Supplementary Material of HOIAnimator: Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models

Wenfeng Song[1], Xinyu Zhang[1], Shuai Li[2,3*], Yang Gao[3], Aimin Hao[3,5], Xia Hou[1], Chenglizhao Chen[4], Ning Li[1], Hong Qin[6†]

[1]Beijing Information Science and Technology University [2]Zhongguancun Laboratory, China

[3]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University

[4]College of Computer Science and Technology, China University of Petroleum (East China)

[5]Research Unit of Virtual Human and Virtual Surgery (2019RU004), Chinese Academy of Medical Sciences

[6]Department of Computer Science, Stony Brook University (SUNY at Stony Brook), Stony Brook, New York 11794-2424, USA

https://zxylinkstart.github.io/HOIAnimator-Web/

1. Overview

In this supplementary material, we commence by delineating the detailed HOIAnimator used for computing the metrics (Section 2). Next, we provide an in-depth presentation of the user study details (Section 3). Subsequently, we delve into additional experiments (Section 4), and provide an expanded set of results (Section 5). Finally, we provide a comprehensive illustration of the network architecture (Section 6).

2. Evaluation Criteria

In this section, we detail the text and Human-Object Interaction (HOI) animation feature extractors, which are crucial components of our metric computation framework. Additionally, our evaluation criteria for the HOIAnimator encompasses six key metrics: the Fréchet Inception Distance (FID), Diversity, MultiModality Diversity (MM-Dist), R Precision, and Penetration. The comprehensive testing protocol is meticulously designed to rigorously evaluate the efficacy and robustness of our approach in generating HOI animations across diverse scenarios.

Text and HOI Animation Feature Extractor. Following the approach outlined in T2M [2], our HOIAnimator involves utilizing a text extractor to convert raw text into a semantic feature vector, denoted as (s). Concurrently, HOI animations are processed through an HOI animation extractor, resulting in another feature vector (m). In this process, we aim to minimize the distance between matched pairs of text and HOI animation feature vectors, ensuring feature vectors' close correspondence.

FID. FID calculates the distance between real samples and generated samples in latent space. Following the animation generation work [4], we define $FID(x, \hat{x}) = \| \mu_x - \mu_{\hat{x}} \|_2^2 + Tr(\Sigma_p + \Sigma_{\hat{x}} - 2(\Sigma_p \Sigma_{\hat{x}})^{0.5})$. Here, x and $\hat{x}$ represent the real HOI animations and the generated HOI animations. FID is an objective metric calculating the distance between features extracted from real and generated motion sequences, which reflects the generation quality.

Diversity. Diversity evaluates the variability of the generated HOI animations across a range of descriptions. To measure this, we randomly sample two subsets of equal size (S_d), from the entire collection of motions generated from various descriptions. We extract respective sets of HOI animation feature vectors $\{m_1, ..., m_{S_d}\}$ and $\{m'_1, ..., m'_{S_d}\}$. The diversity of the set of motions is defined as, $Diversity = \frac{1}{S_d} \sum_{t=1}^{S_d} \|m_i - m_i{}'\|$, where $S_d = 300$ is used in experiments.

MM-Dist. Distinct from the concept of diversity, MM-Dist quantifies the extent to which the generated motions vary within each individual text description. It is measured across a dataset containing motions paired with C distinct descriptions. For c-th description, we randomly sample two subsets with the same size S_m and then extract two subsets of feature vectors $\{m_{c,1}, ..., m_{c,S_m}\}$ and $\{m'_{c,1}, ..., m'_{c,S_m}\}$. The multimodality of the motion set is formalized as, $MM - Dis = \frac{1}{C*S} \sum_{c=1}^{C} \sum_{t=1}^{S_m} \|m_{c,i} - m_{c,i}{}'\|$, where $S_m = 10$ is used in experiments.

R Precision. The R precision metric evaluates the similarity between the textual description and the generated motion sequence. It represents the likelihood that the actual text ranks within the top k positions after sorting. In this study, we set k to be 1, 2, and 3.

Vertex Distance. Vertex distance evaluates generation quality by comparing distances between vertices in real and generated objects.

Penetration. Similar to prior work [3], the penetration

*,† Corresponding authors

score evaluates whether the human gets close to the object during the interaction. We define the approach phase as the initial motion frames from a sequence N_A. Then the penetration distance for a trajectory is $\frac{1}{N_A}\sum_v \sum_i^{N_A} \mathbf{sdf}(v) \cdot \mathbb{1}_{\mathbf{sdf}(v)>0}$, where $\mathbb{1}$ is indicator function. $\mathbf{sdf}_i$ is the signed distance function of the human in the i^{th} frame and v is one of 2K points on the object's surface. We report the percentage of trajectories with penetration distance $\leq 2cm$, ignoring trajectories with distance to object $> 2cm$, since trajectories that do not approach the object will trivially avoid penetration.

3. User Study Details

In this section, we provide more details of our user study. We use the WenJuanxing [1] website to design and collect our questionnaires. We show our designed user interface, where users should rate each HOI animation as shown in Fig. 1. We invite 40 participants from varied backgrounds, comprising 22 students, 3 salespeople, 6 software engineers, 2 teachers, 3 managers, and 4 individuals from other professional fields. Among all participants, 65% are male, and a significant majority of 79% fell within the age range of 18 to 25 years.

4. Additional Experiments

4.1. PMP Modules and Their Analysis

In this section, we explore the impact of the Object Passage and Dual Flow modules within our network's architecture. To assess modules' contributions to PMP, we conducted a series of comparative experiments, which included removing or altering the positions of modules. The results, presented in Table 1, indicate that placing the Object Passage module at the beginning of the network significantly improves performance, increasing precision from 0.699 to 0.771. In contrast, positioning the Dual Flow module towards the end of the network is more beneficial, as demonstrated by an increase in precision from 0.768 to 0.772, compared to its initial placement. 'DF' refers to our approach of combining two features through concatenation. The PMP is more effective than concatenation, as evidenced by the improvement in accuracy from 0.768 to 0.781. Our method ensures text congruence and enhances the overall quality of the generated HOI animations.

4.2. ICF Modules and Their Analysis

At the same time, it can be seen from Fig. 3 that ICF has great guidance on HOI animation.

4.3. Importance of Object Motion guided Human Motion

In this section, we demonstrate the symbiotic guidance relationship between humans and objects within HOIAnimator. We dynamically influence the features of objects by in-

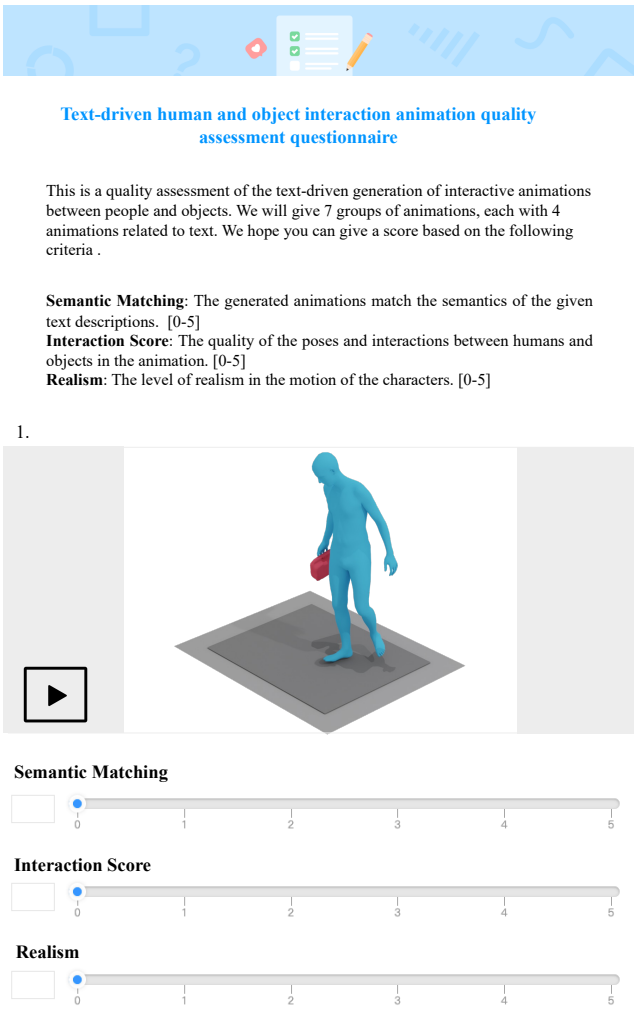

Figure 1. **User interface in our user study.** We ask users to rate the synthetic animations on three aspects: Semantic Matching, Interaction Score, and Realism.

Methods	Precision↑	FID ↓	Penetration↑
Real motions	$0.821^{\pm 0.005}$	$0.012^{\pm 0.002}$	–
– | DF	$0.772^{\pm 0.007}$	$0.631^{\pm 0.027}$	$0.623^{\pm 0.003}$
DF | –	$0.768^{\pm 0.003}$	$0.632^{\pm 0.027}$	$0.622^{\pm 0.003}$
– | OP	$0.699^{\pm 0.004}$	$0.634^{\pm 0.047}$	$0.621^{\pm 0.002}$
OP | –	$0.771^{\pm 0.004}$	$0.632^{\pm 0.032}$	$0.627^{\pm 0.005}$
DF | OP	$0.778^{\pm 0.007}$	$0.625^{\pm 0.052}$	$0.638^{\pm 0.003}$
Ours	$\mathbf{0.781}^{\pm 0.005}$	$\mathbf{0.623}^{\pm 0.063}$	$\mathbf{0.643}^{\pm 0.001}$

Table 1. **PMP module analysis.** We adjust the position of Object Passage (OP) and Dual Flow (DF) or remove a certain part. Our configuration can achieve the best results. Here, '|' indicates the order. '–' means do not use any operation.

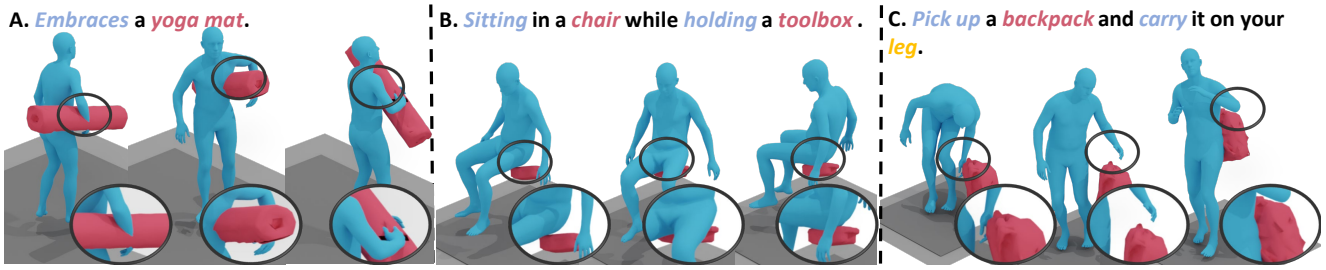

Figure 2. **Failure cases.** We present cases in more diverse scenarios.

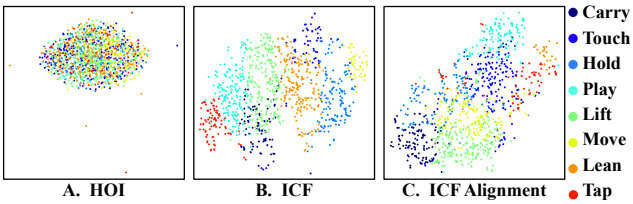

Figure 3. **Tsne [5] of HOI animation.** The feature is by HOI encoder (A), ICF (B), and HOI animation after ICF alignment (C).

terchanging and comparing them with the motion features of humans in the Object Passage. As shown in Table 2, network performance is significantly enhanced when object features guide human motion features. Specifically, precision increases from 0.772 to 0.781, the FID improves from 0.689 to 0.623, and the Penetration metric rises from 0.631 to 0.643. The results evidence suggests that object motion features are more effective in guiding human motion features. We attribute results to the more sensitive nature of object location features. Conversely, using human features rich in information content to guide object features tends to be counterproductive. We produce high-quality HOI animations by steering human motion features using the nuanced aspects of object motion features.

Methods	Precision↑	FID ↓	Penetration↑
Real motions	$0.821^{\pm 0.005}$	$0.012^{\pm 0.002}$	–
Human | Object	$0.772^{\pm 0.005}$	$0.689^{\pm 0.027}$	$0.631^{\pm 0.002}$
Object | Human	$\mathbf{0.781}^{\pm 0.005}$	$\mathbf{0.623}^{\pm 0.063}$	$\mathbf{0.643}^{\pm 0.001}$

Table 2. **Object passage.** We adjust the order of human and object centric diffusion model passage.

5. More Results

In this section, we showcase an expanded range of our experimental results. The results encompass three key areas: firstly, the generation of diverse HOI animations from textual descriptions; secondly, the interaction of various motions with the single object label; and thirdly, the exchange of diverse objects with the same motion label. We demonstrate the versatility and adaptability of our approach in creating varied and dynamic HOI animations.

Text to Diverse HOI Animations. Our HOIAnimator can generate diverse object interaction animations from simple text descriptions. When the text prompt is "A person moves a stool with his hands", HOIAnimator can generate diverse animations of the stool moving in diverse directions, as shown in Fig 4.

Interaction of Diverse Objects with the Same Motion. We demonstrate HOIAnimator's capability to generate a range of animations when the motion label prompt is "Holding". We show how it adeptly depicts holding diverse objects, such as a yoga ball, a backpack, and a box, as shown in Fig. 5. Our HOIAnimator can generate diverse object interaction motions based on the same motion label.

Interaction of Diverse Motions with the Same Object. We demonstrate how our HOIAnimator, using only the simple prompt "wooden chair", can generate a range of animations that interact with the wooden chair. The HOI animations encompass diverse actions like moving, lifting, and sitting, as depicted in Fig. 6. The results showcase the HOIAnimator's capability to create diverse interactive motions from a singular object label.

Failure Cases. We have conducted extensive experiments to verify that our method could be generalized to most scenarios except deformable objects (e.g., a yoga mat in Fig 2-A) and multiple humans/objects simultaneously (e.g., 'Sitting in a chair while holding a toolbox' in Fig 2-B). The 1st case comes from the deformation objects. Currently, our model does not support nonrigid object animations, but it is a key goal for our future work. The 2nd case is that our dataset is all for a single object and human. Multiple humans or multiple objects are what we expect to solve in the future.

Our HOIAnimator fits the text of HOIs in most seen scenarios by employing diverse fine-grained textures during dataset annotation. Even when faced with contradictory descriptions, our model remains robust. For instance, it can correct descriptions to produce reasonable motions, even with minor positional misalignment as shown in Fig 2-C (e.g., generating 'carry it on the leg' similar to 'carry it on the back'). This adaptability results from the similarity of the two text feature spaces learning from the dataset.

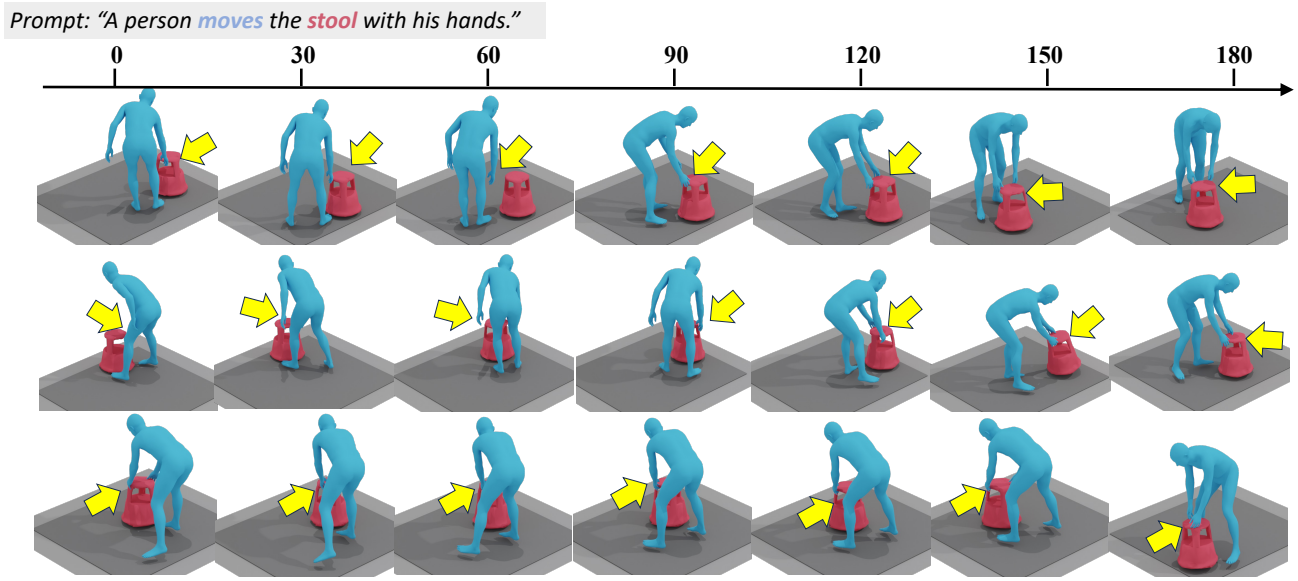

Figure 4. **Text to diverse HOI animations.** HOIAnimator can produce highly consistent HOI animations that align seamlessly with the given text, featuring rational interactions with high diversity. The yellow arrows indicate interaction areas.

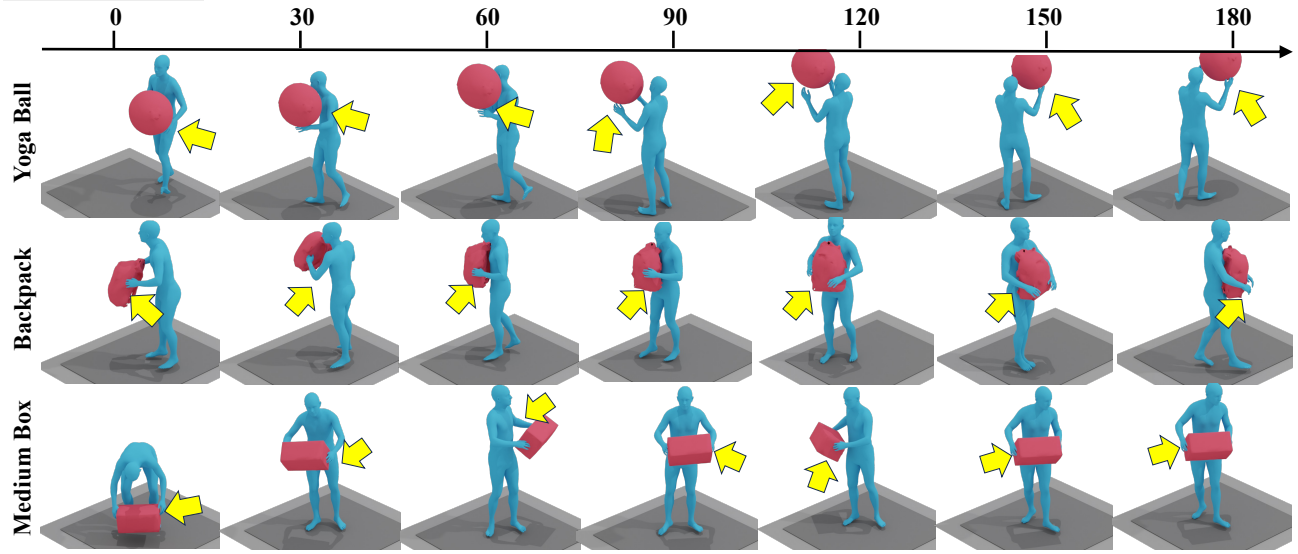

Figure 5. **Interaction of various objects with the same motion.** HOIAnimator can generate diverse object interaction motions based on the same motion label.

6. Network Architecture

In order to enable our method to be successfully reproduced, we elaborate our HOIAnimator network structure in Table 3.

References

[1] Wenjuanxing. https://www.wjx.cn/. 2

[2] 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*, pages 5152–5161, 2022. 1

[3] 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. *arXiv preprint arXiv:2307.07511*, 2023. 1

[4] Mathis Petrovich, Michael J Black, and Gül Varol. Action-

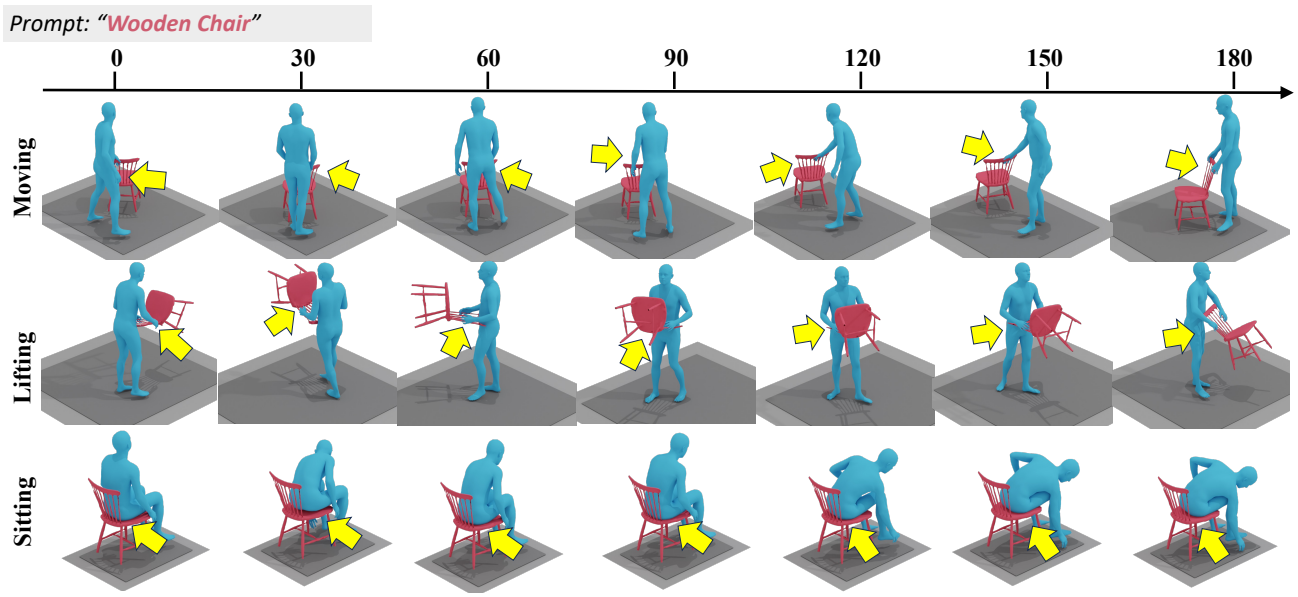

Figure 6. **Interaction of diverse motions with the same object.** HOIAimator can generate diverse interactive motions based on the same object label.

Text Encoder	Frozen CLIP ViT-B/32 TransformerEncoderLayer(d_model=256, num_heads=4, dim_feedforward=1024) × 2
Times Encoder	Linear(in_features=1000, out_features=512) Mish()
HOI Encoder	Linear(in_features=165, out_features=1024)
Object Passage	Linear(in_features=1024, out_features=1024) Linear(in_features=1024, out_features=1024)
Dual Flow	TransformerEncoderLayer(d_model=256, num_heads=4, dim_feedforward=2048) × 2 Linear(in_features=2048, out_features=1024)
Latent Encoder	Linear(in_features=512, out_features=512)
TransformerEncoder	Linear(in_features=1024, out_features=1024, bias=True) (Vertice) LeakyReLU(negative_slope=0.2, inplace=True) Linear(in_features=7, out_features=1024, bias=True) (Emotion) LeakyReLU(negative_slope=0.2, inplace=True) Conv1d(in_channels=1024, out_channels=1024, kernel_size=5, stride=1, padding=2, padding_mode='replicate') LeakyReLU(negative_slope=0.2, inplace=True) InstanceNorm1d(num_features=1024) Linear(in_features=1024, out_features=1024, bias=True) Transformer(in_size=1024, hidden_size=1024, num_hidden_layers=6, num_attention_heads=8, intermediate_size=1536) Linear(in_features=1024, out_features=1024, bias=True)
Latent Decoder	Linear(in_features=512, out_features=165)

Table 3. **Architecture of our method.** We provide detailed network architecture of our key components, including HOI Encoder, Object Passage, and Dual Flow.

conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on*

Computer Vision, pages 10985–10995, 2021. 1

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9: 2579–2605, 2008. 3