

Appendix of “ChainHOI: Joint-based Kinematic Chain Modeling for Human-Object Interaction Generation”

A1. Summary

In the appendix, we first introduce more details of our method and experiments in Sec. A2. Then, we present the full experimental results using additional metrics and analyze these results in Sec. A3. Moreover, we conduct extensive ablation studies to evaluate the impact of each design choice and hyperparameter in Sec. A4 and Sec. A6. In Sec. A5, we introduce failure cases generated by our method. Finally, we discuss the limitations of our ChainHOI in Sec. A7.

A2. More Details of Our Method and Experiments

A2.1. Implementation Details

T is set to 1000 as the maximum diffusion step, and the variances β_t vary from 0.0001 to 0.02. We use DDIM [22] with 50 time steps during sampling. The number of blocks N is set to 6. D_m and D_t are set to 64 and 256, respectively. We downsample the number of object points to 16 using PointNet [19]. Our ChainHOI is optimized using AdamW [15] on two RTX 3090 Ti GPUs in parallel with a learning rate of $1e-4$ and a batch size of 32. The model is trained for 200 epochs. During testing, the guidance scale is set to 2. λ_1 and λ_2 are set to 2 and 1, respectively.

A2.2. Details of Our Loss

We note that non-watertight objects do not affect the computation of $\mathcal{G}()$ during training because $\mathcal{G}()$ computes the square of the minimum absolute distance from the joint to all triangles.

Why is the $\mathcal{G}()$ calculated from human joints to the ground truth object rather than the generated object? Because the object consists of a large number of triangular facets, using the generated object information results in a 3.6-fold increase in GPU memory usage (1.8 GB vs. 6.6 GB when the batch size is 1). We evaluate the performance using generated objects, reducing the batch size to one-fourth of the original due to GPU memory constraints. Tab. A1 shows that performance was inferior compared to using GT objects.

	FID↓	R-Top1↑	OCD↓	PS↓
gen obj.	0.098	0.437	0.089	0.081
GT obj.	0.095	0.435	0.091	0.081

Table A1. Comparisons of using the generated object or the ground truth object to calculate the distance.

A2.3. Details of Our OCD

LLM-assisted label generation and evaluation have been extensively utilized in recent studies [2, 4, 12, 26]. Specifically, we first filter grouping candidates by object category and contacting body parts (avg. 5.1 candidates on average in each group). Next, we utilize ChatGPT-4o to determine which instructions are semantically identical by evaluating action intent and the specific human body part involved in the contact. Consequently, this task is relatively straightforward for ChatGPT-4o. Table A12 shows the prompt used to group semantically identical HOIs.

We also assess the quality of the labeling process through two methods: manual evaluation and LLM-assisted evaluation. First, we review a 10% sample of the labels, achieving an accuracy rate of 94%. Second, following the approach in [12], we employ ChatGPT-4o to evaluate all labels to ensure consistency within each instruction group, resulting in a mean consistency score of 0.906 (on a scale from 0 to 1). Table A13 shows the prompt used to evaluate group labels to ensure consistency within each instruction group.

A2.4. Details of Our User Study

We initiate the evaluation by randomly selecting 20 test prompts from the BEHAVE dataset. Subsequently, for each of these 20 prompts, we instruct each method to generate 5 samples. This process yields a corpus of 100 samples, which are then used in a pairwise user study. The evaluation score is calculated as the ratio of votes received to the total votes cast.

A2.5. Details of Our Evaluator

As mentioned in our main manuscript, we adopt the metrics from T2M [6] to evaluate motion generation quality. However, computing these metrics requires a pre-trained model

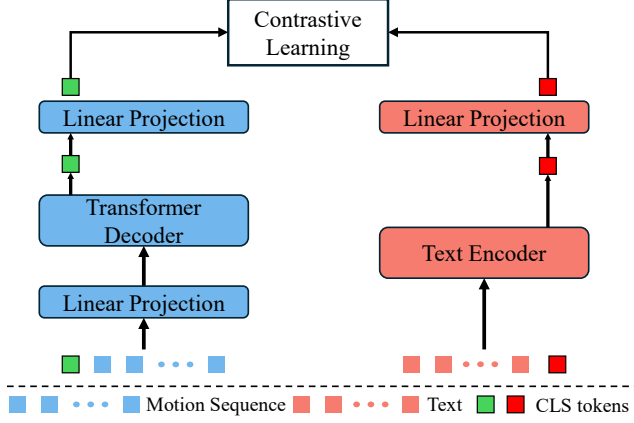


Figure A1. **Overview of Our Evaluator.** Inspired by CLIP [20], our evaluator incorporates a motion branch and a text branch. The motion branch takes motion sequences and a CLS token as inputs. The text branch takes texts and a CLS tokens as inputs. Then the output CLS tokens of two branches are passed into a linear projection and then are used to calculate the contrastive learning loss.

to extract features from both motion sequences and text descriptions. Since there is a domain gap between text-driven motion generation and HOI generation, and because HOI-Diff does not provide such a model, we trained a new feature extractor for evaluation.

Inspired by CLIP [20], we design and train our evaluator using a contrastive learning method. As shown in Figure A1, our evaluator consists of a motion branch and a text branch. The motion branch takes motion sequences and a CLS token as inputs, while the text branch takes texts and a CLS token as inputs. The output CLS tokens from both branches are then passed through a linear projection and used to calculate the contrastive learning loss.

Implementation Details. Following previous works [3, 11, 16, 23], our evaluator is used to assess the quality of human motion only (excluding object motion). During training, human motions are represented using the HumanML3D representation [6] $\bar{\mathbf{m}} \in \mathbf{R}^{L \times D}$, where $D = 263$. The motion branch consists of 8 Transformer Decoders [25], while the text branch uses the pretrained RoBERTa [13]. The dimensionality of the motion branch is 384. The output dimensions of the linear projections for both branches are 512. Our evaluator is optimized using the Adam optimizer [9] with a learning rate of 1×10^{-4} on an RTX 3090 Ti GPU. The batch size and number of training epochs are set to 64 and 200, respectively. Specifically, our evaluator is trained in two stages. In the first 8 epochs, the text branch, except for the linear projection, is fixed, and only the motion branch is trained. Then, all branches are trained together.

A2.6. Details of Compared Baselines

Apart from the methods compared in our main manuscript, we have included a modified version of CHOIS [11] for comparison, as it recently released its source code. Due to space constraints, we provide a brief description of these modifications in the main manuscript. Below, we provide a detailed description of the modified methods:

- **CHOIS*** [11]: Since CHOIS aims to generate HOI sequences conditioned on both text descriptions and object waypoints, we removed the object waypoints from the CHOIS model and modified the input and output dimensions. Note that our HOI representation is easily compatible with the one used in CHOIS.
- **InterDiff** [27]: InterDiff is designed for HOI prediction conditioned on past HOI sequences. To adapt it to text-driven HOI generation, we replaced the past HOI sequences with text descriptions. Specifically, we modified the feature dimensions and utilized the CLIP text encoder to extract text features.
- **MDM^{finetuned}** [24]: As MDM is a text-driven motion generation method, we directly fine-tuned the MDM model pretrained on the HumanML3D dataset [6] to generate human motion only. Note that MDM^{finetuned} does not generate object motion, and metrics for evaluating interaction quality are not included.
- **MDM*** [24]: To adapt MDM to the text-driven HOI generation task, we concatenated object motion (6-DoF) and human motion as the model’s inputs and outputs. This modification allows MDM* to generate HOI sequences.
- **PriorMDM*** [21]: PriorMDM introduces a ComMDM block for two-person motion generation. To adapt it to the text-driven HOI generation task, we replaced one of the two persons with the object and modified the input and output dimensions accordingly.

A3. Experiments Using More Metrics

In this section, we present comprehensive results evaluated using additional metrics to demonstrate the effectiveness of our ChainHOI. In addition to the metrics introduced in our manuscript, we utilize the following metrics to assess generation quality:

- **MultiModal Distance (MM. Dist.):** MM. Dist. calculates the average distance between the motion features of each generated motion and the text features of their corresponding descriptions in the test set. Note that the features for both motion and text are extracted by our evaluator.
- **Diversity (Div.):** Div. measures the variance in the generated motions. We randomly sample two equal-sized subsets from all motions and then compute the average distance between these subsets.
- **Contact Distance (CD):** We also report the original con-

Methods	FID↓	R-Precision↑			MM. Dist.↓	Div.↑	OCD ↓	PS ↓	FSR ↓	CD ↓
		Top1	Top2	Top3						
<i>On the BEHAVE dataset</i>										
Real motion.	0.001±.000	0.287±.011	0.443±.013	0.544±.011	1.055±.002	1.346±.011	-	-	-	-
MDM ^{finetuned} [24]	0.246±.006	0.223±.011	0.378±.005	0.488±.015	1.118±.006	1.350 ±.008	-	-	-	-
MDM* [24]	0.257±.004	0.220±.007	0.355±.001	0.451±.001	1.071±.003	1.307±.006	0.458±.013	0.095±.007	0.098±.002	0.481±.014
PriorMDM* [21]	0.328±.018	0.243±.009	0.329±.009	0.385±.013	1.203±.012	1.142±.017	0.215±.012	0.116±.001	0.066±.004	0.232±.010
InerDiff [27]	0.170±.002	0.310±.003	0.480±.005	0.599±.001	1.045±.002	1.325±.006	0.191±.027	0.078 ±.000	0.069±.002	0.206±.003
CHOIS* [11]	0.157±.001	0.301±.002	0.488±.003	0.606±.003	1.090±.002	1.265±.013	0.187±.002	0.086±.001	0.118±.003	0.202±.003
HOI-Diff [16]	0.457±.003	0.295±.003	0.441±.005	0.539±.006	1.119±.009	1.204±.017	0.148±.003	0.102±.000	0.125±.002	0.214±.003
HOI-Diff + AIC [16]	0.437±.004	0.312±.002	0.467±.003	0.563±.006	1.107±.003	1.235±.020	0.101±.001	<u>0.081</u> ±.001	0.098±.002	0.117±.003
Our ChainHOI	<u>0.095</u> ±.001	<u>0.435</u> ±.009	<u>0.621</u> ±.011	<u>0.717</u> ±.008	<u>0.967</u> ±.001	<u>1.337</u> ±.015	0.091±.001	<u>0.081</u> ±.001	<u>0.063</u> ±.000	<u>0.096</u> ±.001
Our ChainHOI + AIC	0.093 ±.001	0.444 ±.008	0.623 ±.010	0.722 ±.011	0.964 ±.004	<u>1.337</u> ±.010	0.072 ±.001	0.099±.011	0.058 ±.001	0.078 ±.001
<i>On the OMOMO dataset</i>										
Real motion.	0.001±.001	0.247±.006	0.398±.004	0.504±.005	1.050±.001	1.356±.013	-	-	-	-
MDM ^{finetuned} [24]	0.164±.004	0.123±.006	0.208±.006	0.278±.007	1.228±.004	1.333±.002	-	-	-	-
MDM* [24]	0.169±.005	0.120±.004	0.208±.006	0.281±.009	1.191±.004	1.319±.001	0.560±.003	0.022±.006	0.134±.001	0.686±.002
PriorMDM* [21]	0.329±.001	0.147±.004	0.219±.007	0.277±.005	1.200±.005	1.181±.003	0.588±.019	0.025±.001	0.115±.007	0.755±.022
InerDiff [27]	0.253±.007	0.118±.009	0.210±.009	0.281±.007	<u>1.167</u> ±.001	1.227±.003	0.472±.002	0.015 ±.001	0.139±.001	0.585±.003
CHOIS* [11]	0.251±.013	0.133±.003	0.254±.002	0.343±.003	1.193±.003	1.334±.014	0.323±.002	0.021±.001	0.151±.004	0.433±.001
HOI-Diff [16]	0.480±.001	0.114±.002	0.198±.003	0.268±.002	1.221±.008	1.124±.020	0.678±.005	0.022±.002	0.161±.001	0.763±.014
HOI-Diff + AIC [16]	0.245±.001	0.140±.002	0.253±.004	0.340±.001	1.183±.005	1.303±.014	0.301±.027	<u>0.017</u> ±.001	0.136±.004	0.331±.015
Our ChainHOI	<u>0.112</u> ±.004	<u>0.264</u> ±.005	<u>0.431</u> ±.008	<u>0.545</u> ±.008	1.023 ±.007	1.350 ±.002	<u>0.263</u> ±.002	0.019±.001	0.089 ±.009	<u>0.283</u> ±.009
Our ChainHOI + AIC	0.098 ±.002	0.266 ±.005	0.434 ±.008	0.549 ±.008	1.023 ±.007	<u>1.348</u> ±.018	0.120 ±.001	0.021±.001	<u>0.090</u> ±.002	0.136 ±.009

Table A2. **Quantitative evaluation of the BEHAVE [1] and OMOMO [10] test sets.** We repeated evaluation 20 times to calculate the average results with a 95% confidence interval (denoted by \pm). The best result is in bold, and the second best is underlined. Affordance-guided Interaction Correction (AIC) [16] is a post-processing method.

tact distance used in previous HOI generation methods. In contrast to our optimal contact distance, the original contact distance uses contact labels from a single ground-truth label.

The complete experimental results are presented in Tab. A2. These findings demonstrate that our ChainHOI performs well on the newly introduced metrics, namely MM Dist., Div., and CD. For the metrics OCD and CD, different models exhibit similar trends. The gap between OCD and CD indicates that using a single ground-truth contact label to calculate contact distance in a generative model is inappropriate. In contrast, our OCD provides a more accurate evaluation of contact distance. On the other hand, we note that it is not surprising that some models' R-Precisions outperform those of real motions, as such phenomena have been reported in many works, such as [7, 17, 18, 23].

A4. More Ablation Studies

In this section, we conduct extensive ablation studies to evaluate the effectiveness of each component and design choice.

A4.1. Impact of the Design of HOI Graph

To evaluate the effectiveness of our HOI graph, we compare our HOI graph with the following designs:

- **Discrete Graph:** The HOI graph is a discrete graph with no edge connections between any two joints.

	FID↓	R-Top1↑	OCD↓	PS↓
Discrete Graph	0.138	0.457	0.121	0.084
Complete Graph	0.154	0.443	0.106	0.086
Our HOI Graph	0.095	0.435	0.091	0.081

Table A3. **Evaluations of different HOI graph designs on the BEHAVE dataset.**

	FID↓	R-Top1↑	OCD↓	PS↓
Design A	0.179	0.450	0.113	0.090
Design B	0.095	0.428	0.103	0.084
Our Kinetic Chain	0.095	0.435	0.091	0.081

Table A4. **Evaluations of different kinetic chain designs on the BEHAVE dataset.**

- **Complete Graph:** The HOI graph is a complete graph with edges between every pair of joints.

The experimental results are shown in Tab. A3. The results indicate that, compared to the discrete graph and complete graph, our HOI graph achieves the best performance across all metrics except for R-Top1. Although the discrete graph and complete graph perform better on R-Top1, the FID of both designs is significantly lower than that of our HOI graph.

λ_1	λ_2	FID↓	R-Top1↑	OCD↓	PS↓
0	1	0.126	0.449	0.094	0.085
0.5	1	0.130	0.443	0.108	0.084
1	1	0.142	0.466	0.095	0.084
1.5	1	0.096	0.447	0.089	0.083
2	1	0.095	0.435	0.091	0.081
2.5	1	0.175	0.414	0.089	0.082
2	0	0.126	0.482	0.092	0.087
2	0.5	0.109	0.436	0.080	0.081
2	1	0.095	0.435	0.091	0.081
2	1.5	0.142	0.432	0.078	0.089

Table A5. **Impact of the training loss.** The gray line represents the configuration used in our ChainHOI model.

A4.2. Impact of the Design of Kinetic Chains

To evaluate the effectiveness of our kinetic chain design, we propose two alternative configurations for both the internal kinetic chains and the human-object chain:

- **Design A:** This design modifies the internal kinetic chains by reducing the number from five to two. The two remaining kinetic chains represent the upper and lower body, respectively. Note that the human-object chain remains unchanged in this design.
- **Design B:** This design alters the human-object chain by replacing it with a fully connected graph. Specifically, every human joint is connected to the object node, rather than only the potential interaction joints. The internal kinetic chains remain the same as in our original design.

As shown in Tab. A4, Design A achieves higher performance on R-Top1. However, the FID, OCD, and PS metrics all decrease significantly compared to our original design. In contrast, Design B, which connects the object node to all human joints, outperforms Design A in both FID and interaction-related metrics. Overall, the design used in our ChainHOI achieves better FID, OCD, and PS while maintaining good R-Top1 performance.

A4.3. Impact of the Training Loss

As mentioned in Section 3.5 of our main manuscript, we propose two loss functions to improve the quality of human-object interactions. To analyze the impact of our proposed losses, we evaluate the performance by varying the weight of each loss.

The evaluation results are shown in Tab. A5. As illustrated in the upper part of Tab. A5, constraining the distance between the predicted human joints and the ground-truth object mesh significantly improves FID and PS, and slightly improves the performance on OCD. In the lower part of Tab. A5, explicitly constraining the object’s 6-DoF significantly enhances the performance on OCD and PS.

	FID↓	R-Top1↑	OCD↓	PS↓
Shared Decoder	0.174	0.447	0.091	0.085
Independent Decoders	0.095	0.435	0.091	0.081

Table A6. **Evaluations of the design of Semantic-consistent Module and Context-aware Decoder on the BEHAVE dataset.**

#Points	FID↓	R-Top1↑	OCD↓	PS↓
8	0.159	0.364	0.083	0.091
16	0.095	0.435	0.091	0.081
32	0.109	0.446	0.095	0.087
64	0.124	0.424	0.094	0.081

Table A7. **Effect of the number of points sampled by PointNet.** The gray line represents the configuration used in our ChainHOI.

A4.4. Impact of the Design of Semantic-consistent Module and Context-aware Decoder

As shown in Figure 4 of our main manuscript, both the Semantic-consistent Module and the Context-aware Decoder adopt two Transformer decoders to separately obtain information from object geometry and text (denoted as Independent Decoders). To demonstrate the necessity of using two different Transformer decoders, we evaluate the performance when using a single decoder to obtain information from both object geometry and text simultaneously (denoted as Shared Decoder). Specifically, the object geometry tokens and text tokens are concatenated and then input into a Transformer decoder. The experimental results are shown in Tab. A6. When using the Shared Decoder, the FID drops significantly, demonstrating the necessity of using Independent Decoders.

A4.5. Impact of PointNet

Our ChainHOI adopts PointNet [19] to extract features from object geometry. To evaluate the impact of the number of points sampled by PointNet, we conducted experiments with varying point counts. The experimental results are shown in Tab. A7. The results indicate that using 8 points results in the worst performance on FID, R-Top1, and PS, while performing well on OCD. Furthermore, we find that increasing the number of points does not lead to higher performance. Therefore, we use 16 points in our ChainHOI.

A4.6. Impact of Inference Steps

We also evaluate the impact of the number of inference steps. Note that we use DDIM [22] to generate HOI sequences during inference. The experimental results are shown in Tab. A8. As the number of sampling steps increases, the model’s performance also improves. However, considering the inference time cost, we set the number of

Inference Steps	AIT	FID↓	R-Top1↑	OCD↓	PS↓
20	0.28s	0.101	0.434	0.093	0.085
50	0.61s	0.095	0.435	0.091	0.081
100	1.41s	0.093	0.436	0.090	0.081
200	2.92s	0.093	0.438	0.089	0.080

Table A8. **Impact of the number of inference steps.** The gray line represents the configuration used in our ChainHOI model. The Average Inference Time (AIT) is the mean over 100 samples on an RTX 3090Ti.

Guidance Scale	FID↓	R-Top1↑	OCD↓	PS↓
1	0.102	0.344	0.100	0.086
2	0.095	0.435	0.091	0.081
3	0.095	0.460	0.094	0.083
4	0.094	0.482	0.102	0.085
5	0.094	0.498	0.114	0.084

Table A9. **Impact of the guidance scale.** The gray line represents the configuration used in our ChainHOI model.

#Blocks	FID↓	R-Top1↑	OCD↓	PS↓
2	0.207	0.248	0.171	0.091
4	0.141	0.342	0.104	0.088
6	0.095	0.435	0.091	0.081
8	0.135	0.423	0.088	0.083

Table A10. **Impact of the number of blocks.** The gray line represents the configuration used in our ChainHOI model.

inference steps to 50 to balance generation quality and inference efficiency.

A4.7. Impact of Guidance Scale

We conduct experiments to evaluate the impact of the guidance scale during generation. We adopt the classifier-free method [8] to achieve conditional generation. The evaluation results are presented in Tab. A9. When the guidance scale is set to 1, the performance on FID, OCD, and PS is satisfactory, likely because the model utilizes the input object geometry to guide HOI generation. However, without text guidance, the generated HOIs may not correspond to the provided text, leading to a lower R-Top1 score. Conversely, as the guidance scale increases, the performance on FID and R-Top1 improves, while the quality of human-object interactions declines, since the quality of human-object interactions does not depend on text guidance.

	FID↓	R-Top1↑	OCD↓	PS↓
HOI-Diff	0.514	0.097	0.281	0.023
CHOIS*	0.368	0.107	0.269	0.026
Our	0.154	0.186	0.238	0.022

Table A11. **Generalization Performance Evaluation.** All methods are trained on the OMOMO dataset and evaluated on the 3D-FUTURE dataset.

A4.8. Impact of the Number of Blocks

Furthermore, we evaluate the impact of the number of blocks in our ChainHOI model. As shown in Tab. A10, both increasing and decreasing the number of blocks lead to a performance drop. Therefore, the ChainHOI model using six blocks is our final model.

A4.9. Generalization Performance Evaluation

To evaluate ChainHOI’s generalization performance on unseen objects, we tested our model, pre-trained on the OMOMO dataset, on the 3D-FUTURE dataset [5] using the protocol established by CHOIS [11]. Results in Table A11 reveal that although a performance drop is observed for all methods, our approach maintains superior performance compared to others.

A5. Failure Case Analysis

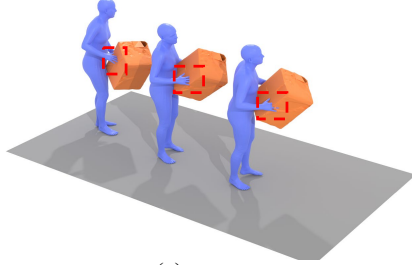
As shown in Fig. A2, we present two typical failure cases encountered by our method:

- *Issues with Finger-Object Clipping:* Subfigures (a) and (b) illustrate problems related to clipping between the fingers and objects. Since the input data is represented using the SMPL model [14], which does not include finger joint information, our ChainHOI is unable to accurately model the fingers, resulting in clipping between the fingers and objects.
- *Large Contact Distances:* Subfigure (c) shows that the contact distance may be excessively large when interacting with certain complex objects. For instance, with objects such as chairs, our model struggles to learn the correct contact points and appropriate contact distances.

A6. Visualization of Attention Scores in Our Kinematic-aware Decoder.

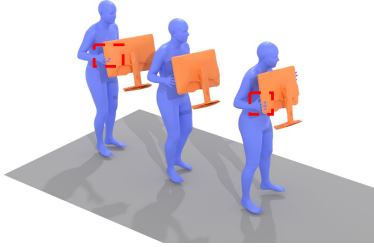
As shown in Fig. A3, we present two examples to demonstrate that our ChainHOI can adaptively focus on the joints interacting with the target object. For other potential interaction joints that have low relevance to the target object, lower attention scores are assigned. The results show that our method effectively captures the relationship between the target object and the precise interaction joints.

A person **clutches the box** large from the front.



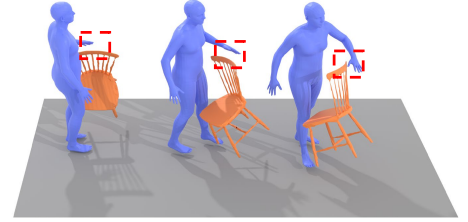
(a)

A person is **gripping the monitor** from the front.



(b)

A person is **raising a chair** wood with his left hand.



(c)

Figure A2. **Visualization of Failure Cases.** We present two typical failure cases encountered by our method. *Issues with Human-Object Clipping:* Subfigures (a) and (b) illustrate problems related to clipping between the fingers and objects. Since the input data is represented using the SMPL model [14], which does not include finger joint information, our ChainHOI is unable to accurately model the fingers, resulting in clipping between the hands and objects. *Large Contact Distances:* Subfigure (c) demonstrates that the contact distance may be excessively large when interacting with certain complex objects. For instance, with objects such as chairs, our model struggles to learn the correct contact points and appropriate contact distances.

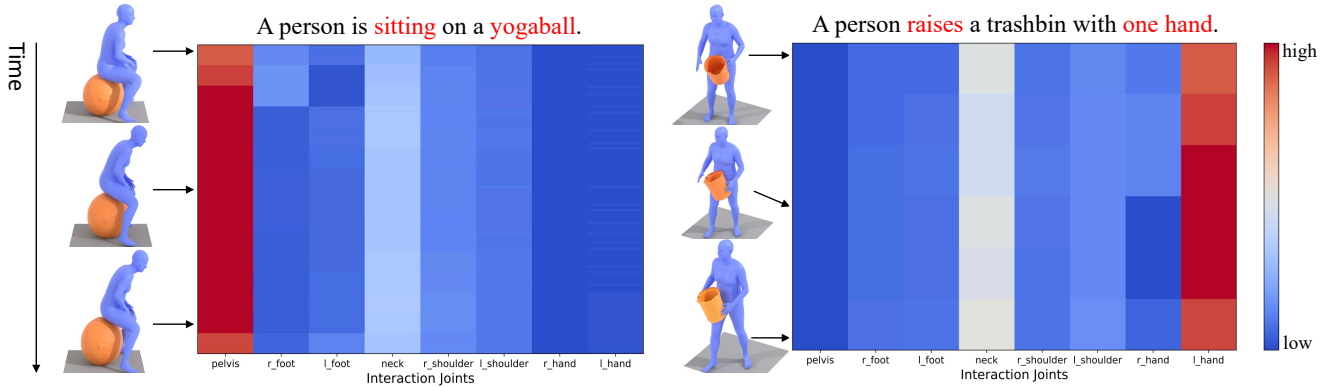


Figure A3. **Visualization of attention scores in our Kinematic-aware Decoder.** We present two examples to demonstrate that our ChainHOI can adaptively focus on the joints interacting with the target object. For other potential interaction joints that have low relevance to the target object, lower attention scores are assigned. The results show that our method effectively captures the relationship between the target object and the precise interaction joints.

A7. Limitations

Although our ChainHOI is capable of generating realistic and coherent human-object interactions, it still has certain limitations. First, due to the SMPL human representation used in the BEHAVE [1] and OMOMO [10] datasets, our ChainHOI is unable to accurately model the fingers and prevent clipping between the fingers and objects. Second, as analyzed in Sec. A5, our ChainHOI struggles to learn the correct contact points and appropriate contact distances for complex objects. Furthermore, due to the physical geometry information extraction method adopted in our approach, ChainHOI is unable to handle interactions between humans and non-rigid objects.

References

- [1] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 3, 6
- [2] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 852–862, 2024. 1
- [3] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 2
- [4] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang,

“System Prompt:” You are an assistant of understanding and grouping human motion instructions. Given several groups of instructions that describe interactions between humans and objects, the instructions in each group represent semantically consistent actions. Your task is to identify which groups represent semantically consistent actions. Please provide your output in the following format: `[[1, 2], [3]]`, which represents group 1 and 2 are consistent.

Here is an example:

###Input###:

Group 1:

The person is gripping the yogamat from the front.
The person has a firm grasp on the yogamat from the front.
The person is clutching the yogamat from the front.

Group 2:

The person is clutching a yogamat against his body with his right hand.
The individual is clasping a yogamat near his body with his right hand.
The person is gripping a yogamat close to his body with his right hand.

Group 3:

A person is gripping a yogamat in front.
A person is carrying a yogamat in front.
A person is clutching a yogamat in front.

Group 4:

The individual is clutching a yogamat with his left hand, keeping it firmly against his body.
A person is grasping onto a yogamat, holding it tightly against his body with his left hand.
Someone holds a yogamat close to his body, with his left hand gripping onto it tightly.

Group 5:

The person is grasping the yogamat from the front.
A person has ahold of the yogamat from the front.
The person has taken possession of the yogamat from the front.

Output ###:

`[[1, 3, 5], [2], [4]]`

Table A12. The prompt used to group semantically identical HOIs.

- Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36: 30039–30069, 2023. 1
- [5] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 5
- [6] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 1, 2
- [7] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 3
- [8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
- [9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2
- [10] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 3, 6
- [11] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV) 2024*, 2024. 2, 3, 5
- [12] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. In *European Conference on Computer Vision*, pages 109–127. Springer, 2024. 1
- [13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 2
- [14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 5, 6
- [15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay

“System Prompt:” You are an assistant responsible for evaluating and checking the consistency of human motion instructions. Given several groups of instructions that describe interactions between humans and objects, your task is to assess the semantic similarity among all groups. The instructions in one group are semantically consistent. You should output a similarity score between 0 and 1, where 1 indicates all groups are semantic similar, and 0 indicates complete inconsistency. Only when groups show completely inconsistent semantics and you are very certain, score 0. Do not output any content other than scores.

Here is an example:

#####Example 1#####

###Input###

Group 1:

The person is pushing the chairwood back and forth.

The person is moving the chairwood back and forth.

The person is exerting force on the chairwood, moving it back and forth.

Group 2:

A person is sitting on the chairwood.

A person is occupying the chairwood.

A person is positioned on the chairwood.

###Output###

0

#####Example 1#####

###Input###

Group 1:

The person is propelling the tablesquare with his foot.

The person is nudging the tablesquare using his foot.

The person is pressing the tablesquare forward by his foot.

Group 2:

A person is nudging the tablesquare with his foot.

A person is shoving the tablesquare with his foot.

A person is prodding the tablesquare with his foot.

###Output###

1

#####Example 1#####

###Input###

Group 1:

A person clutches the boxlarge from the front.

A person firmly grasps the boxlarge from the front.

A person tightly holds the boxlarge from the front.

Group 2:

A person is grasping the boxlarge using only his left hand.

A person has lifted the boxlarge using only his left hand.

A person is clutching the boxlarge using only his left hand.

###Output###

1

Table A13. The prompt used to evaluate group labels to ensure consistency within each instruction group.

- regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1
- [16] 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. 2, 3
- [17] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. *arXiv preprint arXiv:2403.19435*, 2024. 3
- [18] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 1546–1555, 2024. [3](#)

- [19] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. [1](#), [4](#)
- [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#)
- [21] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. [2](#), [3](#)
- [22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. [1](#), [4](#)
- [23] Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2024. [2](#), [3](#)
- [24] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. [2](#), [3](#)
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. [2](#)
- [26] Yi-Lin Wei, Jian-Jian Jiang, Chengyi Xing, Xiantuo Tan, Xiao-Ming Wu, Hao Li, Mark Cutkosky, and Wei-Shi Zheng. Grasp as you say: Language-guided dexterous grasp generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. [1](#)
- [27] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. [2](#), [3](#)