2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

OASIS: Object-guided Attention for Text-conditional Diffusion Synthesis of Human Interaction Sequences

Chih-Chun Yang*[1] Tianhui Cai*[1] Zoltán Milacski[1] Aayush Prakash[2]
Shingo Takagi[2], Daeil Kim[2], Fernando de la Torre[1]
[1] Carnegie Mellon University, USA
[2] Meta, USA

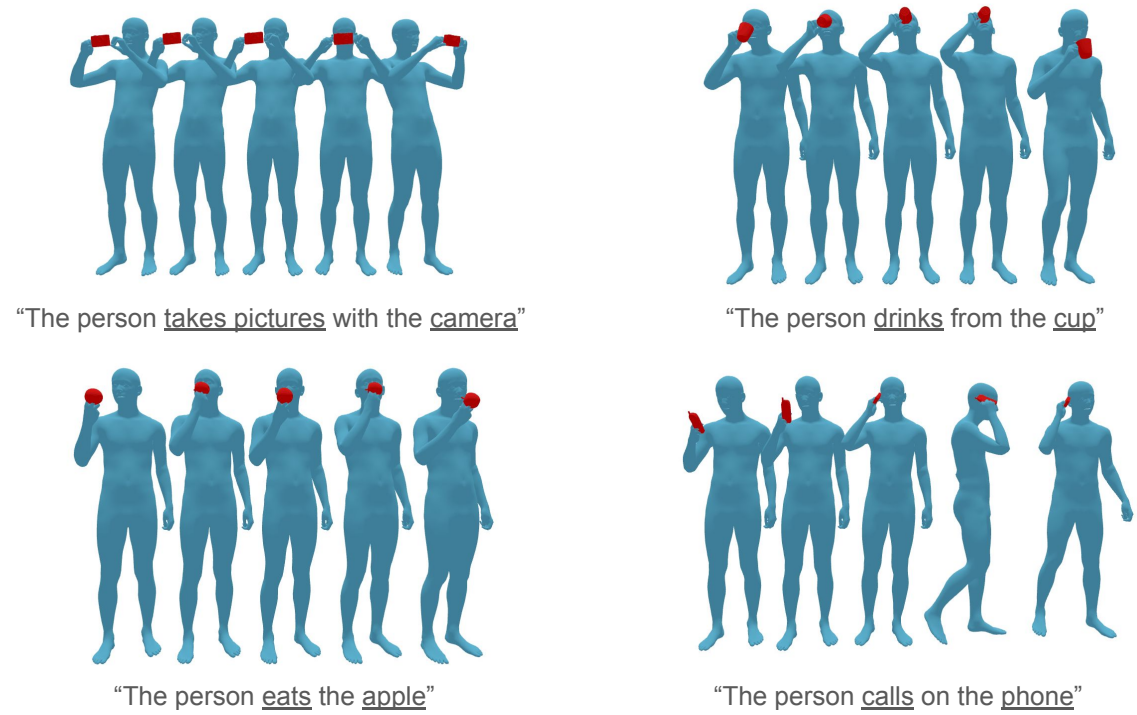


Fig. 1: Illustration of text-driven 3D human-hand-object interactions. The sequences depict different interactions based on textual descriptions. Each sequence demonstrates how the model generates 3D human poses and object manipulations corresponding to the given textual commands.

***Abstract*— Analyzing and synthesizing human-object interaction is crucial for advancing intelligent systems that engage with the physical environment. However, simultaneous tracking of human and object data presents inherent challenges, resulting in limitations in dataset scale, diversity, and annotation quality within this domain, thereby hindering the generalization ability of trained models. This study introduces OASIS, a novel framework that extends pretrained text-conditional human motion diffusion models to address the complex task of full-body 3D hand-object interaction generation. Specifically, we freeze the parameters of the pretrained motion diffusion model, while incorporating additional object-guided attention layers, which we train to adapt the human motion latents to match the input object motion sequence and the text. Our method can be understood as a ControlNet[38] for interaction. Through extensive experimentation, we demonstrate the effectiveness and robustness of our framework in generating realistic hand-object interactions from textual descriptions. Our method surpasses the state-of-the-art performance in FID and accuracy interaction fidelity metrics compared to the prior best method IMoS [10], with improvements of 0.08 in FID and 2% in accuracy for body motion synthesis, and 0.15 in FID and 10% in accuracy for hand motion synthesis.**

* denotes equal contribution

I. Introduction

Understanding human-object interaction is crucial for developing intelligent systems capable of effectively engaging with the physical world. Despite significant advancements in computer vision and machine learning, datasets specifically tailored for full-body human-object interaction remain scarce. This scarcity presents a significant bottleneck, hindering the development and evaluation of models and algorithms aimed at understanding and synthesizing human-object interactions.

Human-object interaction is often closely tied to hand-object interaction, as hands are the primary means through which humans manipulate objects. Datasets such as HO-3D

979-8-3315-5341-8/25/$31.00 ©2025 IEEE

[14], DexYCB [5] and Obman[15], which capture or synthesize hand interacting with objects, has been built to facilitate the study in this field. For example, GraspTTA model [18] is trained on Obman dataset [15] to generate grasp pose of hand on objects, and D-Grasp [6], which is trained on DexYCB dataset [5], adopts reinforcement learning to synthesize hand motions of grasping and moving objects. However, these works are limited to hand interactions and do not encompass full-body motion.

Collecting datasets for full-body human-object interaction poses several challenges. Unlike static image datasets, capturing dynamic interactions between humans and objects requires careful planning, sophisticated equipment, and meticulous annotation efforts. Consequently, existing datasets such as GRAB [30] often suffer from limitations in scale, diversity, and annotation quality, further exacerbating the challenges faced by researchers in this field.

The most common solution to tackling this problem is to decompose the task into two steps: i) start-conditional goal pose and contact generation, and ii) contact-conditional interpolation between the start and goal poses. Both GOAL [29] and SAGA [34] follow this approach, training the first step using the small GRAB dataset [30] . However, GOAL uses deterministic autoregressive inpainting in the second stage, while SAGA employs a stochastic variational encoder to capture diversity. Both methods exhibit some generalization capability to unseen objects, but only consider motion synthesis until the moment of the grasp.

Concurrently, recent advancements in object-free, text-conditional human motion synthesis [22], [32], [35], [37] have demonstrated promising results, enabled by the decently sized AMASS dataset [21] and its BABEL [24] and HumanML3D [11] annotations. These advancements have paved the way for unprecedented levels of realism and diversity using open-vocabulary, zero-shot generalization. However, these methods are limited to pure human motion generation and do not facilitate human-object scenarios.

To the best of our knowledge, the only method that facilitates whole-body object grasping with text-conditioning is IMoS [10]. Unlike the aforementioned works, IMoS considers post-grasp dynamic object motion, synthesizing it jointly with human motion. The synthesis is conditioned on text, as well as the starting object and body poses. IMoS learns body and arm motion separately using two conditional variational auto-regressors, then optimizes the synthesized object motion to fit the synthesized hand motion. However, they still train their model from scratch using only the small GRAB dataset, limiting its generalization to novel objects and interactions.

In this paper, we propose a novel framework that extends the capabilities of pretrained text-driven human motion diffusion models to tackle the challenging task of full-body 3D hand-object interaction synthesis. Specifically, we freeze the pretrained MDM [32] human motion diffusion model, and introduce additional trainable object-guided attention layers that guide the motion sampling to match the given object motion sequence. In other words, we propose a ControlNet[38] to adapt human motion generators to human-object interaction. By leveraging the pretrained models, our framework can effectively generalize to more diverse interactions. Through comprehensive experiments and evaluations, we demonstrate the effectiveness and robustness of our proposed framework in generating realistic hand-object interactions from textual descriptions. Our approach outperforms IMoS[10] in terms of FID and interaction accuracy metrics. Specifically, it achieves an improvement of 0.08 in FID and 2% in accuracy for body motion synthesis, and 0.15 in FID and 10% in accuracy for hand motion synthesis. These results highlight the significant advancements our framework offers in both the quality and accuracy of generated human-object interactions.

TABLE I: **Comparison of human hand-object interaction methods across four criteria: Full Body, Text-Driven, Till Grasp, and Post Grasp.** Our method and IMoS emphasize the more complex task of post-grasp interaction while SAGA and GOAL focus on human approaches before grasping.

Method	Full Body	Text-Driven	Till Grasp	Post Grasp
GOAL [29]	✓	-	✓	-
SAGA [34]	✓	-	✓	-
IMoS [10]	✓	✓	-	✓
OASIS (ours)	✓	✓	-	✓

Here, we summarize our contributions:

- We propose a framework that enables a pretrained text-conditional human motion diffusion model to perform full-body hand-object interaction synthesis with strong generalization capability.
- We achieve this by freezing the pretrained diffusion model and incorporating additional object-guided attention mechanism layers that adapt the motion denoising (and thus, the sampling) to follow and interact with objects.
- We also propose a hand cVAE to generate detailed hand motion from the generated full-body motion, facilitating fine-grained grasping.

II. Related Work

Text-Driven Human Motion Synthesis. In this task, the synthesis is conditioned on a text prompt, in order to alleviate the constraints posed by the limited number of categories in class-conditional generation [22], [13], [2], by leveraging the compositionality of free-form natural language. Earlier works in this category employed multimodal autoencoders [1], [31] or conditional Variational Autoencoders [12], [23], [2]. More recently, denoising diffusion probabilistic models (DDPMs) [32], [35], [39] have gained attention by fitting the data distribution more accurately without compression, using a stable denoising MSE training objective and a multi-pass sampling procedure for diversity. These methods typically condition their synthesis on semantic, vision-aware text representations from CLIP[26]. Another approach, T2M-GPT [37] adopts a two-stage strategy involving a VQ-VAE model [33] for discretized motion encoding and a transformer decoder for

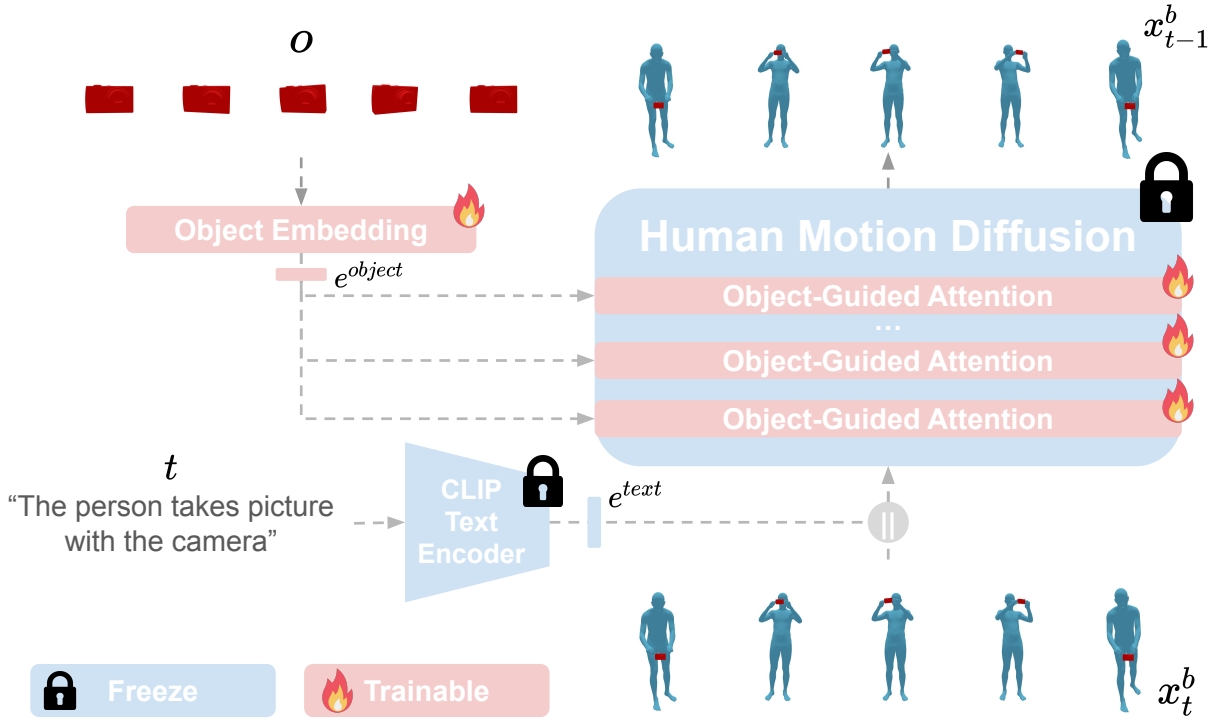


Fig. 3: **Object-guided Human Body Motion Synthesis.** This figure shows our module for human body motion synthesis, where text descriptions t are encoded using a frozen CLIP Text Encoder. Object sequences o are embedded and combined with human motion sequences x_b via object-guided attention layers within a diffusion model. Trainable components are marked with a flame icon, while frozen components are indicated with a lock icon.

autoregressive generation over the discrete motion tokens. However, none of these methods consider human-object interaction. Towards this end, in this paper, we extend frozen pretrained motion diffusion models to facilitate full-body hand-object interaction.

Hand-Object Interaction Generation Synthesizing interactions between objects and human hands is a challenging task due to the diversity and complexity. Some works address this problem by generating static grasp poses for objects and hands [7], [18], [3], [19], while more recent approaches focus on generating dynamic motions of hand-object interactions [41], [40], [36], [4]. For instance, GraspTTA [18] trains a conditional variational autoencoder (CVAE) model [28] conditioned on object point clouds to generate static grasping hand poses, while it also uses a ContactNet to predict the contact map on the object for pose refinement. Text2HOI [4] decomposes interaction generation into two subtasks: mesh-conditional contact generation with a cVAE, and contact-conditional motion generation with a DDPM. While these works focus solely on generating hand grasp poses or hand motions, our approach has a broader scope by generating full-body motion.

Full Body Human Hand-Object Interaction. Text-driven full-body hand-object interaction focuses on synthesizing realistic human motion, either before the human makes contact with the object or after it is grasped [29], [34], [10], [9], [20], [8], [4]. Approaches like GOAL [29] and SAGA [34] generate full-body motion leading up to the point of contact with the object, utilizing CVAEs [28] for motion generation. In contrast, IMoS [10] extends this approach by synthesizing motion sequences occurring after the object is grasped, addressing post-grasping dynamics. However, these methods struggle with limited datasets and less robust generative models, leading to lower-quality and less diverse interactions. Methods like InterFusion [8] aim to synthesize static 3D human-object interactions through a two-stage framework addressing challenges in text-to-3D generation but lack temporal dynamics. Text2HOI [4] focuses on generating hand-object motion using contact maps and diffusion models, but it is limited to hand motions rather than full-body interactions. CG-HOI [9] models human-object motion interdependently with a joint diffusion process guided by explicit contact information but does not leverage pretrained motion models. In comparison, our work addresses these challenges by utilizing a frozen pretrained motion diffusion model with fine-tuned attention layers for object interaction, improving robustness and generalization. Unlike TOHO [20], which generates task-specific human-object interactions using predefined task parameters, our method generalizes to diverse text prompts without requiring task-specific priors.

III. Approach

We define the notations used in this paper. Our dataset, $\mathcal{D}$, consists of N instances of text-body-hand-object quadru-

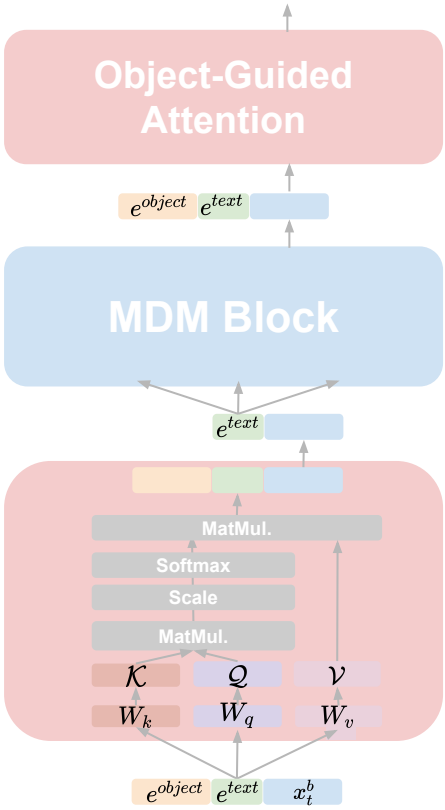


Fig. 4: **Object-Guided Attention.** This figure illustrates the attention mechanism where e^{objet}, e^{text}, and previous human motion state x_t^b are processed. Queries Q, keys K, and values V are generated using weight matrices W_q, W_k, and W_v, respectively. Attention scores are computed, followed by softmax normalization and scaling, to produce the attention output guiding the motion synthesis process.

plets, represented as $\{(text_i, x_i^b, x_i^h, o_i)\}_{i=1}^N$. Here, $text_i$, x_i^b, x_i^h, and o_i denote the text description, human body motion, human hand motion, and object motion for each quadruplet, respectively. Each object pose o and text description $text$ are embedded into embeddings e^{object} and e^{text} via a learnable embedding layer and freezed CLIP text encoder respectively. To achieve more detailed and fine-grained interactions, we separate the interaction synthesis into human body motion synthesis and human hand motion synthesis and optimize them independently. Our primary objective is to generate 3D full-body human poses that realistically interact with the specified object, guided by the provided textual instructions.

A. Object-guided Human Body Motion Synthesis.

Each component of object-guided human body motion synthesis module is illustrated in Fig. 3.

Diffusion Process. We use diffusion model as our base generative model to generate human body motion. Diffusion is modeled as a Markov noising process. During the forward process, noise is incrementally introduced into real data, gradually transforming it into pure Gaussian noise. Conversely, the reverse diffusion process entails iteratively removing noise at each step, resulting in denoised body motion. Specifically, to model the distribution $x^b \sim q(x_0^b)$, the forward diffusion process unfolds as a Markov chain over T steps, yielding a sequence of time-dependent distributions $q(x_t^b|x_{t-1}^b)$. Formally, this process is formulated as follows,

$$q(x_t^b|x_{t-1}^b) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}^b, (1-\alpha_t)I) \tag{1}$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. In our context, text-conditioned motion synthesis models the distribution $p(x_0^b|t)$ as the reversed diffusion process of gradually cleaning x_T^b. Instead of predicting ϵ_t as formulated by [17], we follow [27] and predict the signal itself.

Object-Guided Attention. As shown in Fig. 4, object-guided attention comprises a cross-attention layer that integrates both human motion and object motion as input. This layer enables the model to dynamically adjust its focus during the generation process, allowing it to align the generated human motion with the characteristics and dynamics of the specified object. This dynamic allocation of attention facilitates the synthesis of more contextually relevant and visually coherent human-object interactions. Formally, we first concatenate the motion sequence x^b, the object sequence embedding e^{object}, and the text embedding e^{text} in temporal dimension. Then, attention is applied to the entire sequence, facilitating the interaction between body and object representation.

$$\begin{aligned} \mathcal{Q} &= W_q\{e^{object}, e^{text}, x^b\} \\ \mathcal{K} &= W_k\{e^{object}, e^{text}, x^b\} \\ \mathcal{V} &= W_v\{e^{object}, e^{text}, x^b\} \\ Attention(\mathcal{Q},\mathcal{K},\mathcal{V}) &= softmax(\mathcal{Q}\mathcal{K}^\mathcal{T}/\sqrt{d})\mathcal{V} \end{aligned} \tag{2}$$

A causal attention mask is used to maintain the sequence's causal relationship. An object-guided attention layer is inserted before each of the 8 MDM blocks to control the diffusion model's generation process.

B. Hand Motion Synthesis

We build a hand motion synthesis module based on CVAE [28] as shown in Fig. 5 to synthesize a detailed hand motion conditioned on object motion o, generated human body motion x^b and the text description $text$.

Training. During training, we input hand embedding e_{hand} , object embedding e_{object}, human body motion pose sequence x^b and text embedding e_{text} of the text description $text$ into the motion encoder, along with two learnable distribution parameters, μ and Σ. More specifically, we form e_{hand} by concatenating the hand pose sequence x^h with PointNet [25] embedding of the hand point cloud, and form e_{object} by concatenating the object pose sequence with PointNet [25] embedding of the object point cloud. The CVAE encoder learns to represent these inputs in a latent space by outputting the learned distribution parameters, μ and Σ, which define a Gaussian distribution. A latent vector z is sampled from this distribution and concatenated with the object embedding (e_{object}), the human body motion pose sequence (x^b), and the

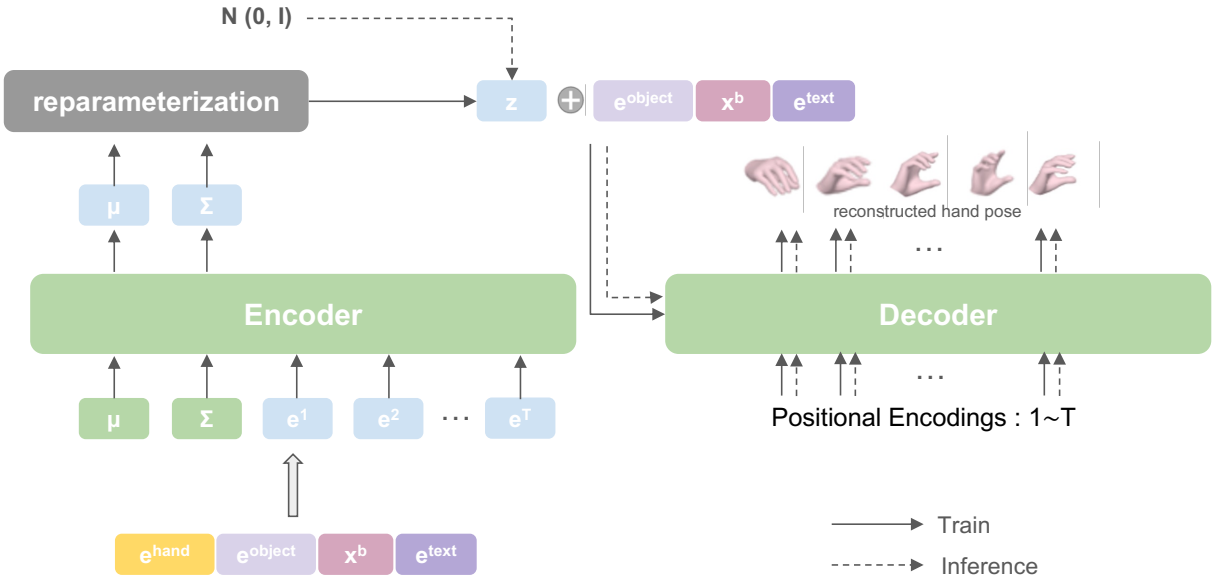


Fig. 5: **Hand Motion Synthesis Module.** The module is built upon a CVAE [28] model with hand embedding e_{hand}, object embedding e_{object}, human body motion pose sequence x^b and text embedding e_{text} of t as input during training. During inference, a latent vector z is sampled from this distribution and concatenated with the object embedding (e_{object}), the human body motion pose sequence (x^b), and the text embedding (e_{text}) to condition the CVAE [28] decoder, which outputs the hand motion that corresponds to the text.

text embedding (e_{text}) to condition the CVAE decoder. We follow [22] by inputting time information as a sinusoidal positional encoding in the decoder, which then outputs the reconstructed hand pose sequence $\hat{x}^h_{\text{recon}}$, ensuring the generated motions are realistic and coherent.

We use two losses for training: the KL divergence loss $\mathcal{L}_{\text{KL}}$ and the reconstruction loss $\mathcal{L}$recon. The KL divergence loss measures the difference between the learned distribution and a standard normal distribution, while the reconstruction loss measures the L1 distance between the reconstructed hand pose $\hat{x}^h_{\text{recon}}$ and the ground truth hand pose x^h. These losses are defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{recon}} &= \|\hat{x}^h_{\text{recon}} - x^h\|_1 \\ \mathcal{L}_{\text{KL}} &= \text{KL}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(0, I)) \\ \mathcal{L}_{\text{train}} &= \lambda\mathcal{L}_{\text{recon}} + (1-\lambda)\mathcal{L}_{\text{KL}} \end{aligned} \tag{3}$$

with $\lambda \in (0, 1)$.

Inference. During inference,we sample z from a standard normal distribution $\mathcal{N}(0, I)$ and concatenate it with the object embedding (e_{object}), the human body motion pose sequence (x^b), and the text embedding (e_{text}). This concatenated vector is then input into the decoder alone with the positional encoding to generate the hand pose sequence $\hat{x}^h$.

Optimization. To ensure a realistic contact between the object and the hand, we follow IMoS [10] and implement an object pose optimization module. From the generated hand pose sequence $\hat{x}^h$, we first identify the initial grasp frame (t = g_0) by analyzing the object's movement across frames and selecting the frame just before a significant movement occurs. We obtain the hand vertices of this frame $V^h_{g_0} = \text{SMPLX}(\hat{x}^h_{t=g_0})$ and calculate the distance between the hand vertices $V^h_{g_0}$ and the object vertices $V^o_{g_0}$ by $D = \text{dist}(V^h_{g_0}, V^o_{g_0})$. To ensure coherent contact between the hand and object during the grasping, we then optimize the object's pose by maintaining this initial distance between the hand and object vertices in subsequent frames. To achieve this, for each subsequent time step t, we optimize R^o_t, T^o_t by minimizing the least squares objective:

$$\begin{aligned} R^{o*}_t, T^{o*}_t &= \min_{R^o_t, T^o_t} \left\|\text{dist}(V^h_t, V^o_t) - D\right\|_2 \\ &= \min_{R^o_t, T^o_t} \left\|\text{dist}(V^h_t, R^o_t \cdot V^o_0 + T^o_t) - D\right\|_2 \end{aligned} \tag{4}$$

with V^o_0 as the object vertices of the initial frame.

IV. Experiments

Implementation Detail. We implement our framework using Pytorch. For the human motion synthesis module, we use the diffusion mdoel pretrained on the HumanML3D dataset released by [32]. We exclusively optimize the newly introduced modules, maintaining the pretrained model in a frozen state. We train our model for 5000 epochs using the Adam with a base learning rate of 1×10^{-4}. Our models have been trained with T = 1000 noising steps and a cosine noise schedule on a single NVIDIA GeForce RTX A4000 GPU for a period of about 8 hours.

For the data representation, we utilize the SMPL-X parametric body model to depict the human pose. SMPL-X characterizes the entire human body, including the hands

TABLE II: **Quantitative comparison with IMoS [10] in body motion synthesis.** Lower FID indicates a closer match to ground truth distribution. Higher accuracy reflects better alignment with specified interactions. For diversity and multimodality metrics, closer alignment with the real data indicates better performance.

Method	FID (↓)	Accuracy (↑)	Diversity (→)	Multimodality (→)
Real Motions (GT)	-	0.99 ± 0.0001	1.21 ± 0.0369	0.23 ± 0.0264
IMoS [10]	0.26 ± 0.0009	0.82 ± 0.0118	1.11 ± 0.0250	0.25 ± 0.0255
OASIS (ours)	$\mathbf{0.18 \pm 0.0046}$	$\mathbf{0.84 \pm 0.0271}$	$\mathbf{1.16 \pm 0.0377}$	$\mathbf{0.22 \pm 0.0314}$

TABLE III: **Quantitative comparison with IMoS [10] in hand motion synthesis.** We compare our hand motion synthesis output, using the human body motion generated by the human motion synthesis module as input, with the hand motions generated by IMoS [10]. Top-1s are highlighted in **bold**.

Method	FID (↓)	Accuracy (↑)	Diversity (→)	Multimodality (→)
GT Hand Motion	-	0.99 ± 0.0001	1.12 ± 0.0202	0.21 ± 0.0057
IMoS	0.57 ± 0.0005	0.73 ± 0.0047	1.05 ± 0.0087	0.23 ± 0.0056
OASIS (ours)	$\mathbf{0.41 \pm 0.0001}$	$\mathbf{0.83 \pm 0.0001}$	$\mathbf{1.07 \pm 0.0104}$	$\mathbf{0.21 \pm 0.0042}$

and face, as a function that can be differentiated. This function is defined by body shape parameters $\beta \in \mathbb{R}^{10}$, root translation $t \in \mathbb{R}^3$, axis-angle rotations for 55 body joints $r \in \mathbb{R}^{55\times3}$, and facial expressions $f \in \mathbb{R}^{10}$. To represent the human body excluding hand, we initially convert SMPL-X to SMPL and then apply the pre-processing pipeline outlined in MDM[32] to extract the body motion representation. Each body pose is described by a composite vector $b_i = (\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^p, j^v, j^r, c^f)$, where $\dot{r}^a \in \mathbb{R}$ denotes the global root angular velocity along the Y-axis, $\dot{r}^x$ and $\dot{r}^z$ represent root linear velocities on the XZ-plane, r^y indicates root height, j^p and j^v in $\mathbb{R}^{3\times22}$, j^r in $\mathbb{R}^{6\times22}$ denote local joints positions, velocities, and rotations in root space, respectively. Binary features $c^f \in \mathbb{R}^4$ are obtained by thresholding heel and toe joint velocities to accentuate foot ground contacts. For human hand representation, we adhere to the pre-processing pipeline utilized in IMoS [10]. Hand poses h_i are characterized by the concatenation of a 6-DOF pose extracted from rotation matrix $r^h \in \mathbb{R}^9$ and translation vector $t^h \in \mathbb{R}^3$. Object pose o_i follows the same method of representing the hand pose.

Dataset. We utilize the GRAB dataset [30], which includes comprehensive grasping sequences performed by ten distinct individuals. These participants interact with 51 diverse objects, each representing one of four fundamental intents: "use," "pass," "lift," and "offhand." The "use" category is further divided into 26 specific actions, illustrating plausible interactions between intent and object, such as drinking from or pouring a cup, taking a picture with a camera, or browsing its functionalities. Following the partitioning strategy outlined in the dataset, we designate subject 'S1' for validation, subject 'S10' for testing, and subjects 'S2' through 'S9' for training our model. This approach ensures that our testing phase includes individuals with unique body shapes, introducing variability akin to real-world scenarios. Notably, our testing also involves novel intent-object pairs that are absent from the training set, such as offhanding a water bottle. Due to inconsistencies in the dataset's motions depicting lifting actions, we exclude the "lift" intention from our experiments, same as IMoS[10]. This results in 789 sequences for training, 157 for validation, and 115 for testing.

Evaluation Metric. Following previous works [10], we use the GRAB dataset to train an action recognition classifier, enabling the computation of several evaluation metrics. These metrics include FID (Fréchet Inception Distance) [16], which measures the discrepancy between the distributions of generated and ground truth data in latent space; Recognition Accuracy, which assesses the consistency between the specified interaction and the generated output; Diversity, which evaluates the breadth and variability within the generated motion distribution; and MultiModality, which quantifies the average variance observed across multiple samples generated from a single text prompt. By leveraging the final layer of the classifier as the motion feature extractor, we can calculate FID, Diversity, and MultiModality, ensuring a comprehensive evaluation of our framework's performance. Following the evaluation process of IMoS [10], we repeat our experiments 20 times and report a statistical interval with 95% confidence.

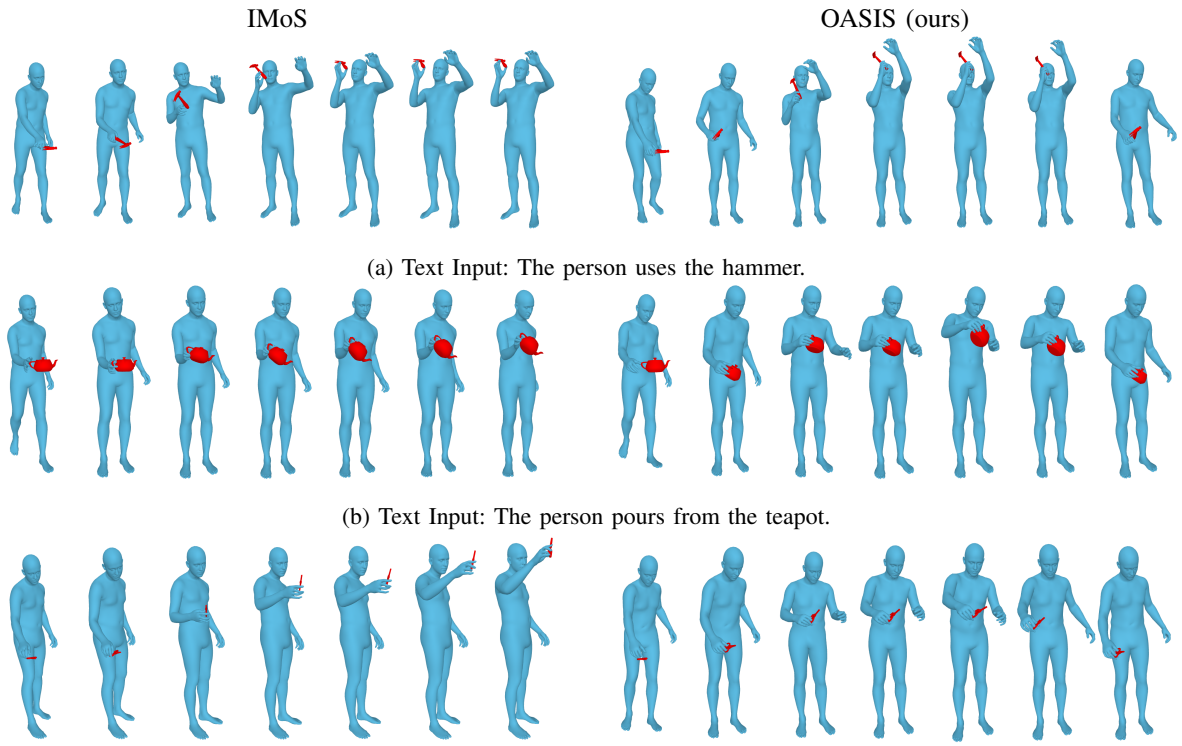
A. Quantitative Comparison

We compare our framework with the previous SOTA method, IMoS [10], which is currently the only approach that focuses on full-body post-grasping human hand-object interaction similar to ours.

Human Body Motion Synthesis. As shown in Table II, our method outperforms the previous SOTA, IMoS [10], across various evaluation metrics. Our framework achieves a lower Fréchet Inception Distance (FID), indicating a closer alignment with the ground truth data distribution. Additionally, our method significantly surpasses IMoS [10] in terms of Recognition Accuracy, demonstrating higher consistency between the specified interactions and the generated outputs.

Our method achieves superior Diversity and Multimodality scores compared to IMoS [10], demonstrating its effectiveness in generating high-quality, varied, and contextually accurate text-guided human-object interactions.

Hand Motion Synthesis. To evaluate the performance of the hand motion synthesis module, we trained a hand motion



(a) Text Input: The person uses the hammer.

(b) Text Input: The person pours from the teapot.

(c) Text Input: The person uses the scissors.

Fig. 6: **Qualitative comparison of full body motion synthesis between IMoS [10] and our work.** Our framework generates more realistic and contextually appropriate motions compared to IMoS. Each row shows sequences of actions based on text instructions (a) "The person uses the hammer," (b) "The person pours from the teapot," and (c) "The person uses the scissors."

action recognition classifier using the GRAB dataset [30], following the same process used for evaluating human body motion synthesis. We compared the hand motion synthesis output of our model, which utilizes the human body motion generated by the human body motion synthesis module as input, with the hand motions produced by IMoS [10]. As shown in Table III, our model outperforms IMoS [10] across all metrics. This demonstrates the robustness and superior performance of our approach in generating high-quality hand motions.

B. Qualitative Comparison

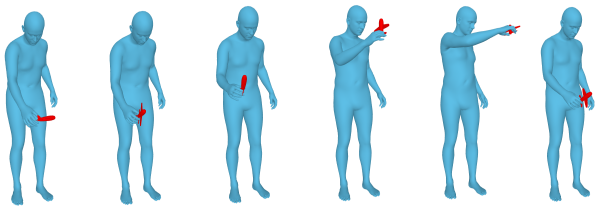
In this section, we present qualitative comparisons between our model and IMoS [10], focusing on human body motion synthesis, hand-object interaction generation, and the integrated results of both.

Human Body Motion Synthesis. Fig. 6 presents a qualitative comparison between IMoS [10] and our model. Across various text instructions, our model consistently generates more realistic and contextually appropriate motions compared to the state-of-the-art IMoS [10]. Our approach particularly excels in capturing the nuances of each interaction, resulting in full-body motions that are not only visually coherent but also exhibit natural transitions between different stages of the interaction. Additional motions involving different objects generated by our model are shown in Fig. 7.

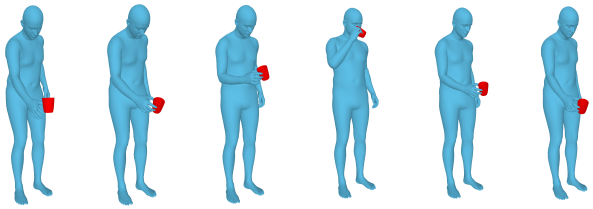
Hand Object Interaction Analysis. In addition to full-body motion synthesis, our framework demonstrates significant improvements in hand-object interaction synthesis. Fig. 8 illustrates the precision and naturalness of hand poses generated by our method compared to the SOTA IMoS [10]. Our approach produces realistic motions that align with the actions specified in the text input. It ensures that the hand's grip and interaction with objects such as a hammer, teapot, and scissors are contextually accurate and realistic, providing better contact and interaction with the objects.

V. Discussion and Limitations

Our OASIS framework shows significant advancements in generating realistic hand-object interactions, surpassing state-of-the-art performance in fidelity and accuracy. However, there are several limitations that outline areas for future work. First, our method is limited to hand-object interactions and does not incorporate interactions involving other body parts, such as the feet, torso, or head. Second, we did not focus on accurately synthesizing the motions of other body parts, such as the feet. While OASIS effectively models hand-object interaction, the motion of other body parts may be less precise, which can affect the overall realism in interactions where multiple body parts are involved. Third, our approach do not address long or complex interactions involving multiple actions or extended sequences with ob-

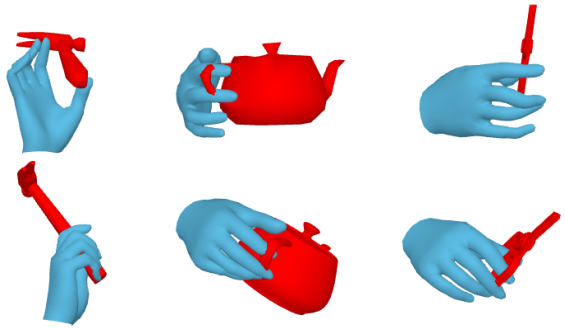


Text Input: The person flies the airplane.



Text Input: The person drinks from the cup.

Fig. 7: **Qualitative results for various text inputs.** Our model demonstrates its ability to generate natural motions that align with the text prompt.



(a) (b) (c)

Fig. 8: **Qualitative comparison of hand-object interaction synthesis between IMoS (upper) [10] and our work (lower).** Each row shows sequences of hand poses based on text instructions: (a) "The person uses the hammer," (b) "The person pours from the teapot," and (c) "The person uses the scissors." Our method demonstrates more accurate and natural hand grips and transitions, enhancing the realism of the interactions compared to IMoS[10].

jects. Additionally, our method does not support texture or fine appearance details, and therefore cannot produce photo-realistic or deceptive content. The results generated by our model are not designed to resemble real-world scenes and cannot be confused with reality. However, future work combining our technique with methods that support more realistic textures might raise ethical concerns, especially if such systems are used to create highly realistic, misleading content.

VI. Conclusion

In conclusion, our paper presents a novel framework that extends diffusion-based text-driven human motion synthesis models to address the complex task of hand-object interaction. By leveraging the generalization capabilities of pretrained models, our framework shows robustness and versatility in generating diverse and realistic interactions from textual descriptions. Our contributions not only advance the state-of-the-art in human-object interaction synthesis but also set the stage for future research in this challenging domain. We believe that our work will inspire further exploration , ultimately leading to the development of more intelligent and capable systems for interacting with the physical world.

VII. Ethical Impact Statement

This work focuses on developing a framework for synthesizing human-object interaction via text instruction. While the ability to generate realistic 3D hand-object interactions from textual descriptions holds significant potential for advancements in human-computer interaction, virtual reality, and robotics, it is important to consider the ethical implications associated with such technologies.

Positive Impacts: Our framework aims to advance the field of human-object interaction modeling, leading to positive impacts in areas such as assistive technology, healthcare, and education. By enabling more intuitive human-computer interaction and facilitating training in virtual environments, our technology has the potential to improve accessibility and inclusivity across various domains.

Data Privacy and Security: Our framework relies on publicly available datasets and does not collect or utilize personal or sensitive data. We recognize the importance of ensuring that datasets used in AI research are ethically sourced, adequately anonymized, and used in compliance with privacy standards. Future applications of this technology must adhere to these principles to prevent privacy violations.

Bias and Representation: The quality and generalizability of our model are inherently tied to the diversity and scale of the training data. For example, if certain demographics or object types are underrepresented in the data, the generated interactions may not generalize well to those cases. To mitigate these risks, we have focused on improving diversity in our training process. However, further research is needed to ensure the inclusion of a broader range of human-object interactions, considering factors such as cultural and social diversity.

Misuse and Unintended Consequences: As with any generative model, there is potential for misuse, such as generating deepfakes or misleading content. While our framework is designed for legitimate applications like human-computer interaction and robotics, malicious actors could repurpose this technology for harmful or deceptive purposes. It is critical for developers, regulators, and policymakers to work together in establishing guidelines and controls to limit the misuse of synthetic content generation systems.

In summary, while this research presents significant advancements, we acknowledge the potential risks and encourage ongoing discourse around the ethical use of generative models. We advocate for the responsible deployment of this technology in ways that prioritize fairness, security, and transparency.

REFERENCES

[1] C. Ahuja and L.-P. Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.

[2] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022.

[3] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019.

[4] J. Cha, J. Kim, J. S. Yoon, and S. Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024.

[5] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] S. Christen, M. Kocabas, E. Aksan, J. Hwangbo, J. Song, and O. Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022.

[7] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020.

[8] S. Dai, W. Li, H. Sun, H. Huang, C. Ma, H. Huang, K. Xu, and R. Hu. Interfusion: Text-driven generation of 3d human-object interaction. In *ECCV*, 2024.

[9] C. Diller and A. Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. 2024.

[10] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023.

[11] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. 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.

[12] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. 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, June 2022.

[13] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[14] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[15] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021.

[19] H. Li, X. Lin, Y. Zhou, X. Li, Y. Huo, J. Chen, and Q. Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv:2210.09245*, 2022.

[20] Q. Li, J. Wang, C. C. Loy, and B. Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024.

[21] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[22] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, October 2021.

[23] M. Petrovich, M. J. Black, and G. Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

[24] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021.

[25] C. R. Qi, H. Su, K. Mo, and L. 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 (CVPR)*, July 2017.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[28] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[29] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.

[31] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[32] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[33] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[34] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang. Saga: Stochastic whole-body grasping with contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[35] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[36] H. Zhang, Y. Ye, T. Shiratori, and T. Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 2021.

[37] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[38] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[39] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[40] J. Zheng, Q. Zheng, L. Fang, Y. Liu, and L. Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023.

[41] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022.