ContA-HOI: Towards Physically Plausible Human-Object Interaction Generation via Contact-Aware Modeling

Supplementary Material

This supplementary material provides additional implementation details, experimental analysis, and visualization results for ContA-HOI. We organize the content as follows: Section 6 provides extended evaluation metrics analysis,Section 7 covers detailed implementation aspects, and Section 8 discusses limitations and future work.

6. Extended Evaluation Metrics

6.1. Contact-Related Metrics Details

While the main paper presents standard evaluation metrics, here we provide detailed explanations of our contact-related metrics and introduce an enhanced evaluation protocol.

Contact and Collision Definitions. Following ROG [34], we define contact as occurring when the minimum distance between human joints (left hand, right hand, left foot, right foot) and object surface is below 0.05m. The original ROG evaluation defines collision/penetration as human-to-object distance below 0.04m. We identified this threshold as too permissive, potentially missing subtle but perceptually noticeable penetrations. Therefore, we additionally compute collisions using a stricter threshold of 0.01m to provide a more conservative assessment of physical plausibility.

6.2. Enhanced Evaluation Results

Table 3 presents the evaluation results using our enhanced mesh-based contact metrics alongside the standard joint-based metrics from the main paper.

Table 3. Comparison of collision evaluation with different thresholds on the FullBodyManipulation dataset. Collision@0.04m uses the original ROG threshold, while Collision@0.01m uses our stricter threshold for more conservative evaluation.

Method	Contact% ↑	Collision004% ↓	Collision001% ↓
CHOIS [20]	0.44	0.25	0.20
ROG [34]	0.45	0.22	0.17
ContA-HOI (Ours)	**0.49**	0.23	**0.15**

The stricter threshold better differentiates between methods - while all methods show similar performance with the 0.04m threshold, the 0.01m threshold reveals clearer differences in their ability to avoid subtle penetrations. Our method maintains competitive performance even under the stricter evaluation, achieving the lowest collision rate (0.15) among all baselines with the 0.01m threshold.

7. Detailed Implementation

7.1. Contact Affordance Predictor (CAP)

The Contact Affordance Predictor identifies task-relevant contact regions on object surfaces by jointly reasoning about text semantics, human pose configuration, and object geometry. Here we provide additional implementation details beyond those in the main paper.

Architecture Design. CAP employs a hierarchical attention mechanism that creates a causal chain: language → human understanding → object interaction. This design is motivated by how humans naturally plan interactions: first understanding the task from language, then determining which body parts to use, and finally identifying where to contact the object.

Given input features: text embedding $\mathbf{e}_{\text{text}} \in \mathbb{R}^{512}$ from CLIP encoder, human keypoints $\mathbf{h} \in \mathbb{R}^{24\times 3}$ representing 24 SMPL-X joints, and object point cloud $\mathbf{O} \in \mathbb{R}^{1024\times 3}$ from PointNet++ encoding. The object point cloud is preprocessed to ensure uniform coverage of the object surface, combining both boundary points for overall shape and Poisson disk sampled points for fine details.

The hierarchical processing follows two steps. First, we compute language-contextualized human features:

$$\mathbf{h}_{\text{proj}} = \text{Linear}(\mathbf{h}) \in \mathbb{R}^{24\times d}, \tag{10}$$

$$\mathbf{h}_{\text{context}} = \text{CrossAttn}(\mathbf{h}_{\text{proj}}, \mathbf{e}_{\text{text}}, \mathbf{e}_{\text{text}}). \tag{11}$$

This allows language to modulate which aspects of human pose are important for the specific interaction. For instance, ”kick the box” would emphasize foot joints, while ”lift the box” would highlight hand configurations. The cross-attention mechanism learns these task-specific associations during training.

Second, we apply human-object attention to identify potential contact regions:

$$\mathbf{O}_{\text{enc}} = \text{PointNet++}(\mathbf{O}) \in \mathbb{R}^{1024\times d}, \tag{12}$$

$$\mathbf{A}_{\text{contact}} = \text{CrossAttn}(\mathbf{h}_{\text{context}}, \mathbf{O}_{\text{enc}}, \mathbf{O}_{\text{enc}}), \tag{13}$$

$$\mathbf{P}_{\text{contact}} = \text{Softmax}(\text{MLP}(\mathbf{A}_{\text{contact}})), \tag{14}$$

where $\mathbf{P}_{\text{contact}} \in \mathbb{R}^{1024}$ represents contact probability for each object point. The attention weights in $\mathbf{A}_{\text{contact}}$ capture which object regions are most relevant for the contextualized human features, effectively learning an affordance map conditioned on both the action semantics and current body configuration.

Contact Point Selection. From the predicted contact probabilities $\mathbf{P}_{\text{contact}}$, we select the top-$K$ points with highest probabilities as contact anchors for CRF construction:

$$\mathcal{C} = \text{TopK}(\mathbf{P}_{\text{contact}}, K) = \{c_1, c_2, \ldots, c_K\}, \tag{15}$$

where K is adaptively determined based on the interaction complexity, typically ranging from 4 to 24 points. These selected contact points serve as the object-side anchors in our sparse CRF representation, significantly reducing computation compared to using all 1024 points while maintaining the most task-relevant information.

Contact Validity Loss. To ensure predicted contacts are physically feasible, we introduce a contact validity loss that enforces consistency between predicted object contacts and designated human joints. This loss operates in world coordinates to account for actual spatial relationships:

$$\mathcal{L}_{\text{validity}} = \sum_{l \in \mathcal{L}_{\text{active}}} \lambda_l \cdot \max(0, d^l_{\text{min}} - \tau_{\text{reach}}), \tag{16}$$

where $\mathcal{L}_{\text{active}}$ denotes active limbs based on contact labels from the dataset (e.g., left hand, right hand, left foot, right foot), d^l_{min} is the minimum distance between limb l and its nearest predicted contact point, and $\tau_{\text{reach}} = 0.03m$ is the reachability threshold determined empirically from biomechanical constraints.

The distance d^l_{min} is computed as:

$$d^l_{\text{min}} = \min_{c \in \mathcal{C}} \|\mathbf{q}_l - \mathbf{p}_c\|_2, \tag{17}$$

where $\mathbf{q}_l$ is the 3D position of limb l and $\mathbf{p}_c$ is the world coordinate of contact point c after applying object transformations.

For training, we combine this with a binary cross-entropy loss for contact prediction:

$$\mathcal{L}_{\text{CAP}} = \mathcal{L}_{\text{BCE}}(\mathbf{P}_{\text{contact}}, \mathbf{C}^{\text{gt}}) + \alpha \cdot \mathcal{L}_{\text{validity}}, \tag{18}$$

where $\mathbf{C}^{\text{gt}} \in \{0, 1\}$ are the ground-truth contact labels obtained from the dataset annotations, where each frame of an action sequence is labeled with binary indicators (0/1) for contact of the left hand, right hand, left foot, and right foot.

Training Strategy. CAP is pre-trained independently for 100 epochs before being integrated into the full framework. This staged training ensures stable contact predictions that provide reliable anchors for CRF construction. During pre-training, we use data augmentation including random rotations and translations to improve generalization to different object orientations and positions.

7.2. Contact Dynamics Model (CDM)

The Contact Dynamics Model learns to predict realistic CRF evolution over time, providing a learned prior for physically plausible interactions. Unlike the motion generation model that operates in full pose space, CDM focuses specifically on learning the dynamics of contact relationships, enabling more targeted and effective guidance during inference.

Architecture. CDM employs a conditional diffusion model with spatio-temporal attention to capture CRF dynamics. Given the sparse CRF representation from Section 3.1, the model learns how contact relationships evolve throughout an interaction sequence.

The forward diffusion process progressively adds noise to the clean CRF:

$$q(\text{CRF}_t|\text{CRF}_{t-1}) = \mathcal{N}(\text{CRF}_t; \sqrt{1-\beta_t}\text{CRF}_{t-1}, \beta_t \mathbf{I}), \tag{19}$$

where β_t follows a cosine schedule from 10^{-4} to 0.02 over $T = 1000$ timesteps, providing smooth noise addition. The model learns to reverse this process:

$$p_\theta(\text{CRF}_{t-1}|\text{CRF}_t, \mathbf{c}) = \mathcal{N}(\text{CRF}_{t-1}; \mu_\theta(\text{CRF}_t, t, \mathbf{c}), \sigma_t^2 \mathbf{I}), \tag{20}$$

where $\mathbf{c} = \{\mathbf{e}_{\text{text}}, \mathbf{h}_0, \mathbf{O}\}$ are conditioning variables comprising text embeddings, initial human pose, and object geometry.

Network Architecture. The denoising network μ_θ consists of three key components:

Spatial Attention: Models dependencies between different joint-contact pairs within each frame:

$$\mathbf{S}_n = \text{SelfAttn}(\text{CRF}_n) \in \mathbb{R}^{M \times d}, \tag{21}$$

where M is the number of selected contact pairs and d is the feature dimension. This captures which contacts are correlated—for instance, when grasping, multiple fingers contact simultaneously.

Temporal Attention: Captures the temporal evolution of contact relationships across frames:

$$\mathbf{T} = \text{TemporalAttn}([\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N]) \in \mathbb{R}^{N \times M \times d}. \tag{22}$$

This models how contacts form, maintain, and release over time, learning typical interaction patterns like approach → contact → manipulation → release.

Conditional Cross-Attention: Incorporates semantic and geometric conditions:

$$\mathbf{F} = \text{CrossAttn}(\mathbf{T}, [\mathbf{e}_{\text{text}}, \mathbf{h}_0, \mathbf{O}]). \tag{23}$$

This ensures the predicted CRF dynamics align with the intended action and are compatible with the object's geometry and initial human configuration.

Training Objective. The CDM is trained to predict the clean CRF from noisy inputs using a reconstruction loss:

$$\mathcal{L}_{\text{CDM}} = \mathbb{E}_{t,\epsilon}\left[\|\text{CRF}_0 - \hat{\text{CRF}}_\theta(\text{CRF}_t, t, \mathbf{c})\|^2\right], \tag{24}$$

where $\text{CRF}_t = \sqrt{\bar{\alpha}_t}\text{CRF}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ is the noisy CRF at timestep t, with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

Implementation Details. The CDM network uses a U-Net backbone with skip connections to preserve fine-grained contact information across diffusion timesteps. Each attention block consists of 4 heads with dimension 256. The model is trained for 500 epochs with a learning rate of 10^{-4} using the AdamW optimizer. During training, we apply dropout with rate 0.1 to the attention layers to improve generalization.

8. Limitations and Future Work

8.1. Failure Cases

While ContA-HOI significantly improves contact modeling in human-object interactions, we identify several failure modes:

Complex Multi-Contact Scenarios. When interactions involve simultaneous contacts with multiple body parts (e.g., sitting on a chair while manipulating an object), our method occasionally misses secondary contacts. This occurs because CAP tends to focus on primary interaction points identified from text descriptions.

Dynamic Object Deformation. Our framework assumes rigid objects and does not model deformable objects like cloth or soft furniture. The rigid assumption limits applications to scenarios with deformable objects where contact dynamics differ significantly.

Fine-Grained Manipulation. For interactions requiring precise finger-level control (e.g., typing on a keyboard, playing piano), the SMPL-X joint representation lacks sufficient granularity. While SMPL-X includes hand joints, the resolution is insufficient for capturing detailed finger contacts.

8.2. Future Directions

Several promising directions emerge from this work:

Hierarchical Contact Modeling. Extending CRF to capture multi-scale contacts from full-body to finger-level interactions would enable more diverse applications.

Physics-Based Refinement. Incorporating physics simulation as a post-processing step could further improve physical plausibility, especially for complex multi-object scenarios.

Interactive Generation. Developing user interfaces that allow interactive specification of contact constraints during generation would enhance controllability for animation and VR applications.

Learning from Partial Observations. Extending the framework to learn from RGB videos without full 3D supervision would greatly expand the available training data.