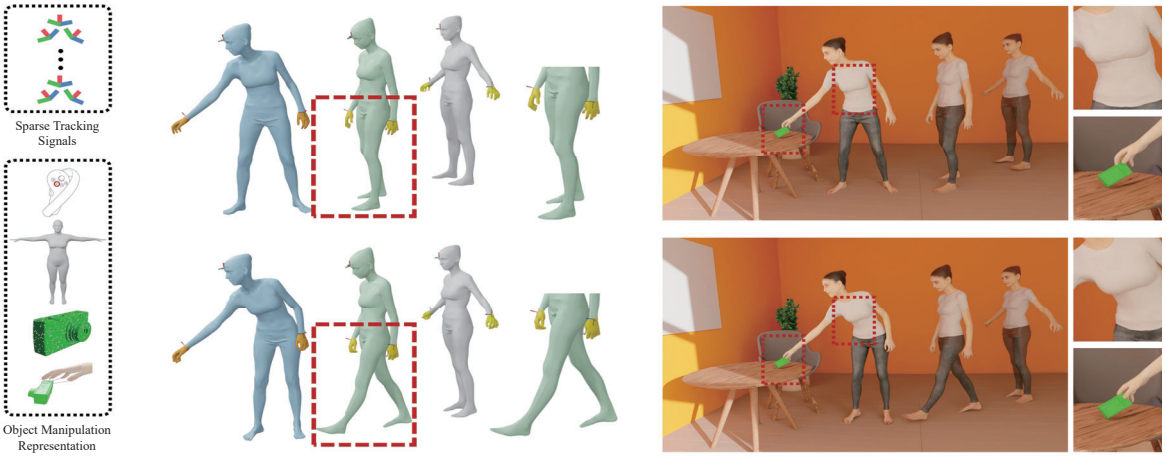
MMG: Manipulation-aware Holistic Human Motion Generation from Sparse Tracking Signals

Xuehuai Shi[*†] Renzhi Xiao[‡] Yilun Sheng[*] Lili Wang[‡] Jian Wu[‡] Xiaobai Chen[*] Jieming Yin[*]

Qingshan Liu[*‡]



(a) Inputs (b) Holistic Motion Reconstruction (c) Generated Manipulation-aware Holistic Motion

Figure 1: **Holistic human motion generation with/without object manipulation from the same sparse tracking signals.** Given the inputs in (a) that include sparse tracking signals and the object manipulation representation, MMG reconstructs the initial human body and hand (holistic) motion using the sparse tracking signals at the top of (b), refines the human motion with the constraint of the object manipulation representation at the bottom of (b), and finally generates the manipulation-aware holistic human motion with manipulation disabled (top) and enabled (bottom) in (c).

Abstract

Generating realistic avatar motion via sparse tracking signals through VR devices is essential for enhancing the immersive user experience. Human-object manipulation behaviors not only affect hand motion but also significantly impact body motion. However, existing motion generation methods for human-object interactions overlook the coordinated coupling between body and hand motions during manipulations. Due to the diversity and complexity of holistic motion (body and hand motions simultaneously) in the latent motion space, generating physically plausible and temporally consistent holistic motion in real time, via the joint constraints imposed by sparse tracking signals and manipulation content, is a major challenge in the human motion generation task. We propose the manipulation-aware holistic human motion generation method (MMG) to help resolve this issue. In MMG, first, we construct a manipulation-aware holistic human motion generation framework that serially compresses the latent motion space distribution of the body and hand to generate realistic holistic human motion with object manipulation enabled. Second, to enhance the impact of object manipulation on holistic motion generation, MMG designs a novel object manipulation representation to extract effective manipulation features. Third, MMG is trained by an elaborate progressive manipulation-guided training algorithm to improve motion generation robustness and inference performance. Compared to state-of-the-art methods, MMG achieves up to a 39% improvement in the generated holistic motion quality with a 3.55× speedup in generation performance. In manipulation-enabled scenes, MMG generates holistic motion in real time ($\geq$24fps). Compared to the state-of-the-art methods, its perceived quality is significantly improved, and the task performance of holistic motion-required VR manipulation is high-significantly improved. This paper's code is at https://github.com/XRZ-BUAA/MMG.

Index Terms: Human Motion Generation, Manipulation Awareness, Real-time Holistic Motion, Virtual Reality

1 Introduction

In VR, generating realistic human motions to simulate user actions in real time bridges the physical and virtual worlds, which is essential for achieving natural and immersive experiences. However, mainstream VR devices, such as the Meta Quest Pro, Apple Vision Pro, or PICO 4 Pro, typically provide only sparse tracking signals, leaving most of the user's body unmonitored. This sparsity makes generating real-time, coordinated human body and hand (holistic) motions a challenging task. Object manipulation, a core foundation of VR interactions, enables users to grasp, move, and rotate virtual objects like they would in the real world. When combined with sparse tracking conditions, the constraints of manipulation make it more difficult to synthesize realistic holistic motion in real time.

[*]The State Key Laboratory of Tibetan Intelligent Information Processing and Application, School of Computer Science, Nanjing University of Posts and Telecommunications, Jiangsu, China, 210023.

[†]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beihang University, Beijing, 100191.

[‡]Corresponding author. E-mail: qsliu@njupt.edu.cn.

Current research on real-time human-object interaction (HOI) motion generation predominantly focuses on either reconstructing body motions [1, 15, 13] from sparse signals or independently estimating hand motions [57, 42]. This fragmented approach lacks a coordinated mechanism for generating holistic motions in VR object manipulation scenarios. Additionally, existing motion generation techniques that incorporate manipulation content have two key limitations: they struggle with real-time dynamic body motion reconstruction from sparse signals, and they cannot synthesize integrated body-hand motions simultaneously [40, 51, 44]. Naive stacking independent body and hand motion modules creates physical implausibility by ignoring object-manipulation constraints, while also creating computational overhead that breaks real-time performance requirements ($\geq 24 fps$). As shown in Fig. 2 (a), when users attempt to manipulate toothpaste from a distance, compared with the ground truth (GT), the naive stacking method (NS) generates unnatural, stiff leg movements that fail to convey the intended spatial manipulation. The problem persists during the object manipulation. As demonstrated in Fig. 2 (b), when a user takes a photo with a camera, the method produces obvious body posture errors due to its inability to accurately interpret manipulation intent. These unnatural holistic postures and poor frame rates significantly degrade the VR user experience.

The above observations formulate three major challenges that need to be addressed to efficiently synthesize realistic holistic motions in object-manipulation-enabled scenes. The first challenge is compressing the distribution of latent holistic motion space with manipulation constraints to enhance the physical plausibility of the synthesized holistic motion. The second challenge involves efficiently extracting manipulation features to regulate holistic motion generation. The third challenge is improving the performance and robustness of the motion generation model to enhance the user experience in manipulation-enabled VR environments.

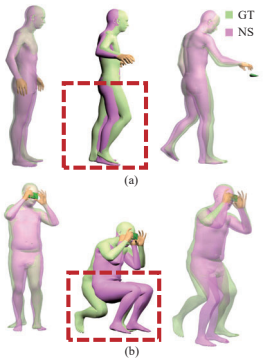


Figure 2: Comparison between ground truth (GT) and holistic motions generated by the naive stacking body and hand motion method (NS) [15, 42] in manipulation-enabled scenes, showing (a) approaching the manipulated object and (b) during manipulation.

We propose the manipulation-aware holistic human motion generation method (MMG) to address the aforementioned challenges. For the first challenge, we construct a manipulation-aware holistic motion generation framework. This framework utilizes sparse tracking signals and manipulation content to serially compress the latent motion space from the human body to the hands, thereby generating complex and natural holistic motions. To tackle the second challenge, we design a novel object manipulation representation. This representation efficiently encodes the manipulated object and manipulation motion features based on avatar shapes, achieving manipulation-aware motion standardization. For the third challenge, we introduce a progressive manipulation-guided training algorithm. This algorithm incrementally trains the model from initial to manipulation-constrained holistic motion and incorporates a distillation mechanism to simplify model complexity, resulting in robust holistic motion inference with significant performance improvements. Fig. 1 illustrates an example of holistic motion generation by MMG. The generated holistic motion in Fig. 1 (c) shows that when the user manipulates the camera, not only does the hand motion dynamically update with manipulation behaviors, but the avatar's body motion, such as the shoulder, also changes significantly depending on whether camera manipulation is enabled. The quality of holistic motion generated by MMG surpasses the state-of-the-art method by 39%, achieving up to a 3.55× generation performance improvement.

The contributions of this paper are as follows:

- A manipulation-aware holistic motion generation framework to synthesize realistic holistic motion in manipulation-enabled VR scenes;
- A novel object manipulation representation method to extract effective manipulation features for enhancing holistic motion synthesis accuracy;
- A progressive manipulation-guided training algorithm to improve synthesis quality with reduced model complexity for holistic motion generation;
- A dedicated user study that proposed for the first time to evaluate the user experience of the generated holistic motion through actual VR manipulation tasks.

2 Related Work

This section reviews recent studies related to our work.

2.1 Full-body Motion Estimation from Sparse Signals

Recent work focuses on estimating full-body motion through sparse inertial measurement units (SIMUs) attached to the body, avoiding complex sensor systems. Marcard et al. [48] first employ an anthropometry-constrained statistical body model with six SIMUs to enhance outdoor motion capture accuracy. Subsequent works apply various generation models, e.g., RNN-based networks [22, 55] and transformer-based networks [53, 26], to tackle unclear spatial dependencies and limited real-time prediction. Diffusion models [38, 20, 39] further advance motion reconstruction, including text-conditioned and action-conditioned generation [45, 27, 56, 58].

With the proliferation of VR applications, SIMUs remain cumbersome, prompting research on generating full-body motion from sparse tracking via VR/AR devices. Variational autoencoder-based methods [12, 9] are introduced to improve the generation accuracy of full-body motion from noisy head and hand posture signals in VR devices. A reinforcement learning framework [50] is implemented to enhance the physical plausibility of valid full-body motions from sparse user signals. For real-time capture and localization, Yi et al. [54] mitigate translation drift from missing global positioning and SLAM failures. Constraint-based diffusion frames sparse-to-full-body tracking as conditional sequence generation [59, 7, 43]. Stratified designs improve lower-body accuracy [13, 15], and context-aware conditioning enhances temporal consistency [33]. Efforts also improve motion synthesis quality in diverse 3D scenes, including fixed-trajectory movements [33], body-object interaction [29], and multi-person interactions [30]. Lightweight generative networks [2, 3] model conditional distributions in latent motion space for 3D upper-body pose accuracy . To address the limitation of ambiguous lower-body motion generation, many kinematics-based works are introduced to reconstruct human motions [52, 24, 15].

Recently, researchers focus on enriching the user experience of avatar generation in VR. Ahuja et al. [1] propose the first emphasized, stylized full-body motion synthesis system in VR, fusing user's actual motions with stylized examples from limited inputs from HMD and both-hand positions, synthesizing coordinated and expressive full-body avatar motions in real time. Wang et al. [49] propose a hierarchical dressed human representation method via physically decoupled diffusion models, enabling reusable, physically layered human generation with complex clothing. For human-object interaction, Hu et al. [21] use encoder-residual graph convolutional networks and multi-layer perceptrons to predict human motions during interactions, but neglect hand motion during the process.

2.2 Hand-Object Interaction Pose Estimation

Research on hand-object interaction pose estimation advances human motion generation by focusing on enhancing hand motion in human-object manipulation.

Numerous datasets advance hand motion estimation. Many RGB-D video datasets capture precise 3D hand poses during hand-object interaction [17, 19, 5, 6, 14]. Recent datasets [41, 23, 4] provide 3D full-body shape and pose sequences during object interactions, enabling deeper research into full-body human-object interaction.

Leveraging these datasets, various hand motion estimation methods have been proposed. Tzionas et al. [46] combine generative and discriminative models for reasonable estimation under occlusion and data loss. Zhang et al. [57] integrate voxel occupancy with geometric details to regress finger motion from wrist and object trajectories. Taheri et al. [42] apply a two-stage inference pipeline to synthesize realistic and temporally consistent bimanual motions. Cheymol et al. [8] propose a virtual-hand adaptive avoidance algorithm, simulating user responses to flames in real environments, enhancing the overall perception of avatar vulnerability in VR. Qu et al. [36] generate semantically consistent non-human hand motions from hand skeletal animations to enrich the user's exploration experience in VR. Conditional diffusion improves temporal coherence and context consistency in interaction-related motion synthesis [45, 7, 43, 33]. However, these methods neglect physical continuity between body and hands during manipulation, reducing overall plausibility, and overlook manipulation-aware holistic motion. Naively combining state-of-the-art body and hand generation methods ignores manipulation context, degrading global consistency and user immersion. MMG generates realistic manipulation-aware holistic motion efficiently via the joint constraints of sparse tracking signals and effectively extracted object manipulation representation.

3 Manipulation-aware Holistic Motion Generation

3.1 Problem Statement

To generate realistic holistic motion in manipulation-enabled VR scenes, MMG is designed to generate manipulation-aware body and hand motion simultaneously using sparse tracking signals and manipulation content.

Inputs. The inputs of MMG include two parts: sparse tracking signals and object manipulation representation. Sparse tracking signals originate from a common HMD and its two accompanying controllers, which are denoted by a time-dependent vector $M(t) = [mh(t), ml(t), mr(t)]$. $mh(t)$, $ml(t)$, and $mr(t)$ drive the sparse movements of the user's head, left wrist, and right wrist at the timestamp t, respectively [13, 15]. Each of these functions possesses 3 degrees of freedom for 3D translation and the corresponding translation velocities, as well as global rotation angles and angular velocities based on the 6D representation in [24]. The sparse tracking signals are represented as $M \in \mathbb{R}^{T_{in} \times 54}$, where T_{in} is the time interval for sparse tracking signal sampling.

The object manipulation representation consists of the manipulation state label $\eta(t)$, the avatar model shape β, the sparse vertex set F_o of the manipulated object, and the manipulation motion feature $F_d(t)$. $\eta(t)$ is a binary value indicating whether the avatar is in the manipulation mode at the timestamp t. β describes the shape of the avatar. F_o contains N_o vertices uniformly sampled from the manipulated object O. $F_d(t)$ is the set of shortest distances from N_d sampled points on the left and right hands, respectively, to the manipulated object at the timestamp t. N_o and N_d are hyperparameters that determine the trade-off between computational complexity and representation fidelity. By combining sparse tracking signals and manipulation content, we form the complete inputs, represented as $X(t) = [M(t), \eta(t), \beta, F_o, F_d(t)]$.

Serialized Diffusion-based SMPL Representation. We use the standard skeletal rig SMPL-X to represent the human body and hands [34]. For the joint j, the local rotation $y^j(t)$ is defined within the set function $Y(t)$ at the timestamp t. The global rotation $G(y^j(t))$ is calculated by cumulatively multiplying back to the root joint, as shown in Equation 1 [15]:

$$G(y^j(t)) = \prod_{i \in A(j)} y^i(t) \tag{1}$$

where $A(j)$ is the ordered joint ancestor set of j, and $G(y^j(t))$ denotes the final joint motion of the joint j.

In generating natural human motion during manipulations in VR, the spatial and shape characteristics of the manipulated object, as well as the relative position between the human and the object, impose kinematic constraints on the human body and hand postures. These constraints determine the reachable space and possible postures of the body and hands. The complexity of manipulation features poses challenges for learning human motion representations. Therefore, we introduce the object manipulation representation to constrain the human body and hand motion, enhancing the physical plausibility of the holistic motion generated by MMG.

Slight deviations in manipulation behavior lead to significant visual differences in body motion. Physically plausible body postures depend on sparse tracking signals and the spatial relationship with the manipulated object. Additionally, in the SMPL-X model, body joints are connected to the hands through wrist joints. Therefore, a serialized diffusion-based approach is constructed to generate human motion from the body to the hands, and a progressive manipulation-guided training algorithm is introduced to enhance the robustness and performance of the MMG model.

Outputs. MMG outputs the holistic motion estimation, denoted by the set function $Y(t)$. MMG first generates the latent code of body $\ddot{Y}_{body}(t)$ and both hands $\ddot{Y}_{hands}(t)$ serially, which is defined as: $\ddot{Y}(t) = \ddot{Y}_{body}(t) \cup \ddot{Y}_{hands}(t)$. Then, MMG leverages the motion decoders to generate $Y(t)$ with the input of $\ddot{Y}(t)$. *body* and *hands* denote the number of joints in the body and both hands of the avatar model, respectively, which are 22 and 30 in SMPL-X [34]. Therefore, the motion outputs of MMG have a dimensionality of $(22+30) \times 6 = 312$, i.e., $Y(t) \in \mathbb{R}^{T_{out} \times 312}$, where T_{out} is the number of output frames generated in a single generation process.

3.2 Manipulation-aware Holistic Motion Generation Framework

In this section, we first illustrate the procedure of proposed MMG framework, and then give the network structure details of MMG.

Fig. 3 illustrates the framework of the proposed MMG. MMG has three stages: stage 1, the holistic motion latent learning stage; stage 2, the initial holistic motion code generation stage; and stage 3, the manipulation-aware holistic motion generation stage.

Stage 1 learns precise body and hand motion latent codes from sparse tracking signals and holistic motion data to guide subsequent stages. While standard autoencoders have limited generative ability and discontinuous latent spaces, variational autoencoders provide stronger generative capabilities, continuous latent spaces, and reduced overfitting risks [28]. Therefore, stage 1 employs a conditional variational autoencoder (CVAE) to learn these motion latent codes. The CVAE architecture comprises four modules: body motion encoder, hand motion encoder, body motion decoder, and hand motion decoder.

The stage 1 process begins by combining holistic motion data with sparse tracking signals through a cross-attention mechanism [47]. The body and hand motion encoders then encode these fused features to generate precise motion latent codes for guiding the motion latent codes generation in later stages. Finally, the decoder modules reconstruct precise holistic motions using these latent codes.

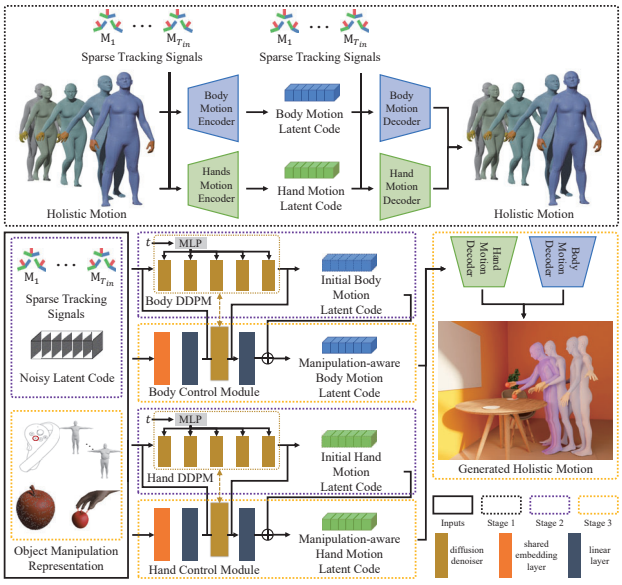


Figure 3: **Illustration of the MMG framework.** It consists of three stages. In stage 1 (marked in a black dotted box), a conditional variational autoencoder (CVAE) learns the latent holistic motion code from sparse tracking signals. The CVAE's body and hand motion decoders then reconstruct the holistic motion from this latent code. In stage 2 (marked in purple dotted boxes), a serialized diffusion model reconstructs the latent motion code from noisy latent codes without manipulation content, using the sparse tracking signals as guidance. In stage 3 (marked in yellow dotted boxes), a serialized control network refines the stage 2 latent motion code by incorporating object manipulation representation. The final holistic motion is then decoded using the stage 1 motion decoders.

Stage 2 aims to generate initial body and hand motion latent codes from real-time sparse tracking signals. We design a serialized diffusion model consisting of a body denoising diffusion probabilistic module (DDPM) and a hand DDPM. Each DDPM contains five denoising layers and incorporates timestep embedding t processed through a multilayer perceptron (MLP). This enables the model to dynamically adapt to different diffusion stages [20]. To ensure feature dimension alignment and information integrity, we implement dedicated embedding modules for three input types:

1) the sparse signal embedding module, 2) the noisy body latent code embedding module, and 3) the concatenated embedding module for initial body motion and noisy hand latent codes.

The initial motion latent codes in stage 2 are generated in two steps. First, the sparse tracking signals and noisy body latent codes are embedded through their respective modules and concatenated before being input into the body DDPM to generate the initial body motion latent code. Second, this initial body motion latent code is concatenated with the noisy hand latent code for embedding, then combined with the embedded sparse signals as input to the hand DDPM to produce initial hand motion latent codes.

Stage 3 refines the initial body and hand motion latent codes from stage 2 by applying object manipulation representation constraints, producing manipulation-aware motion latent codes. This stage employs a serialized control network with body and hand control modules, adding object manipulation representation as additional input. To maintain proper feature dimensions and information integrity, we use two embedding modules: 1) the object manipulation feature embedding module, and 2) the concatenated embedding module for manipulation-aware body motion and noisy hand latent codes.

The stage 3 workflow proceeds as follows. Firstly, we concatenate the embedded sparse tracking signals, noisy body latent codes, and object manipulation representation features. This combined input enters the body control module to generate the residual manipulation-aware body latent code. Adding this residual to the initial body latent code creates the refined manipulation-aware body motion latent code. Secondly, we combine the body motion latent code and noisy hand latent code through concatenated embedding, then merge them with the embedded sparse tracking signals and object manipulation representation features. The hand control module processes these concatenated features to generate hand latent code residuals. Adding these residuals to the initial hand latent codes produces the manipulation-aware hand motion latent codes. Thirdly, stage 3 utilizes copies of the body and hand motion decoders from stage 1, maintaining fixed parameters without additional training. These decoders use the manipulation-aware body and hand motion latent codes to synthesize the final manipulation-aware holistic motion in real time.

Network Structure Details. In stage 1, MMG employs a conditional variational autoencoder (CVAE) with a Transformer architecture and skip connections to learn latent holistic motion from sparse tracking signals, similar to S^2Fusion [43]. The CVAE's encoder and decoder each contain 9 layers, with 4 attention heads per layer.

In stage 2, the architecture consists of several embedding modules: the sparse signal embedding module uses three sequential layers (convolutional, linear, convolutional); the noisy body latent code embedding module uses a single convolutional layer; and the concatenated embedding module for initial body motion and noisy hand latent codes combines linear and convolutional layers. For both body and hand motion generation, the DDPMs use a 6-layer DiT backbone [35].

In stage 3, the object manipulation feature embedding module uses a residual network with 4 residual blocks and a Transformer architecture. The concatenated embedding module for manipulation-aware body motion and noisy hand latent codes combines a linear layer and a convolutional layer. Both body and hand control modules share the same structure: a linear layer, a fixed DDPM copied from stage 2, and another linear layer. The control modules use linear layers to transform manipulation-aware features into a more suitable feature space for downstream processing. Following other diffusion-based motion generation works, our denoisers in stages 2 and 3 directly predict the final denoised result instead of predicting noise as in conventional diffusion [37, 15].

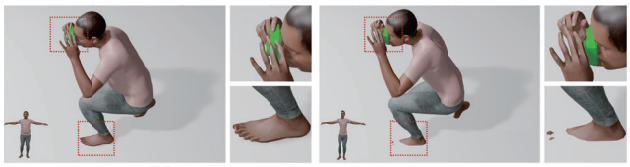
We set the length of each input motion sequence to 20, meaning both T_{in} and T_{out} are 20. Each latent code has dimensions of $(1,\ 256)$ and represents the motion across all 20 frames. The CVAE encoder converts motion sequence data $\mathbb{R}^{(n_{seq}, n_{joints}*6)}$ into a latent code in $\mathbb{R}^{(1,\ 256)}$, while the decoder reverses this process. Note that only the decoder is used in the motion inference process. Body joints and hand joints are trained separately, with 22 body joints and 30 hand joints (15 for each hand).

3.3 Object Manipulation Representation

There is a lack of efficient methods to quickly extract and accurately represent object manipulation content, which is crucial for synthesizing natural and realistic holistic motion during object manipulation. In this section, we extract effective manipulation features to guide the manipulation-aware holistic motion encoding. The object manipulation representation $I(t)$ is

$$I = [S(t), \beta, F_o, F_d(t)], \tag{2}$$

which consists of the manipulation state label $S(t)$ at the timestamp t, the shape feature β of the avatar model SMPL-X, the manipulated object feature F_o, and the manipulation motion feature $F_d(t)$ at the timestamp t.



(a) (b)

Figure 4: Visualization of holistic motion with different shapes when the avatar is manipulating the camera.

$S(t) \in \mathbb{R}^1$ is a binary parameter that indicates whether the user is engaged in manipulation at the timestamp t. MMG allows users to actively confirm manipulating objects via a controller button. Upon confirmation, MMG uses the three other features mentioned above to refine human body and hand motion, ensuring that the generated motion appears more natural and realistic during object manipulation.

MMG incorporates the body shape parameter $\beta \in \mathbb{R}^{10}$ of SMPL-X into the object manipulation representation. In a physically plausible holistic motion, local rotations of avatars with different shapes vary for the same manipulation behaviors and sparse tracking signals. Fig. 4 visualizes the holistic motion moment of an avatar manipulating the camera with different body shapes. Compared to the holistic motion with the shape shown in Fig. 4 (a), the slimmer avatar in Fig. 4 (b) exhibits penetration of the hands and feet under the same sparse tracking signals.

To provide MMG with the geometric shape and spatial position of the interacted object for synthesizing precise motion during interactions, we incorporate the sampled interacted object's vertices $F_o \in \mathbb{R}^{2N_o \times 3}$ into the object manipulation representation. To reduce the feature complexity and enhance the representation effectiveness, we sample N_o vertices of the interacted object based on the left-hand and right-hand coordinate systems instead of directly integrating all vertex information or using uniform vertex sampling. First, we project the interactive object O onto the unit sphere surface. For any vertex v on O, its projection on the unit sphere's surface is denoted as v':

$$v' = \frac{v - mean(V)}{arg_{v_i \in O} max(v_i - mean(V))} \tag{3}$$

where V denotes the set of all vertices on O, and $mean(V)$ represents the average value of all vertices in V. Next, we uniformly sample N_o vertices $\{v'_1, ..., v'_{N_o}\}$ on the unit sphere, and transform them into both the left-hand and right-hand coordinate systems L, R to obtain F_o:

$$\begin{aligned} F_o = &[(r^L_1, \theta^L_1, \phi^L_1), ..., (r^L_{N_o}, \theta^L_{N_o}, \phi^L_{N_o}), \\ &(r^R_1, \theta^R_1, \phi^R_1), ..., (r^R_{N_o}, \theta^R_{N_o}, \phi^R_{N_o})] \end{aligned} \tag{4}$$

where $(r^{L(R)}_i, \theta^{L(R)}_i, \phi^{L(R)}_i)$ is the polar coordinate representation of v'_i in the left(right)-hand coordinate system.

To achieve responsive and stable holistic motion and to promote smooth motion transitions during manipulation, $F_d(t) \in \mathbb{R}^{2N_d \times 1}$ provides continuous information about proximity using the minimum distance from hands to the object's surface at the timestamp t. We uniformly sample N_d points on both the left and right hands at the timestamp t, denoted as $p^l_1(t), ..., p^l_{N_d}(t), p^r_1(t), ..., p^r_{N_d}(t)$. For each sampled point $p(t)$, we calculate the nearest distance $d(t)$ to the interacted object's surface, and obtain $F_d(t)$:

$$F_d(t) = [d^l_1(t), ..., d^l_{n_d}(t), d^r_1(t), ..., d^r_{n_d}(t)] \tag{5}$$

3.4 Progressive Manipulation-guided Training

The proposed progressive manipulation-guided training algorithm trains MMG in three steps: 1) latent motion training, 2) initial holistic motion code training, and 3) manipulation-aware holistic motion code training.

Latent motion training trains the CVAE in stage 1 of MMG. The human body and hand motion codes in the latent space within the CVAE serve as references for generating motion codes in stages 2 and 3. The body and hand motion decoders within the CVAE rely on accurate motion codes to reconstruct holistic motion with minimal deviation. The CVAE is trained using the loss function shown in Equation 6:

$$\begin{aligned} L_{cvae} = &\lambda_{kl} D_{kl}\left(N(\mu, \rho) || N(0,1)\right) \\ &+ \lambda_{rec} L_1(\hat{R}, R) \\ &+ \lambda_{jp} L_1(\hat{P}, P) \end{aligned} \tag{6}$$

which includes the KL divergence, the reconstructed joint rotation, and the joint position loss terms. λ_{kl} is the weight assigned to the KL divergence loss; D_{kl} denotes the KL divergence loss term, which aligns the latent space in the CVAE with a standard normal distribution; μ and ρ are the mean and standard deviation predicted by the CVAE encoder. λ_{rec} and λ_{jp} are weights of the reconstructed joint rotation and the joint position loss terms, respectively. L_1 is the smooth L1 loss function [18]. $\hat{R}$ and R are the predicted and ground truth joint rotations, respectively. $\hat{P}$ and P are joint positions calculated using forward kinematics [24] based on $\hat{R}$ and R, respectively.

Initial holistic motion code training focuses on enabling the serialized diffusion model in stage 2 to learn complex body postures and detailed hand postures. Given the input of the noisy body motion latent code and sparse tracking signals, the loss function for training the body DDPM is shown in Equation 7:

$$L^{ddpm}_{body} = L_1(\hat{Z}_{body}, Z_{body}) \tag{7}$$

where $\hat{Z}_{body}$ and Z_{body} are the predicted and ground truth body motion latent codes, respectively. Next, we fix the parameters in body DDPM and train the hand DDPM with the loss function L^{ddpm}_{hands} shown in Equation 8:

$$L^{ddpm}_{hands} = L_1(\hat{Z}_{hands}, Z_{hands}) \tag{8}$$

where $\hat{Z}_{hands}$ and Z_{hands} are the predicted and ground truth hand motion latent codes, respectively.

Manipulation-aware holistic motion training leverages the manipulation representation to train the serialized control network to refine the holistic motion code. First, we fix the parameters of the body and hand DDPM. Then, we train the body control module using the loss function shown in Equation 9:

$$\begin{aligned} L^{cn}_{body} = &L_1(\hat{Z}^{cn}_{body}, Z_{body}) \\ &+ \lambda_{jp} L_1(\hat{P}^{cn}_{body}, P_{body}) \end{aligned} \tag{9}$$

which includes terms for body joint rotation and position loss. $\hat{Z}^{cn}_{body}$ and Z_{body} are the predicted and ground truth manipulation-aware body motion latent codes. $\hat{P}^{cn}_{body}$ and P_{body} are positions of the joints calculated based on $\hat{Z}^{cn}_{body}$ and Z_{body}. Similarly, the hand control module is trained using the loss function shown in Equation 10:

$$\begin{aligned} L^{cn}_{hands} = &L_1(\hat{Z}^{cn}_{hands}, Z_{hands}) \\ &+ \lambda_{jp} L_1(\hat{P}^{cn}_{hands}, P_{hands}) \\ &+ \lambda_{vp} L_1(\hat{V}^{cn}_{hands}, V_{hands}) \end{aligned} \tag{10}$$

which includes the additional hand joints rotation loss term. V_{hands} are positions of the sampled hand vertices in the object manipulation representation, and $\hat{V}^{cn}_{hands}$ are the positions calculated by the output hand motion of MMG.

Model distillation is inspired by MotionLCM [10]. It trains a lightweight online network to fit the control-network-guided serialized diffusion model constructed in stages 2 and 3. This approach allows the model to predict clear latent representations with fewer-

Table 1: Comparison of holistic motion estimation results between state-of-the-art methods and MMG on GRAB [41]. SA. denotes the state-of-the-art body motion generation method SAGE [15], AG. denotes AGRoL [13], and GR. denotes the state-of-the-art hand motion generation method GRIP [42].

Methods	Body MPJRE	Body MPJPE	Hand MPJRE	Hand MPJPE	Body MPJVE	MPJPE	MPVPE
AG.[13] + GR.[42]	2.82	5.68	8.82	0.60	19.84	2.99	3.02
SA.[15] + GR.[42]	3.06	3.51	8.82	0.60	11.03	2.96	2.98
MMG	**2.06**	**2.81**	**6.70**	0.60	**9.57**	**1.82**	**1.89**

step inferences. Specifically, in stages 2 and 3, for each DDPM, we train a 24-layer DiT denoiser as the teacher model, then distill it into a target 6-layer denoiser model. Since the teacher and target models share the same network architecture and feature dimensions, we not only have the target denoiser learn the output of the teacher denoiser, but also have it learn the intermediate outputs of certain layers within the teacher denoiser network. During diffusion inference in stages 2 and 3, we first use 5 denoising steps. Then, we employ the motionLCM method [11] to reduce the number of denoising steps, ultimately requiring only a single denoising step in the inference process. Finally, the well-trained target denoisers are implemented in stages 2 and 3 of MMG.

4 Experiments

4.1 Experimental Setup

Coefficients Settings. The hyperparameters N_o, N_d, T_{in}, and T_{out} are set to 1024, 99, 20 and 20. In this paper, the coefficients of loss functions λ_{kl}, λ_{rec}, and λ_{vp} are set to 0.00001, 1.0, and 0.25, respectively. λ_{jp} is set to 1.0 and 50.0 for the body control module and the hand control module, respectively. For details of dataset preparation and optimization procedures used in MMG training, please refer to the supplementary material.

Real-time Running Setup. We use the Unity platform to perform real-time execution of the program. The optimized MMG is converted into ONNX format and integrated into Unity. The hardware setting of this paper for real-time execution includes a PICO 4 Pro HMD powered by a workstation with a 3.9GHz Intel® Core™ i9-12900K CPU, 32GB RAM, and an NVIDIA GeForce GTX 4090 graphic card. To obtain manipulation content, we use the 'A' button on the controller, allowing the user to determine whether to enter or exit the manipulation state. The object closest to the inner side of the hand (the palm side) is regarded as the manipulated object.

Evaluation Metrics. We first evaluate the accuracy and temporal consistency of MMG in holistic motion generation. Then, we assess the real-time performance of MMG. For motion generation accuracy metrics, we use body and hand **MPJRE** (mean per joint rotation error), and body and hand **MPJPE** (mean per joint position error). For temporal consistency, we employ body **MPJVE** (mean per joint velocity error). Due to the numerous joints and the small unit velocity of many joints in the hand, the hand **MPJVE** tends to be 0 in all methods and is therefore omitted in comparisons. We also use **MPJPE** and **MPVPE** (mean per vertex position error) to evaluate holistic motion.

4.2 Quantitative and Qualitative Results

We combine the state-of-the-art body motion generation methods AGRoL [13] and SAGE [15], along with the hand motion generation method GRIP [42] as benchmarks for the state-of-the-art holistic motion generation.

To fully evaluate the motion generation quality of MMG, quantitative comparisons are performed not only on the manipulation-included holistic motion dataset GRAB [41], but also on manipulation-excluded body motion datasets: AMASS [32], IDEA400 [31], and TRUMANS [25].

For fair comparisons, all methods use a batch size of 16. T_{in} and T_{out} of MMG, AGRoL, and SAGE are both 20, while GRIP retains its original T_{in} and T_{out} of 2 [42]. We reoptimize these methods on the respective datasets until convergence for each dataset comparison.

As shown in Table 1, on the manipulation-included holistic motion dataset GRAB, MMG demonstrates superior accuracy in both holistic and body motion generation compared to state-of-the-art methods while maintaining comparable accuracy in hand motion.

In terms of holistic motion generation quality, compared to both AGRoL+GR. and SA.+GR., MMG achieves a 39% improvement in MPJPE and a 37% improvement in MPVPE. In terms of body motion generation quality, compared to AGRoL+GR., MMG achieves a 27% improvement in body MPJRE, a 51% improvement in body MPJPE, and a 52% improvement in body MPJVE; compared to SA.+GR., MMG achieves a 32% improvement in body MPJRE, a 20% improvement in body MPJPE, and a 13% improvement in body MPJVE. In terms of hand motion generation quality, although MMG maintains consistency with state-of-the-art methods in Hand MPJPE, it still demonstrates a 24% improvement over state-of-the-art methods in Hand MPJRE.

To further validate the superiority of MMG, we compare it with state-of-the-art methods in Fig. 5. We visualize the holistic motion generation results in various manipulation sequences covering both the phase of approaching manipulated objects and the phase of operating them, and compare them with GT.

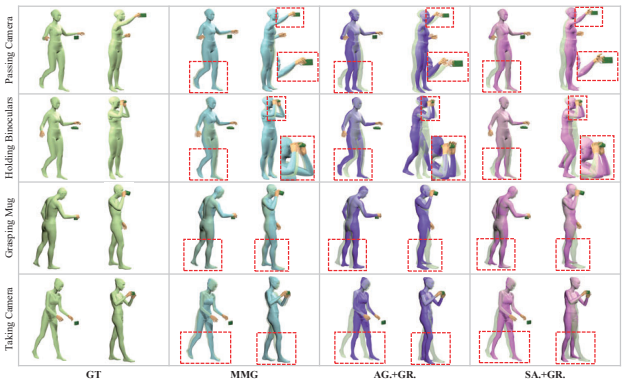


Figure 5: **Qualitative results on GRAB [41].** We compare MMG with state-of-the-art methods (AG.[13]+GR.[42] and SA.[15]+GR.[42]) against GT during the manipulation process. In each column, we set GT to be semi-transparent and overlay it with the results of compared methods, and mark noticeable artifacts with red boxes.

In row 1, when the user prepares to hold the camera with one hand and extend it outward, both state-of-the-art methods exhibit the issue of overly small leg strides, and fail to generate a hand posture that naturally conforms to the camera during handing it out. In row 2, when the user prepares to pick up binoculars with both hands, the final stopping step shows insufficient bending of the lower legs, and the arms as well as the left hand fail to form a natural grasping posture that conforms to the binoculars during manipulating binoculars. In row 3, the user's final step when picking up a teacup has an overly small leg stride, and during the drinking motion, the legs assume an unnatural closed stance lacking coordination with the hand motions. Row 4 shows the user taking the camera with both hands, again displaying the issue of an overly small final step stride, along with an unnatural leg-closing posture during taking camera. These visualizations demonstrate that object manipulation impacts holistic motion generation. Current state-of-

the-art methods, lacking object manipulation guidance, fail to produce natural and realistic holistic motions in two key ways.

First, when approaching an object, they do not properly adjust stride length despite consistent sparse tracking signals: for objects above the waist (rows 1 and 2), they fail to take the necessary larger final step; for objects below the waist (rows 3 and 4), they do not generate sufficient forward lean or downward bend. Second, without semantic guidance during object manipulation, subjects exhibit unnatural poses, including unnecessary leg positions where the legs are too close together (rows 3 and 4); poor coordination between arm and hand movements and body posture leads to unnatural object manipulations (rows 1 and 2). In contrast, MMG effectively refines the holistic motion latent code based on the extracted manipulation feature, enabling the generation of natural manipulation-aware holistic motion that achieves better alignment with GT.

The accuracy metrics and manipulation duration visualizations in the GRAB dataset both demonstrate that, compared to state-of-the-art methods, MMG generates more accurate and smoother holistic motion. In conclusion, compared to state-of-the-art methods, the holistic motions generated by MMG are significantly closer to GT.

Table 2: Comparison of body motion estimation between state-of-the-art methods and MMG across various body motion datasets.

Datasets	Methods	Body MPJRE	Body MPJPE	Body MPJVE	Body MPJJV
AM.[32]	AG. + GR.	2.61	4.80	22.76	3.24
	SA. + GR.	**2.60**	**3.77**	20.10	1.26
	MMG	2.84	4.95	**20.00**	**0.54**
ID.[31]	AG. + GR.	2.67	**4.06**	15.87	2.89
	SA. + GR.	**2.61**	4.50	12.43	0.67
	MMG	2.84	5.30	**11.69**	**0.35**
TR.[25]	AG. + GR.	4.70	6.41	30.92	6.48
	SA. + GR.	**4.60**	**6.10**	18.54	1.32
	MMG	4.65	6.70	**18.18**	**1.12**

Table 2 compares the quantitative results of MMG against state-of-the-art methods on the body motion datasets AMASS and IDEA400. Due to the lack of manipulation content guidance, MMG does not outperform state-of-the-art methods in body-only motion generation, resulting in a decrease in the quality of body motion generation of 1-8% for body MPJRE and 9-24% for body MPJPE.

Table 3: Performance (*ms*) comparisons of MMG with state-of-the-art methods.

Methods	Time Cost (*ms*)				
	Col.	Inf.	Ref.	Tot.	speedup
AG. [13] + GR.[42]	0.65	107.15	0	107.80	2.62×
SA.[15] + GR.[42]	0.65	143.80	1.57	146.02	3.55×
MMG	0.49	40.63	0	41.12	/

To further evaluate the temporal consistency of body motion, we employ body MPJJV (mean per joint jitter value). MPJJV calculates the mean absolute value of the jitter difference between the generated motion and ground truth motion for each joint. Thanks to MMG's use of a unified inference network to synthesize holistic motions while maintaining the semantic connection between the body and hands, the temporal coherence of MMG's body motions is significantly enhanced. As a result, MMG demonstrates improvements over state-of-the-art methods in both body MPJVE and MPJJV, achieving up to a 57% improvement in the temporal coherence metric MPJJV, while degrading the motion generation quality to 1- 8% in MPJRE and 9-24% in MPJPE.

Table 3 compares the time consumption of state-of-the-art holistic motion generation methods with MMG at each step. There are three steps in motion generation: tracking signal collection (Tra.), motion inference (Inf.), and motion refinement (Ref., which is only needed in SA.). Tol. denotes the total time cost. MMG significantly outperforms state-of-the-art methods in motion inference. Specifically, compared to AG.+GR., MMG achieves a 2.64× time cost improvement in this step, and in comparison with SA.+GR., MMG achieves a 3.54× time cost improvement. Furthermore, MMG does not require the additional construction of a refine network to refine generated motions, resulting in further performance enhancements. Overall, MMG achieves real-time generation of manipulation-aware holistic motion in VR ($\geq 24 fps$), delivering a 2.62-3.55× speedup compared to state-of-the-art methods while producing more accurate and temporally stable holistic motions.

4.3 Ablation Studies

We conduct an ablation study to evaluate the effectiveness of different components in MMG. The design of MMG includes three components: the serialized diffusion model (Diff.), the control network module (Con.), and the object manipulation representation (Rep.). Since Diff. serves as the foundation of MMG, we individually investigate the benefits of incorporating Rep. and Con. into MMG.

Table 4: Ablation on various components of MMG on GRAB. **Con.** denotes the control network module in stage 3 of MMG, and **Rep.** denotes the object manipulation representation.

Components		Body	Body	Body	Hand	Hand
Con.	**Rep.**	MPJRE	MPJPE	MPJVE	MPJRE	MPJPE
✗	✗	5.02	5.74	13.25	12.69	2.05
✓	✗	3.76	4.08	12.75	9.91	0.89
✓	✓	**2.06**	**2.81**	**9.57**	**6.70**	**0.60**

Table 4 presents a comparison of quality metrics under different component combinations of MMG. In Table 4, MMG derived from the various components is reoptimized and fully fitted on the GRAB dataset.

Effectiveness of Con. Table 4 compares the impact on MMG's holistic motion generation quality without Con. (row 1) to that with Con. (row 3). All metrics show significant improvement when Con. is implemented, indicating that Con. performs excellently in enhancing the accuracy and temporal stability of holistic motion generation.

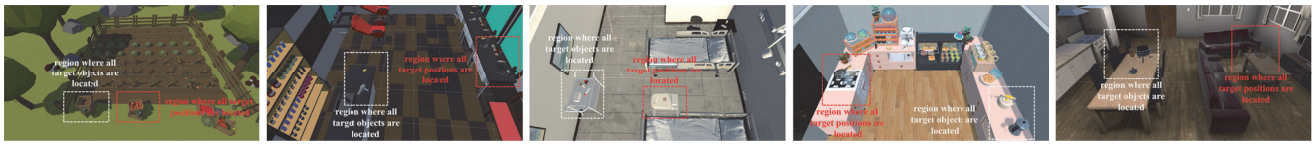
Effectiveness of Rep. In Table 4, we compare the effect of using uniformly sampled points from the manipulated object as input to Con. (row 2) with using Rep. as input for Con. (row 3). Compared to simple uniform sampling, the implementation that uses Rep. as input for Con. outperforms in terms of temporal consistency and, especially, accuracy in MMG's holistic motion generation, as showcased by body and hand MPJRE and MPJVE.

4.4 User Study

We conduct a psychophysical user study to evaluate the generated holistic motion quality and task performance during manipulation tasks in VR using MMG and the state-of-the-art method, respectively. Since SA.+GR. demonstrates a marked advantage over AG.+GR. in holistic motion generation quality, with only a $2 fps$ drop in frame rate, we select SA.+GR. as the state-of-the-art method for this user study.

We formulate two hypotheses for the user study:

H1 Compared to the state-of-the-art method, MMG achieves significantly higher perceived quality in the generated holistic motion during manipulation in VR;



(a) $farm$ (b) $barbershop$ (c) $medical\ room$ (d) $cafe$ (e) $cabin$

Figure 6: Visualization of all tested VR scenes for object manipulation tasks in the user study.

Table 5: 2AFC lists of manipulators and observers in MEQ.

Manipulator-based 2AFCs	
Q1	Do you find it easy to move or rotate objects during manipulation?
Q2	Do you feel that your body and hand motions are natural when performing manipulation tasks?
Q3	Does the framerate of the holistic motion generation allow smooth manipulation?
Q4	Do you feel that your body and hands can adjust postures appropriately when manipulating different objects?
Q5	Do you think the proposed method enhances the user experience during object manipulation?
Observer-based 2AFCs	
Q1	Do you observe that it is easy for the manipulator to manipulate objects?
Q2	Do you think the manipulator's body and hand motions are natural when performing manipulation tasks?
Q3	Does the holistic motion generation delay meet your expectations for manipulation tasks?
Q4	Do you observe that the manipulator's holistic motions can appropriately adjust postures when manipulating different objects?
Q5	Do you think the proposed method enhances the user experience when observing object manipulation?

H2 Compared to the state-of-the-art method, when generating holistic motion during manipulation, using MMG shows highly-significant task performance improvements.

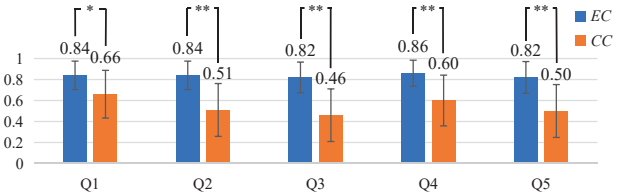
4.4.1 User Study Design

Setup The hardware setup and the operating environment for this study are the same as the 'Real-time Running Setup' in Section 4.1.
Participants We recruit 25 participants, consisting of 13 males and 12 females, aged between 18 and 50, with an average age of 31. None of the participants are in pilot user studies, and 11 of them have prior experience using VR HMDs. All participants have normal hearing and vision or have their vision corrected to normal levels through glasses.
Conditions We use MMG to generate avatar holistic motion during manipulation tasks as the experimental condition (*EC*), and use the state-of-the-art method SA.+GR. as the control condition (*CC*).
Procedure We construct five manipulation-required VR scenes: $farm$, $barbershop$, $medical\ room$, $cafe$, and $cabin$, as shown in Fig. 6. The manipulated objects in these scenes include static and dynamic objects of different sizes and shapes. Participants use either one hand or two hands to manipulate these objects, and they need to move laterally within *cabin* to pass through narrow passages. All participants complete the manipulation tasks in these five scenes using *EC* and *CC*. Each experimental procedure requires two participants, one as the manipulator and the other as the observer. For each pair of participants, each tested scene is presented randomly. In each scene, the manipulator's and observer's initial positions are fixed. The manipulator needs to manipulate all objects to the target positions in the scene, while the observer monitors the process from a third-person perspective. After each trial, we record the total time taken to complete the task and ask both participants to fill out the manipulation experience questionnaire (MEQ), and record the manipulator-based and the observer-based *score*s for each question in MEQ. The details of MEQ are shown in Table 5, where 5 two-alternative forced choice questions (2AFC) are designed for both the manipulator and the observer. From both the 'manipulator' and 'observer' perspectives, MEQ evaluates the holistic motion generation quality in terms of real-time feedback (Q1), physical realism (Q2), smoothness (Q3), motion flexibility (Q4), and overall user-perceived experience (Q5). Then, the pair proceeds to the next scene. Each participant takes one turn as the manipulator and one turn as the observer for all scenes, and completing all trials. Each participant takes an average of 33 minutes to complete all trials. 25 (participants) $\times$ 5 (scenes) $\times$ 2 (conditions) = 250 experimental trials are collected.

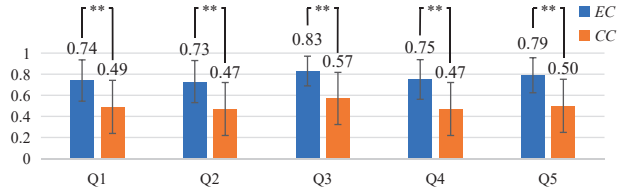
4.4.2 Results and Discussion

We conduct an ANOVA analysis to evaluate the user experience and the task performance between *EC* and *CC*.

We compare the perceived generation quality between *EC* and *CC* in holistic motion through *score*s for all 2AFCs in MEQ. As shown in Fig. 7, *EC* outperforms *CC* across all five core metrics from both the manipulator and observer perspectives: real-time feedback (Q1), physical realism (Q2), smoothness (Q3), motion flexibility (Q4), and overall user-perceived experience (Q5). Tables 6 and 7 demonstrate statistical significance indicators for Q1–Q5 from the two perspectives. Except for Q1 being "significantly better" from the manipulator perspective, *EC* achieves "highly significant" (p-value ≤ 0.01 [16]) superiority over *CC* in all other *score*s. Therefore, the results support **H1**.



(a) average values of manipulator-based $score$



(b) average values of observer-based $score$

Figure 7: Average values of *score* between *EC* and *CC*. A single asterisk indicates significant differences (p-value ≤ 0.05), and double asterisks indicate highly significant differences (p-value ≤ 0.01).

Although *EC* demonstrates significantly superior perceived quality of the generated holistic motion compared to *CC* from the manipulator perspective, both conditions exhibit insufficient flexibility during object manipulation state transitions due to not meeting the immersive frame rate requirement of $90fps$. As a result, *EC* only achieves a "significant improvement" in the manipulator-based Q1. Average *score*s under *CC* across all 2AFCs in MEQ remain below 0.7, primarily due to a reduced interactive experience caused by physical irrationality and inadequate frame rates. When the average

score exceeds 0.65 (indicating that participants tend to favor this technology), *score*s from the observer perspective are consistently lower than those from the manipulator perspective. According to participants' feedback, although holistic motion generation appears natural, Steam VR's positional drift causes intermittent avatar sliding from third-person perspectives. This reflects perceived holistic motion quality degradation from the observer perspective compared to the manipulator perspective.

Table 8 compares the time costs between *EC* and *CC* in completing manipulation tasks across four VR scenes. By leveraging more manipulation-aware holistic motion, *EC* reduces the need for manipulation corrections, while its flexible holistic motion feedback enables more precise interactions, thereby significantly improving manipulation efficiency. The significance metrics in Table 9 indicate that *EC*'s task completion time is statistically significantly shorter than that of *CC* in all tested scenes (p-value ≤ 0.01). Therefore, the results support **H2**.

Table 6: Statistical significance measures between *EC* and *CC* under the manipulator-based *score* of MEQ.

measure	Q1	Q2	Q3	Q4	Q5
$F_{1,240}$	11.91	35.51	40.978	22.67	30.47
η_p^2	0.05	0.13	0.15	0.09	0.11
p-value	0.02	0.00	0.00	0.00	0.00

Table 7: Statistical significance measures between *EC* and *CC* under the observer-based *score* of MEQ.

measure	Q1	Q2	Q3	Q4	Q5
$F_{1,240}$	17.08	18.56	23.89	21.88	25.79
η_p^2	0.07	0.07	0.09	0.09	0.10
p-value	0.00	0.00	0.00	0.00	0.00

Table 8: The average manipulation completion time (s) with the holistic motions generated by *EC* and *CC* under different VR scenes.

Condition	*farm*	*barbershop*	*medical room*	*cafe*	*cabin*
EC	81.5±12.3	36.8±7.5	41.9±7.0	33.1±5.7	109.2±14.7
CC	106.2±20.3	46.1±6.8	58.4±8.5	42.9±8.2	130.5±15.3

Table 9: Statistical significance measures between *EC* and *CC* in manipulation completion times of different VR scenes.

measure	*farm*	*barbershop*	*medical room*	*cafe*	*cabin*
$F_{1,48}$	28.05	21.31	22.70	24.35	25.27
η_p^2	0.37	0.31	0.32	0.34	0.35
p-value	0.00	0.00	0.00	0.00	0.00

5 Limitations and Future Work

MMG is trained on the AMASS, IDEA400, TRUMANS, and GRAB datasets. The datasets incorporate complex body postures such as bending and crouching, along with single-handed and bimanual manipulation poses like pinch grips and palmar support. They also cover 51 manipulated objects with geometric forms ranging from simple cubes to complex wine glasses and aircraft models, encompassing varying contact diameters (2.92-19.39cm). Here, contact diameter refers to the longest distance within the hand-object contact surface. The data foundations enable MMG to generate natural holistic motions for various object manipulation tasks. However, MMG still exhibits limitations.

Firstly, as shown in Fig. 8: (a) when the user manipulates a screwdriver, the hands fail to grip it naturally, instead adopting a pinching posture; (b) when handling the suitcase, the hands penetrate the model's surface. These problems stem from two reasons: 1) the dataset used to train the MMG has limited contact diameters and types of manipulation motions, and there is a lack of skeletal constraint modeling. Specifically, for elongated objects with a contact diameter less than 4.15cm, the holistic motions in the dataset are restricted to pinching, lacking more diverse actions such as gripping; 2) the dataset is deficient in cases involving objects with a contact diameter greater than 19.39cm, which limits the prior knowledge regarding the range of hand opening and, in turn, may cause the MMG to exhibit fixed hand motion patterns or cause penetration artifacts, leading to failed cases. Secondly, the absence of articulated objects prevents MMG from supporting holistic motion generation for rotational-axis manipulations, such as opening boxes or pulling drawers. Thirdly, MMG currently focuses on generating manipulation-aware holistic motion in the realm of Joint Funds of the National Natural Science Foundation of Chinahuman-object interaction, and has not yet explored the motion adaptation problem caused by geometric and physical differences between virtual and real worlds in the realm of human-scene interaction.

To address the above limitations, in future work, a comprehensive, holistic motion dataset that incorporates objects with hinge structures and richer sizes needs to be constructed to enrich the prior knowledge base. Furthermore, the virtual-physical perception alignment model needs to be studied to explore the manipulation-aware motion generation mechanism under virtual-real spatial discrepancies within a human-mix interaction framework.

6 Conclusion

We study the problem of generating human body and hand motions simultaneously, i.e., holistic motion, based on sparse tracking signals in manipulation-enabled VR scenes. Our key finding is that the manipulation content significantly influences the holistic motion, especially the body motion. Based on this, we propose a novel object manipulation-aware holistic human motion generation method that uses a specifically designed object manipulation representation to guide the framework in generating manipulation-aware holistic motions. Our method achieves real-time holistic motion generation ($\geq 24 fps$) in VR. Compared to state-of-the-art methods, our method achieves a 39% improvement in holistic motion generation quality while also delivering a 3.55× speedup in generation performance. Experimental results of the user study demonstrate that, compared to the state-of-the-art method, our method significantly enhances the perceived quality of the generated holistic motion during VR manipulations and completes the manipulation tasks with highly significant performance advantages.

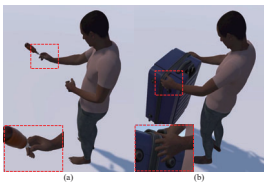


Figure 8: Failed cases of manipulating objects with extremely small (a) and large (b) contact diameters using MMG.

Acknowledgments

This work is supported by the National Natural Science Foundation of China through Project U24B20155, 62402231, 92473205; Jiangsu Provincial Special Funding Project for the Transformation of Scientific and Technological Achievements BA2022026; Natural Science Foundation of the Jiangsu Higher Education 24KJB520027; the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2024C03); Beijing Science and Technology Plan Project Z221100007722004; Natural Science Foundation of Jiangsu Province BK20243051; Open Project Fund of Anhui Provincial Key Laboratory of Bionic Sensing and Advanced Robotics AHFS2024KF07.

REFERENCES

[1] K. Ahuja, E. Ofek, M. Gonzalez-Franco, C. Holz, and A. D. Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021. 2

[2] S. Aliakbarian, P. Cameron, F. Bogo, A. Fitzgibbon, and T. J. Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13253–13262, 2022. 2

[3] S. Aliakbarian, F. Saleh, D. Collier, P. Cameron, and D. Cosker. Hmdnemo: Online 3d avatar motion generation from sparse observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9622–9631, 2023. 2

[4] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022. 3

[5] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8709–8719, 2019. 3

[6] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 361–378. Springer, 2020. 3

[7] A. Castillo, M. Escobar, G. Jeanneret, A. Pumarola, P. Arbeláez, A. Thabet, and A. Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4221–4231, 2023. 2, 3

[8] A. Cheymol, R. Fribourg, A. Lécuyer, J.-M. Normand, and F. Argelaguet. Avatar-centered feedback: Dynamic avatar alterations can induce avoidance behaviors to virtual dangers. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 91–100. IEEE, 2024. 3

[9] V. Choutas, F. Bogo, J. Shen, and J. Valentin. Learning to fit morphable models. In *European Conference on Computer Vision*, pp. 160–179. Springer, 2022. 2

[10] W. Dai, L.-H. Chen, J. Wang, J. Liu, B. Dai, and Y. Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024. 5

[11] W. Dai, L.-H. Chen, J. Wang, J. Liu, B. Dai, and Y. Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408. Springer, 2025. 6

[12] A. Dittadi, S. Dziadzio, D. Cosker, B. Lundell, T. J. Cashman, and J. Shotton. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11687–11697, 2021. 2

[13] Y. Du, R. Kips, A. Pumarola, S. Starke, A. Thabet, and A. Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 481–490, 2023. 2, 3, 6, 7

[14] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12943–12954, 2023. 3

[15] H. Feng, W. Ma, Q. Gao, X. Zheng, N. Xue, and H. Xu. Stratified avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 153–163, 2024. 2, 3, 4, 6, 7

[16] R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer, 1970. 8

[17] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 409–419, 2018. 3

[18] R. Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015. 5

[19] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3196–3206, 2020. 3

[20] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[21] Z. Hu, Z. Yin, D. Haeufle, S. Schmitt, and A. Bulling. Hoimotion: Forecasting human motion during human-object interactions using egocentric 3d object bounding boxes. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[22] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 2

[23] Y. Huang, O. Taheri, M. J. Black, and D. Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pp. 281–299. Springer, 2022. 3

[24] J. Jiang, P. Streli, H. Qiu, A. Fender, L. Laich, P. Snape, and C. Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pp. 443–460. Springer, 2022. 2, 3, 5

[25] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024. 6, 7

[26] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022. 2

[27] J. Kim, J. Kim, and S. Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 8255–8263, 2023. 2

[28] D. P. Kingma, M. Welling, et al. Auto-encoding variational bayes, 2013. 3

[29] N. Kulkarni, D. Rempe, K. Genova, A. Kundu, J. Johnson, D. Fouhey, and L. Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 947–957, 2024. 2

[30] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pp. 1–21, 2024. 2

[31] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

[32] 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*, pp. 5442–5451, 2019. 6, 7

[33] A. Mir, X. Puig, A. Kanazawa, and G. Pons-Moll. Generating continual human motion in diverse 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, pp. 903–913. IEEE, 2024. 2, 3

[34] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019. 3

[35] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. 4

[36] T. Qu, J. Wang, Y. Lin, J. Liu, C. Zhou, B. Zhang, K. Jiang, and Y. Bian. Becoming an animal? exploring proteus effect based on human-avatar hand gesture consistency. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 777–786.

IEEE, 2024. 3

[37] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 4

[38] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015. 2

[39] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[40] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13263–13273, 2022. 2

[41] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 581–600. Springer, 2020. 3, 6

[42] O. Taheri, Y. Zhou, D. Tzionas, Y. Zhou, D. Ceylan, S. Pirk, and M. J. Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *2024 International Conference on 3D Vision (3DV)*, pp. 933–943. IEEE, 2024. 2, 3, 6, 7

[43] J. Tang, J. Wang, K. Ji, L. Xu, J. Yu, and Y. Shi. A unified diffusion framework for scene-aware human motion estimation from sparse signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21251–21262, 2024. 2, 3, 4

[44] P. Tendulkar, D. Surís, and C. Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21179–21189, 2023. 2

[45] 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. 2, 3

[46] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016. 3

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[48] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, vol. 36, pp. 349–360. Wiley Online Library, 2017. 2

[49] Y. Wang, J. Ma, R. Shao, Q. Feng, Y.-K. Lai, and K. Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 436–445. IEEE, 2024. 2

[50] A. Winkler, J. Won, and Y. Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–8, 2022. 2

[51] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision*, pp. 257–274. Springer, 2022. 2

[52] D. Yang, D. Kim, and S.-H. Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In *Computer Graphics Forum*, vol. 40, pp. 265–275. Wiley Online Library, 2021. 2

[53] S. Yang, Z. Quan, M. Nie, and W. Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11802–11812, 2021. 2

[54] X. Yi, Y. Zhou, M. Habermann, V. Golyanik, S. Pan, C. Theobalt, and F. Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4):1–17, 2023. 2

[55] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13167–13178, 2022. 2

[56] 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*, pp. 16010–16021, 2023. 2

[57] H. Zhang, Y. Ye, T. Shiratori, and T. Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2, 3

[58] 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. 2

[59] X. Zheng, Z. Su, C. Wen, Z. Xue, and X. Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14678–14688, 2023. 2