ComGrasp: Component-Aware Whole-Body Grasp Motion Generation via Autoregressive Model

Lin Zhou, Binghui Zuo, Yangang Wang*
School of Automation, Southeast University, Nanjing, China

***Abstract*—Generating long, realistic, and semantically aligned full-body human–object interactions from natural language remains challenging due to complex hand–object contacts, long-range temporal dependencies, and multi-part coordination. Existing methods often struggle to jointly model body motion, hand grasping, and object dynamics. To address these challenges, we propose a language-guided framework for full-body grasp motion generation with explicit contact awareness and part-wise coordination. We introduce a unified representation that encodes SMPL-X body motion, object motion, contact labels, and relative spatial features, and decompose interaction sequences into four semantic components that are discretized using component-specific VQ-VAEs. A masked autoregressive Transformer with a part coordination mechanism is then employed to generate coherent motion tokens conditioned on text. Experiments on the GRAB dataset demonstrate that our method generates longer and more coordinated interaction sequences, outperforming existing approaches.**

***Index Terms*—Text-to-motion synthesis, Whole-body human-object interaction, Discrete motion representation.**

I. Introduction

Generating realistic human motion from natural language descriptions has attracted increasing attention due to its wide applications in animation, virtual reality, robotics, and human-computer interaction. Recent advances in deep generative models have substantially improved the quality and diversity of text-driven human motion synthesis [1]–[5]. However, most existing approaches primarily focus on modeling human motion in isolation, overlooking the rich and dynamic interactions between humans and objects. In real-world scenarios, many everyday actions are inherently interactive and require fine-grained coordination among full-body motion, articulated hand poses, and object dynamics [6], [7].

Whole-body HOI synthesis presents unique challenges beyond general motion generation, requiring unified modeling of the full body, hands, and object dynamics [8]. Achieving realism further demands precise hand-object contact [9] and effective long-term temporal coordination [10]. However, existing methods often simplify these tasks by assuming known object trajectories [11], ignoring full-body context [12], or relying on fixed temporal constraints [13], limiting their applicability in unconstrained settings.

To address these limitations, we propose a language-conditioned framework for synthesizing long, contact-aware, and coordinated whole-body grasp motions. Our approach combines a unified representation with component-aware discretization and coordinated autoregressive generation, building upon recent advances in discrete motion modeling [4], [14].

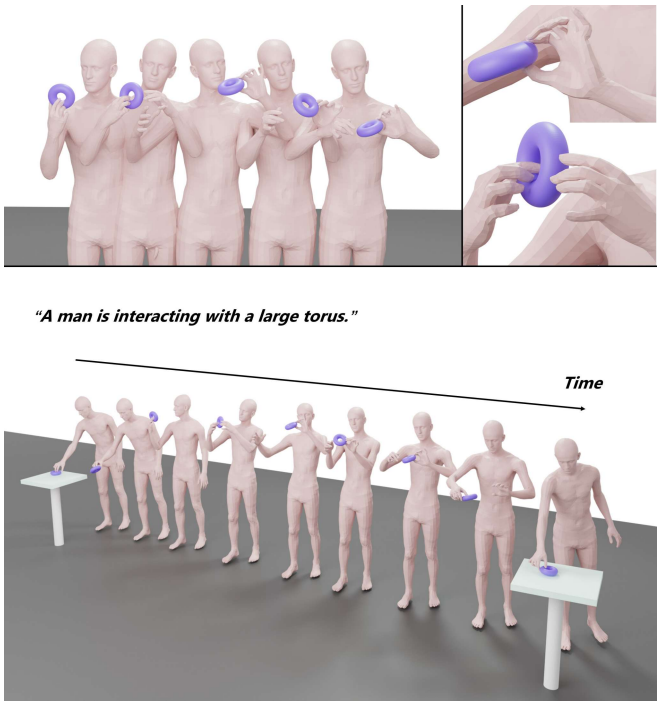


Fig. 1. Language-guided whole-body grasp motion generation. Given a natural language description, our method synthesizes a long and coherent whole-body human-object interaction sequence.

Our contributions are threefold: **Unified HOI representation.** We introduce a unified representation that jointly encodes SMPL-X body motion, object motion, hand-object contact labels, and relative spatial features, providing a comprehensive description of HOI sequences [6], [8]. **Component-aware discrete tokenization.** We decompose the interaction into four semantic components (global motion, body posture, left-hand interaction, and right-hand interaction) and discretize each component using a dedicated VQ-VAE, yielding structured token sequences while preserving motion fidelity [2], [15]. **Coordinated autoregressive generation.** We propose a coordinated autoregressive Transformer that generates discrete motion tokens from text and enforces spatiotemporal consistency across components via a lightweight coordination mechanism.

*Corresponding author: Yangang Wang, e-mail: yangangwang@seu.edu.cn.
This work was supported in part by the National Natural Science Foundation of China (No. 62076061) and the Natural Science Foundation of Jiangsu Province (No. BK20220127).

Experiments on GRAB [6] demonstrate that our approach outperforms state-of-the-art methods in generating longer and more contact-plausible interaction sequences.

II. Related Work

Human-Object Interaction. Human-object interaction generation is a long-standing problem, aiming to capture the complex relationships between humans and objects. The mainstream research in this field can be broadly categorized into regression-based methods and generation-based methods. Regression-based methods mostly estimate the SMPL parameters [16] through neural networks. These approaches [17]–[19] are always designed to reconstruct human-object interactions from input images or videos. Generation-based paradigm [13], [20]–[24], on the other hand, aim to synthesize realistic human-object interactions by modeling the underlying distributions of interaction patterns. These approaches often leverage generative models, such as variational auto-encoders [15] and generative adversarial networks [25], to produce diverse interaction scenarios. With the proposal of the diffusion model [26], [27], a lot of works have applied it to generate high-quality human-object interactions. By integrating various control signals, these methods strive to create more plausible and physically accurate results. Despite significant progress, challenges remain in ensuring temporal coherence in dynamic settings and estimating accurate grasp poses.

Hand-Object Interaction. Compared to general human-object interaction, hand-object interaction focuses more on the detailed modeling of hand poses. With the emergence of benchmarks [28]–[34], research on hand-object interaction has progressed from static to dynamic grasps. For static grasps generation, most methods [9], [35], [36] severed the given object point clouds as conditions and synthesized the plausible hand grasp poses. Although they enhanced the improvement of diversity and physical plausibility, these methods ignored the temporal correlation of grasp poses in real-world scenarios. To address this issue, recent works [12], [37]–[39] paid more attention to the hand grasp sequence generation. D-Grasp [12] applied reinforcement learning to generate hand grasping motions in a simulation environment. Text2HOI [37] employed cascaded diffusion to iteratively generate hand motions from the text prompts. LatentHOI [40] emphasized the importance of modeling fine-grained spatial hand-object interactions and designed a latent diffusion model coupled with a Grasp VAE.

Whole-Body Interaction. It should be noted that whole-body interaction is not a simple fusion of human- and hand-object interaction approaches. It requires not only ensuring the coordination of bi-manual hand poses, but also guaranteeing the smoothness and fluency of body actions. Based on this fact, adopting the parametric SMPL-X model [8] becomes a consensus for accurately representing the full-body poses and hand configurations. With the support of GRAB dataset [6], IMOS [10] separately synthesized the body and arm motions through two different modules. GOAL [7] inferred the entire motions to the goal, which is formulated as a static target body pose. DiffGrasp [11] assumed the known object trajectories and generated the whole-body grasp motions through diffusion models. With the success of auto-regressive models in various sequence generation tasks, some works [1], [2], [4], [5], [14] have explored their potential in human motion generation and inpainting tasks. Attracted by its generativity and editability, in this paper, we introduce it to the whole-body grasp generation task, thereby exploring a meaningful paradigm for this field.

III. Method

A. Problem Definition

Given a natural language description $\mathcal{T}$ containing information on gender, action, and object category, our goal is to generate a sequence of human-object interaction (HOI) motions containing T frames. To maintain a realistic full-body grasping motion, each generated frame should consider the coordination between the human body, hands, and the rigid object. Therefore, the output is a temporally coherent sequence where each frame provides a complete parametric description of the current human and object states.

B. Data Representation

Human Motion. We use the SMPL-X model to parametrically represent the human body, with the pelvis as the root node. The motion parameters include: root rotation $\mathbf{R}_{\text{root}} \in \mathbb{R}^6$ (6D rotation representation), root translation $\mathbf{T}_{\text{root}} \in \mathbb{R}^3$, body pose $\boldsymbol{\theta}_{\text{body}} \in \mathbb{R}^{126}$ (21 joints with 6D representation), left hand pose $\boldsymbol{\theta}_{\text{left}} \in \mathbb{R}^{90}$ (15 joints with 6D representation), and right hand pose $\boldsymbol{\theta}_{\text{right}} \in \mathbb{R}^{90}$ (15 joints with 6D representation).

Object Motion. We use canonical object meshes $\mathbf{V}^{\text{obj}}_{\text{canonical}}$ provided in the GRAB dataset [6] to represent object shapes. At frame t, the object motion is parameterized by rotation $\mathbf{R}^t_{\text{obj}} \in \mathbb{R}^6$ and translation $\mathbf{T}^t_{\text{obj}} \in \mathbb{R}^3$. We additionally define the object center $\mathbf{C}^t_{\text{obj}} \in \mathbb{R}^3$ in the global coordinate system. Accordingly, object vertices are computed as $\mathbf{V}^t_{\text{obj}} = \mathbf{R}^t_{\text{obj}}\mathbf{V}^{\text{obj}}_{\text{canonical}} + \mathbf{T}^t_{\text{obj}}$.

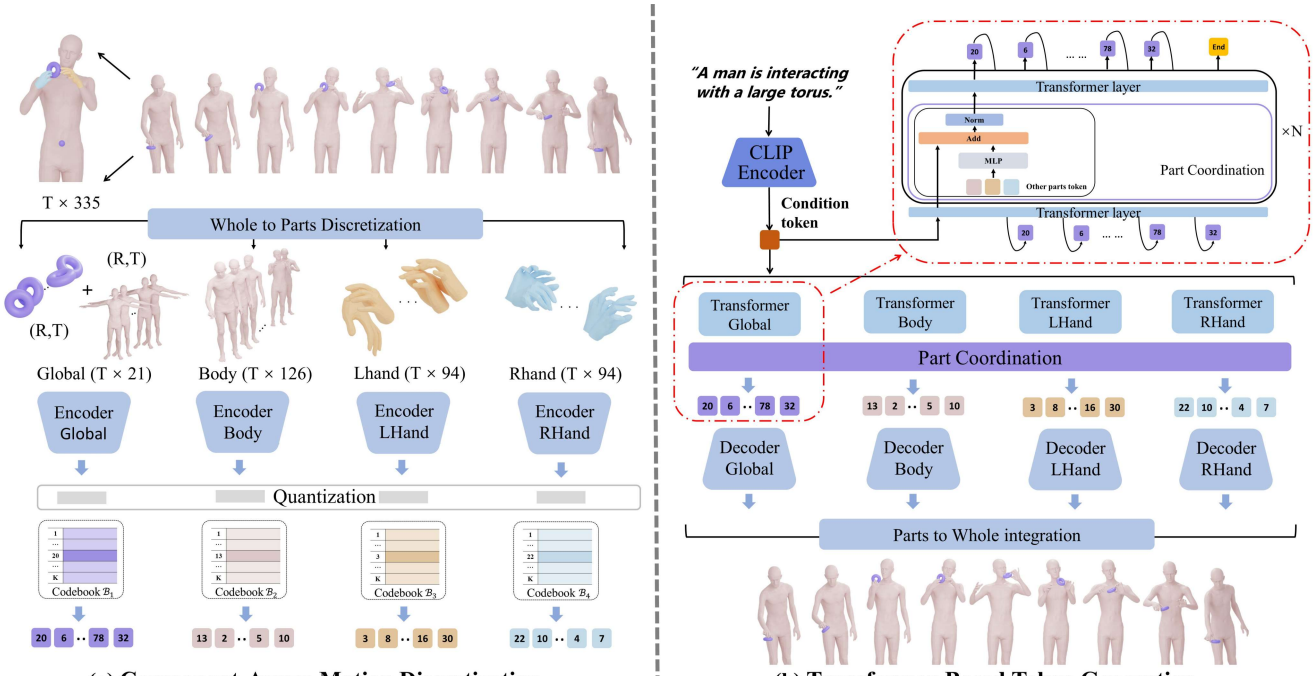
Contact awareness. *Contact awareness* refers to explicitly encoding and generating per-frame hand-object contact states. We derive binary contact labels for both hands from the GRAB frame-wise human-object vertex contact map. Specifically, for each hand (left or right), the label is set to 1 if any vertex in that hand region is in contact with the object, and 0 otherwise. This yields a contact state vector $\mathbf{c}_t = [c^t_l, c^t_r] \in \{0, 1\}^2$ at each frame t. We incorporate these labels into the hand-interaction components together with relative hand-object spatial features to enforce contact-aware generation.

Relative Position. To strengthen the association between the human and the object, we additionally store the relative wrist positions to the object center, denoted as $\mathbf{p}^t_{\text{rel,l}}, \mathbf{p}^t_{\text{rel,r}} \in \mathbb{R}^3$.

Overall Representation. For each frame t, we preprocess all motion-related features into a unified representation $\mathbf{x}_t \in \mathbb{R}^{335}$. Specifically, the state vector $\mathbf{x}_t$ consists of global motion parameters for the human root and the object, body pose parameters, and hand-related interaction features:

$$\mathbf{x}_t = [\mathbf{R}_{\text{root}}, \mathbf{T}_{\text{root}}, \mathbf{R}_{\text{obj}}, \mathbf{T}_{\text{obj}}, \mathbf{C}_{\text{obj}}, \boldsymbol{\theta}_{\text{body}}, \mathbf{x}_{\text{hands}}].$$

$$\text{where} \quad \mathbf{x}_{\text{hands}} = [\boldsymbol{\theta}_{\text{left}}, \mathbf{p}^t_{\text{rel,l}}, c^t_{\text{l}}, \boldsymbol{\theta}_{\text{right}}, \mathbf{p}^t_{\text{rel,r}}, c^t_{\text{r}}].$$



(a) Component-Aware Motion Discretization **(b) Transformer-Based Token Generation**

Fig. 2. **ComGrasp pipeline overview. (a) Component-aware motion discretization.** The 335-D HOI representation is decomposed into four semantic components (Global, Body, LHand-Interact, RHand-Interact) and discretized by component-specific VQ-VAEs to obtain structured token sequences. **(b) Transformer-based token generation.** Text is encoded by CLIP to a condition token. Four parallel autoregressive Transformers generate tokens for each component, while a part coordination module enables cross-component information exchange between Transformer layers. The generated tokens are decoded by the corresponding VQ-VAE decoders and integrated to reconstruct the whole-body grasp motion sequence.

C. Component-Aware Motion Discretization

To obtain discrete representations that preserve fine-grained motion details, we propose a component-aware discretization strategy. Inspired by [14], we decompose HOI motion into semantically independent components, each discretized by a dedicated VQ-VAE. This compositional design facilitates effective modeling of complex interactions.

Motion Decomposition and Formalization. Given an HOI motion sequence $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, we decompose it into four components $\{\mathbf{P}^g, \mathbf{P}^b, \mathbf{P}^l, \mathbf{P}^r\}$. The Global component $\mathbf{P}^g$ captures the absolute motion of the human root and the object $(\mathbf{R}_{\text{root}}, \mathbf{T}_{\text{root}}, \mathbf{R}_{\text{obj}}, \mathbf{T}_{\text{obj}}, \mathbf{C}_{\text{obj}})$, the Body component $\mathbf{P}^b$ encodes the articulated body pose $\boldsymbol{\theta}_{\text{body}}$, and the Left/Right Hand + Interaction components $\mathbf{P}^l$ and $\mathbf{P}^r$ model hand poses $(\boldsymbol{\theta}_{\text{left}}, \boldsymbol{\theta}_{\text{right}})$ augmented with wrist-to-object relative positions $(\mathbf{p}^t_{\text{rel,l}}, \mathbf{p}^t_{\text{rel,r}})$ and binary contact labels $(c^t_{\text{l}}, c^t_{\text{r}})$.

Multi-Component VQ-VAE Discretization. For each motion component $\mathbf{P}^i$ $(i \in \{g, b, l, r\})$, as illustrated in Fig. 2(a), we discretize the continuous motion sequence into a component-specific discrete token sequence using an independent VQ-VAE, which serves as our motion tokenizer. The encoder downsamples the sequence using a stack of 1D convolutions and cross-frame self-attention layers to map $\mathbf{P}^i$ to a latent sequence

$$\mathbf{E}^i = \text{Encoder}^i(\mathbf{P}^i) = \{\mathbf{e}^i_1, \ldots, \mathbf{e}^i_L\}, \tag{1}$$

where $L = T/l$ is the token sequence length after temporal downsampling. Here, $l = 2^{down\_t}$ denotes the temporal downsampling factor. We use a moderate downsampling rate and component-specific codebook sizes to balance long-horizon modeling capabilities with motion reconstruction fidelity (see the Default configuration in Table III).

Each latent vector $\mathbf{e}^i_l$ is quantized by assigning it to the nearest entry in a learnable codebook $\mathcal{B}^i = \{\mathbf{b}^i_j \mid j = 1, \ldots, J_i\}$, yielding the discrete token sequence for component i.

$$\hat{\mathbf{e}}^i_l = \arg\min_j \left\| \mathbf{e}^i_l - \mathbf{b}^i_j \right\|, \tag{2}$$

where J denotes elements in the codebook. By this way, we quantize the continuous representation $\mathbf{E}^i$ into discrete notations $\hat{\mathbf{E}}^i = \{\hat{\mathbf{e}}^i_1, \ldots, \hat{\mathbf{e}}^i_L\}$, which are then decoded by the decoder Decoder^i to reconstruct the corresponding component motion parameters $\hat{\mathbf{P}}^i$. Finally, by integrating the parameters of all components and inputting them into the SMPL-X model and the object model respectively, we can reconstruct the complete human-object interaction motion sequence.

To learn these discrete representations effectively, we train the VQ-VAE by minimizing a composite loss function $\mathcal{L}^i$, which balances reconstruction fidelity and codebook commitment:

$$\mathcal{L}^i = \mathcal{L}^i_{\text{rec}} + \underbrace{\left\| \text{sg}(\mathbf{E}^i) - \hat{\mathbf{E}}^i \right\| + \beta \left\| \mathbf{E}^i - \text{sg}(\hat{\mathbf{E}}^i) \right\|}_{\mathcal{L}_{\text{commit}}}, \tag{3}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation and β is a hyperparameter for the commitment loss $\mathcal{L}_{\text{commit}}$. The reconstruction term $\mathbf{L}^i_{\text{rec}}$ is defined as:

$$\begin{aligned}\mathcal{L}^i_{\text{rec}} = \|\theta - \hat{\theta}\|_1 + \lambda_{\text{joint}} \left\| J_{3D} - \hat{J}_{3D} \right\|_1 \\ + \lambda_{vel.} \left\| (\theta_{i,t+1} - \theta_{i,t}) - (\hat{\theta}_{i,t+1} - \hat{\theta}_{i,t}) \right\|_1, \\ + \mathbb{I}_{i\in\{l,r\}} \lambda_{\text{cont}} \mathrm{BCE}(\mathbf{c}^i, \hat{\mathbf{c}}^i)\end{aligned} \tag{4}$$

which combines parameter reconstruction, joint position reconstruction (computed via SMPL-X), a velocity regularizer, and a specific binary cross-entropy loss for contact labels when training hand interaction components ($i \in \{l, r\}$).

D. Transformer-Based Token Generation

After discretization, each motion sequence is represented by four token index sequences $\mathcal{S} = \{\mathbf{s}^i\}$ with $i \in \{g, b, l, r\}$, where $\mathbf{s}^i = \{s^i_1, \ldots, s^i_L\}$. Our goal is to autoregressively generate all component token streams conditioned on text $\mathcal{T}$, and decode the predicted tokens to reconstruct whole-body grasp motions. The decoded hand-interaction component includes hand poses as well as the contact labels and relative hand–object positions. As shown in Fig. 2(b), we employ four component-wise autoregressive Transformer generators.

Text Conditioning and Sequence Formatting. The text conditions are derived from the GRAB dataset filenames, which contain gender, action, and object information. We generate natural language descriptions by instantiating multiple predefined templates, like "A man is taking a picture with a camera". These text descriptions are then encoded by CLIP [41] to obtain the text embedding c, which is injected as a condition token shared across all component streams. To enable variable-length generation, we append a special learnable `[End]` token to the end of each motion token sequence. During inference, the autoregressive generation for a component terminates once this `[End]` token is predicted. The token sequence length for each sequence is L. For efficient batch processing, we use a special `[Padding]` flag to unify all target tags to a maximum length of $L_{\max}$.

Part Coordination Attention Mechanism. We introduce a *Part Coordination Mechanism* [14] before each Transformer layer except the first one to enable cross-component information exchange. For component i at time step t, the hidden state $\mathbf{h}^t_i$ is updated by aggregating states from other components through a shared MLP:

$$\mathbf{h}^t_{\text{coord},i} = \mathrm{LayerNorm}\big(\mathbf{h}^t_i + \mathrm{MLP}\big(\{\mathbf{h}^t_j\}_{j\neq i}\big)\big). \tag{5}$$

This bidirectional communication enforces spatiotemporal consistency across components and yields coordinated whole-body motion generation.

Training Objective. We optimize the negative log-likelihood of all component streams conditioned on c:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} = \mathbb{E}_{\mathcal{S}\sim p(\mathcal{S})}\left[- \sum_{i\in\{g,b,l,r\}} \log p\left(\mathbf{s}^i \mid c\right) \right]. \tag{6}$$

Since the VQ-VAE tokenizer is trained to preserve contact information within the discrete tokens, optimizing the likelihood of the ground-truth token sequence effectively enforces contact-aware generation.

IV. Experiments

A. Implementation Details

All models were implemented in PyTorch and trained on a single NVIDIA RTX 4070 Ti GPU.

VQ-VAE Tokenizers. We adopt the default component-wise codebook configuration summarized in Table III. Each VQ-VAE tokenizer is trained using AdamW with EMA-reset quantization and a batch size of 64. The commitment loss weight β is set to 0.05. For the reconstruction loss terms, we set $\lambda_{\text{joint}} = 1.0$, $\lambda_{\text{vel}} = 0.5$, and $\lambda_{\text{cont}} = 0.5$. Unless otherwise specified, we set the temporal downsampling exponent to *down_t* $= 4$, resulting in an overall downsampling factor of $l = 16$, and fix the maximum token length to $L_{\max} = 74$ for all components.

Autoregressive Transformers. We train four component-wise Transformers with embedding dimensions 512 (global), 1024 (body), and 768 (hand), coordinated by a shared 3-layer MLP. The vocabulary size for each Transformer is set to $L_{\max}$ plus two special tokens (one for the conditioning text embedding and one for the `[End]` token). We optimize with AdamW using 14 Transformer layers, 16 heads, dropout 0.1, and a block size of 76, with batch size 64.

B. Datasets and Metrics

Dataset. We conduct experiments on the GRAB dataset [6], which collects diverse whole-body human-object interaction sequences with detailed hand grasps. The dataset consists of 10 subjects and provides SMPL-X parameters for the human body and 6-DoF object poses. Additionally, textual descriptions of each interaction are provided based on the folder name in the dataset. We follow the standard split as in prior works [6], who divides 8 subjects for training and 2 subjects for testing.

Metrics. We evaluate generated HOI sequences using metrics grouped into three categories (Table I), following the same evaluation protocol as DiffGrasp.

(1) Motion Accuracy: We report *H-MPJPE* and *MPJPE* to measure 3D joint errors for hands and the full body, respectively, and *MPVPE* to evaluate vertex-level mesh reconstruction accuracy. All motion errors are computed frame-wise between the generated sequence and the corresponding ground-truth sequence within the same clip.

(2) Physical Plausibility: Physical realism is assessed using *Foot Sliding (FS)* and collision-related metrics, including *Collision Percentage (Coll)* and *Collision Depth (C-Depth)*. Following DiffGrasp, collisions are detected based on the object's signed distance field (SDF), where a hand vertex is considered in collision if it penetrates the object surface with a penetration depth below 5 mm. *Coll* measures the fraction of collided hand vertices, and *C-Depth* reports the mean penetration depth.

TABLE I
COMPARATIVE EXPERIMENTAL RESULTS ON THE GRAB DATASET.

Method	Motion Accuracy ↓			Physical Plausibility ↓			Contact Quality	
	H-MPJPE	MPJPE	MPVPE	FS	Coll.	C-Depth	F1 ↑	Cont. dist ↓
DiffGrasp	<u>20.99</u>	<u>12.28</u>	<u>10.09</u>	<u>2.22</u>	0.0023	**0.0001**	**0.7732**	<u>0.04</u>
IMOS	23.73	15.20	11.35	2.82	**0.0012**	<u>0.0002</u>	0.5570	0.09
GOAL	21.93	14.10	10.96	2.53	0.0030	<u>0.0002</u>	0.6865	0.05
Ours	**20.60**	**12.16**	**9.25**	**1.92**	<u>0.0014</u>	<u>0.0002</u>	<u>0.7710</u>	**0.03**

[a]**Bold** denotes the best value, and <u>underline</u> denotes the second best.

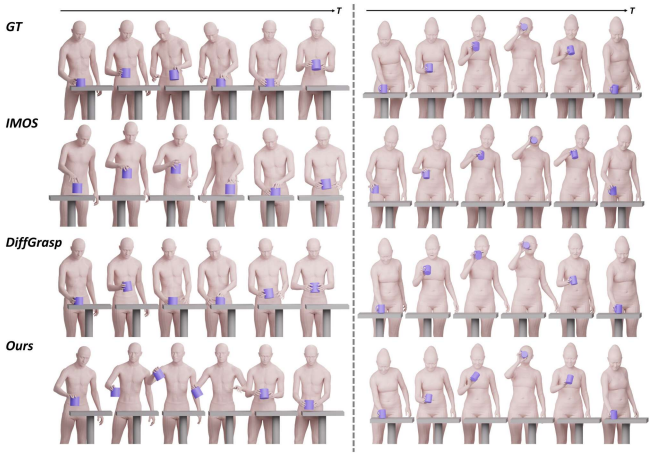


Fig. 3. **Qualitative comparison on the GRAB dataset.** We compare our method with ground truth (GT), IMOS, and DiffGrasp on representative whole-body grasping sequences.

(3) Contact Quality: Hand-object interaction quality is evaluated using the *F1 score*. Following DiffGrasp, contact labels for both generated and ground-truth sequences are extracted by thresholding the minimum *Contact Distance (Cont. dist)* between hand and object vertices at 5 mm, and the F1 score is computed from the resulting precision and recall.

C. Comparison with SOTA Methods

We perform qualitative comparisons with DiffGrasp [11], IMOS [10], and GOAL [7] on the GRAB dataset in Fig. 3, and quantitatively evaluate the methods using the predefined metrics, with results summarized in Table I. Overall, our method achieves consistently strong performance in both quantitative and qualitative experiments.

In terms of **motion accuracy**, our method achieves the best results on H-MPJPE, MPJPE, and MPVPE, indicating more accurate joint-level motion reconstruction and improved mesh-level fidelity compared to all baselines.

Regarding **physical plausibility**, our approach yields the lowest foot sliding (FS), demonstrating more stable and physically consistent long-horizon motion. While IMOS achieves the lowest collision percentage, it often avoids contact altogether, leading to inferior interaction quality. In contrast, our method maintains a low collision rate while preserving meaningful hand-object interactions.

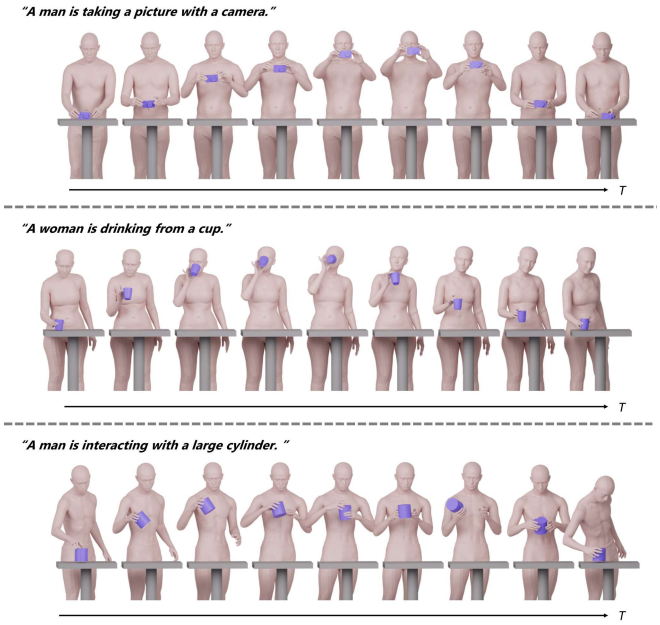


Fig. 4. **More results under different text conditions.** Given diverse natural language descriptions, our method generates long and coherent whole-body human-object interaction sequences.

For **contact quality**, our method achieves the lowest contact distance and the second-best F1 score, reflecting accurate and stable hand-object contacts. Compared to DiffGrasp, which attains the highest F1 score but exhibits larger motion errors and foot sliding, our method provides a better overall balance between motion accuracy, physical realism, and contact correctness.

These results demonstrate that explicitly modeling and coordinating multi-component motions enables more accurate, physically plausible, and contact-aware whole-body human-object interaction generation. Additionally, more qualitative results under different text conditions are presented in Fig. 4.

D. Ablation Studies

We conducted an ablation study on the GRAB dataset to evaluate the effectiveness of key design choices in our framework, focusing on **architecture of tokenizer**, **VQ-VAE codebook capacity**, and **temporal downsampling**. All vari-

TABLE II
ABLATION STUDY RESULTS ON THE GRAB DATASET.

Methods	H-MPJPE↓	MPJPE↓	MPVPE↓	FS↓	Coll.↓	C-Depth↓	F1↑	Cont. dist↓
Architecture of Tokenizer								
single VQ-VAE	31.01	20.28	16.23	2.18	0.0042	0.0013	0.5119	0.13
Codebook Capacity								
Small Codebook	23.31	16.16	11.09	2.07	0.0017	0.0009	0.6211	0.05
Big Codebook	20.71	**12.03**	9.33	1.93	0.0028	0.0003	**0.7519**	**0.03**
Temporal Downsampling								
$down_t = 2$	**20.45**	13.22	10.21	2.34	0.0041	0.0012	0.7420	**0.03**
$down_t = 8$	28.63	14.52	11.36	3.08	0.0042	0.0008	0.5113	0.07
Ours (Full)	20.60	12.16	**9.25**	**1.92**	**0.0014**	**0.0002**	0.7710	**0.03**

[a]**Bold** denotes the best value, and underline denotes the second best.

TABLE III
CODEBOOK CONFIGURATION (CODEBOOK SIZE × CODE DIMENSION) FOR DIFFERENT CAPACITY SETTINGS.

Configuration	Global	Body	Hand-Interact
Small Codebook	256 × 128	512 × 256	256 × 192
Default Codebook	512 × 128	1024 × 256	512 × 192
Big Codebook	1024 × 128	2048 × 256	1024 × 192

ants followed the same training and evaluation protocol as the main comparison, and their quantitative evaluation results are shown in Table II.

Architecture of Tokenizer. We compare our component-aware tokenizer with a unified baseline that encodes the entire 335-D interaction feature using a single VQ-VAE. As shown in Table II, the unified design leads to a substantial performance degradation across all metrics, with particularly large errors in H-MPJPE and MPVPE, as well as a significantly lower contact F1 score. This indicates that jointly discretizing heterogeneous motion signals causes representational aliasing, where dominant global motion patterns suppress fine-grained hand dynamics. By contrast, allocating separate discrete spaces for global motion, body posture, and hand–object interaction better accommodates their distinct motion scales, resulting in improved motion accuracy and more consistent contact modeling.

Codebook Capacity. We study the effect of discrete capacity by comparing the *Small* and *Large* codebook settings against the *default* configuration in Table III. Reducing the codebook capacity consistently degrades both motion accuracy and interaction quality, as evidenced by higher joint and vertex errors and a reduced contact F1 score. Increasing the codebook capacity improves motion reconstruction accuracy and contact quality, but also leads to higher collision rates and substantially increased computational cost. Overall, the *default* codebook achieves the best balance between motion fidelity, contact quality, and efficiency, and is therefore adopted in our final model.

Temporal Downsampling. We evaluate different temporal downsampling rates to analyze the trade-off between temporal resolution and long-term stability. With a smaller downsampling factor ($down\_t = 2$), the model preserves more high-frequency details but suffers from increased foot sliding and higher collision rates. Conversely, excessive downsampling ($down\_t = 8$) leads to significant performance degradation across all metrics, including motion accuracy and contact quality, due to over-compression of temporal dynamics. The default setting ($down\_t = 4$) achieves the best overall performance, yielding the lowest foot sliding, strong motion accuracy, and stable hand–object contacts, and is therefore used in our final model.

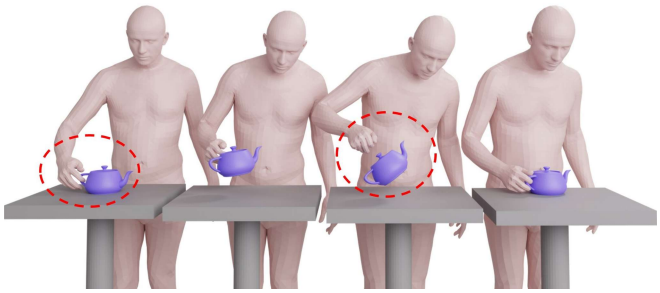


Fig. 5. **Failure case on strict-contact objects.** For strict-contact objects (e.g., teapot handle), contact may drift and drop in later frames (highlighted in red).

V. CONCLUSION

We propose **ComGrasp**, a language-driven framework for generating long-horizon full-body grasping motions. By coordinating componentized motion structures, ComGrasp achieves superior coherence and interaction quality. Experiments on GRAB demonstrate that it consistently outperforms prior methods in physical plausibility and contact accuracy.

Limitations. ComGrasp may exhibit contact drift during long-horizon interactions with constrained objects (e.g., teapot handles, Fig. 5), likely due to insufficient contact stability constraints. Future work will explore incorporating geometry- or physics-aware priors to address this issue.

REFERENCES

[1] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 5152–5161.

[2] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and Y. Shan, "Generating human motion from textual descriptions with discrete representations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 14 730–14 740.

[3] 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. [Online]. Available: https://openreview.net/forum?id=SJ1kSyO2jwu

[4] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 1900–1910.

[5] E. Pinyoanuntapong, P. Wang, M. Lee, and C. Chen, "Mmm: Generative masked motion model," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 1546–1555.

[6] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 581–600.

[7] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "Goal: Generating 4d whole-body motion for hand-object grasping," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 13 263–13 273.

[8] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 10 975–10 985.

[9] S. Liu, Y. Zhou, J. Yang, S. Gupta, and S. Wang, "Contactgen: Generative contact modeling for grasp generation," in *Int. Conf. Comput. Vis.*, 2023, pp. 20 609–20 620.

[10] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "Imos: Intent-driven full-body motion synthesis for human-object interactions," in *Computer Graphics Forum*, vol. 42, no. 2. Wiley Online Library, 2023, pp. 1–12.

[11] Y. Zhang, Q. He, Y. Wan, Y. Zhang, X. Deng, C. Ma, and H. Wang, "Diffgrasp: Whole-body grasping synthesis guided by object motion using a diffusion model," in *AAAI*, vol. 39, no. 10, 2025, pp. 10 320–10 328.

[12] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 20 577–20 586.

[13] X. Peng, Y. Xie, Z. Wu, V. Jampani, D. Sun, and H. Jiang, "Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 2878–2888.

[14] Q. Zou, S. Yuan, S. Du, Y. Wang, C. Liu, Y. Xu, J. Chen, and X. Ji, "Parco: Part-coordinating text-to-motion synthesis," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 126–143.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[17] Z. Weng and S. Yeung, "Holistic 3d human and scene mesh estimation from single view images," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 334–343.

[18] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "Chore: Contact, human and object reconstruction from a single rgb image," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 125–145.

[19] H. Nam, D. S. Jung, G. Moon, and K. M. Lee, "Joint reconstruction of 3d human and object via contact-based refinement transformer," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 10 218–10 227.

[20] S. Xu, Z. Li, Y.-X. Wang, and L.-Y. Gui, "Interdiff: Generating 3d human-object interactions with physics-informed diffusion," in *Int. Conf. Comput. Vis.*, 2023, pp. 14 928–14 940.

[21] C. Diller and A. Dai, "Cg-hoi: Contact-guided 3d human-object interaction generation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 888–19 901.

[22] S. Xu, Y.-X. Wang, L. Gui *et al.*, "Interdreamer: Zero-shot text to 3d dynamic human-object interaction," *Adv. Neural Inform. Process. Syst.*, vol. 37, pp. 52 858–52 890, 2024.

[23] H. Kim, S. Baik, and H. Joo, "David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models," in *Int. Conf. Comput. Vis.*, 2025, pp. 10 330–10 341.

[24] C. Zhu, B. Huang, Z. Wu, B. Zuo, and Y. Wang, "E-react: Towards emotionally controlled synthesis of human reactions," *arXiv preprint arXiv:2508.06093*, 2025.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Adv. Neural Inform. Process. Syst.*, vol. 27, 2014.

[26] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 10 684–10 695.

[28] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 11 807–11 816.

[29] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3196–3206.

[30] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 9044–9053.

[31] Z. Zhao, B. Zuo, W. Xie, and Y. Wang, "Stability-driven contact reconstruction from monocular color images," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 1643–1653.

[32] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 20 953–20 962.

[33] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 013–21 022.

[35] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Int. Conf. Comput. Vis.*, 2021, pp. 11 107–11 116.

[36] B. Zuo, Z. Zhao, W. Sun, X. Yuan, Z. Yu, and Y. Wang, "Graspdiff: Grasping generation for hand-object interaction with multimodal guided diffusion," *IEEE Trans. Vis. Comput. Graph.*, 2024.

[37] J. Cha, J. Kim, J. S. Yoon, and S. Baek, "Text2hoi: Text-guided 3d motion generation for hand-object interaction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 1577–1585.

[38] J. Zheng, Q. Zheng, L. Fang, Y. Liu, and L. Yi, "Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 585–594.

[39] X. Liu and L. Yi, "Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion," *arXiv preprint arXiv:2402.14810*, 2024.

[40] M. Li, S. Christen, C. Wan, Y. Cai, R. Liao, L. Sigal, and S. Ma, "Latenthoi: On the generalizable hand object motion generation with latent hand diffusion." in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 17 416–17 425.

[41] 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 *Int. Conf. Mach. Learn.* PmLR, 2021, pp. 8748–8763.