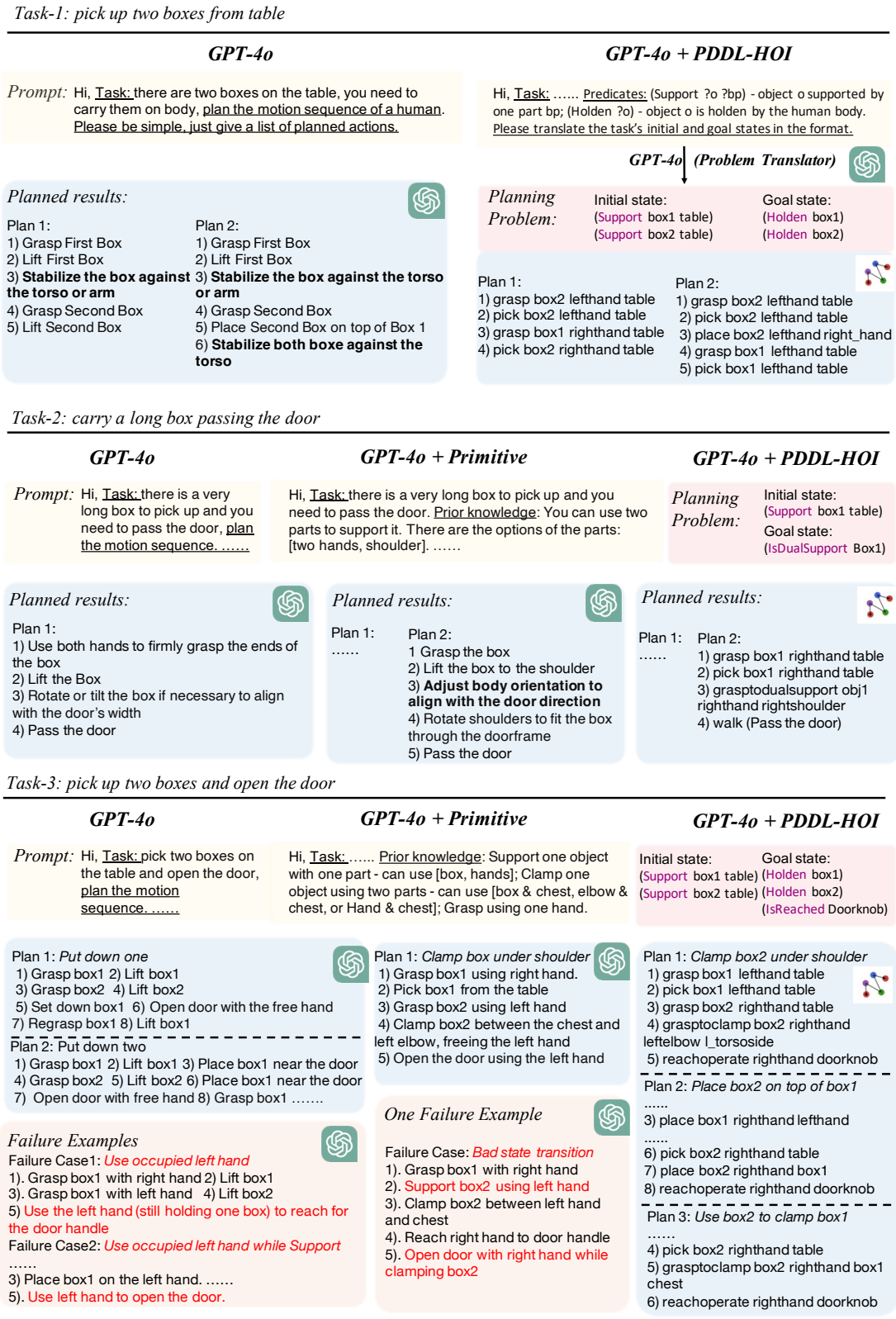


Figure A10. **Comprehensive planning evaluation demonstrates systematic superiority of our structured approach.** Complete prompts and planning outputs across three methods and all evaluation tasks reveal distinct performance patterns, where bold text indicates redundant steps and red text highlights 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.