COLLAGE: Collaborative Human-Agent Interaction Generation using Hierarchical Latent Diffusion and Language Models

Divyanshu Daiya[1], Damon Conover[2], Aniket Bera[1]
[1]IDEAS Lab, Department of Computer Science, Purdue University [2]DEVCOM Army Research Laboratory
{divyanshu, aniketbera}@purdue.edu, damon.m.conover.civ@army.mil

***Abstract*— We propose a novel framework COLLAGE for generating collaborative agent-object-agent interactions by leveraging large language models (LLMs) and hierarchical motion-specific vector-quantized variational autoencoders (VQ-VAEs). Our model addresses the lack of rich datasets in this domain by incorporating the knowledge and reasoning abilities of LLMs to guide a generative diffusion model. The hierarchical VQ-VAE architecture captures different motion-specific characteristics at multiple levels of abstraction, avoiding redundant concepts and enabling efficient multi-resolution representation. We introduce a diffusion model that operates in the latent space and incorporates LLM-generated motion planning cues to guide the denoising process, resulting in prompt-specific motion generation with greater control and diversity. Experimental results on the CORE-4D, and InterHuman datasets demonstrate the effectiveness of our approach in generating realistic and diverse collaborative human-object-human interactions, outperforming state-of-the-art methods. Our work opens up new possibilities for modeling complex interactions in various domains, such as robotics, graphics and computer vision.**
Paper website: https://collagemotion.github.io/

I. Introduction

Modeling human-like agent-object interactions is fundamental in the vision community, enabling applications in gaming, embodied AI, robotics, and VR/AR. While recent works have explored single-person and multi-human object interactions in non-collaborative settings [1]–[6], generating collaborative human-object-human interactions remains largely unexplored. This task requires a complex understanding of human actions and object interactions, as guiding individual agents along with the task involves extensive planning. Given the lack of rich datasets, training a generalized model is challenging. To address this, we propose incorporating the knowledge and reasoning abilities of large language models (LLMs) to guide a generative diffusion latent diffusion model for multi-human-object motion generation in collaborative settings. In the remainder of this paper, we will use the terms 'human' and 'agent' interchangeably, with the specific application determining the appropriate usage. For robotics applications, 'agent' may refer to either a real human or a robotic, human-like entity such as a humanoid.

Pre-trained LLMs, such as GPT-4 [7] and Llama 2 [8], have demonstrated emergent capabilities in reasoning, planning, and motion planning [9]–[12]. We hypothesize that LLMs could provide a general and domain-independent approach to modeling and planning interactive multi-human object and human-object-human task collaboration, given proper learning approaches. Learning to plan without a dataset can help with motion planning in outdoor settings, where currently, no dataset exists with extensive motion capture data. Utilizing humanoid robots in such settings is a significant hurdle, and effective use of planning via

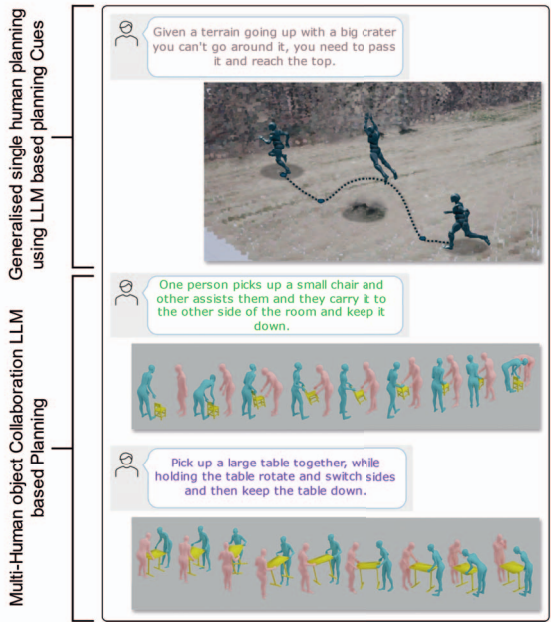


Fig. 1: *Text to collaborative motion and generalized motion generation by COLLAGE, based on user-provided text prompts. In the top image, a simulated humanoid robot adapts to the 3D terrain features based on the input text from the human collaborator. In the bottom image, the two human agents collaborate to handle an object using LLM-based planning via our architecture.*

LLMs for fine-grained motion generation could help with humanoid-based motion and interaction in outdoor environments[1] (Fig. 1). To capture the complex motion dynamics in collaborative settings, we propose a hierarchical motion-specific vector-quantized variational autoencoder (VQ-VAE) architecture that explicitly captures different motion-specific characteristics at different levels of abstraction, addressing the limitations of previous VQ-VAE models [13]–[15]. We incorporate a diffusion model [14], [16] for learning human motion in the latent space and propose a novel architecture to incorporate multi-human object interactions. We augment textual planning cues from LLM with codebook-based associations learned via VQ-VAE training, helping the diffusion model better learn to model according to the task description, given the complex interaction setting and associations. We showcase how plans and cues generated by LLMs can be utilized by the diffusion model for effective generation, demonstrating faster speed and greater generational diversity

[1]Dataset from GDIS Terrain Segmentation Dataset (Havre de Grace, MD), curated by DEVCOM Army Research Lab

compared to other models.

We evaluate our model's generalizability on the human-object-human CORE-4D dataset [17] and multi-human dataset like InterGen [18], comparing our results with existing works in this setting.

The main contributions of our work are:

- We propose a novel approach for generating collaborative human-object-human interactions by leveraging LLMs and hierarchical motion-specific VQ-VAEs, addressing the lack of rich datasets in this domain.
- We introduce a hierarchical motion-specific VQ-VAE architecture that captures different motion-specific characteristics at different levels of abstraction, avoiding redundant repeated concepts across layers and enabling efficient multi-resolution representation.
- We demonstrate the effectiveness of utilizing LLM-generated motion planning cues to guide the diffusion model through the denoising process, resulting in prompt-specific motion generation with greater control and diversity.
- We evaluate our model on multiple datasets, showcasing its generalizability and effectiveness in generating collaborative human-object-human interactions.

II. Related Work

Text-Conditioned Human Motion Generation. Generating human motions based on textual descriptions has been a recent research focus. Early approaches generated motions based on action categories [19]–[22], past motions [23]–[27], trajectories [28]–[32], and scene context [33]–[42]. Recent works have enabled direct generation of human motions from textual inputs [14], [43]–[63], extending to multi-person [64]–[66] and human-scene interactions [38], [67], [68]. However, generating collaborative human-object-human interactions remains largely unexplored.

Human-Object Interaction Generation. Modeling realistic human-object interactions is challenging due to the complexity of capturing both human motions and object dynamics. Prior research has addressed hand-object interactions [69]–[73], single-frame human-object interactions [74]–[79], and zero-shot settings [80]–[82]. Recent studies have explored whole-body dynamic interaction generation through kinematic-based [4], [83]–[94] and physics-based methods [95]–[105], but often suffer from limitations such as a narrow scope of actions, static objects, or lack of comprehensive whole-body motion representation.

Collaborative Multi-Human Interaction Modeling. Collaborative human-object-human interactions remain largely unexplored, despite the study of multi-human interactions in non-collaborative contexts [6]. The complexity arises from modeling intricate coordination between multiple humans and objects, requiring advanced planning and understanding of collective actions. Recent datasets and baselines, such as CORE-4D [17], have begun to address this gap, but further research is needed to develop models capable of handling such complex interactions.

Utilizing LLMs in Motion Generation. Large language models (LLMs) have demonstrated remarkable abilities in reasoning [9], planning [10], and task execution [11]. In the realm of digital humans, LLMs have been employed to guide motion generation [48], [106]–[109]. Our approach extends this line of work by utilizing LLMs to guide the generation of collaborative human-object-human interactions.

Hierarchical VQ-VAE and Diffusion Models in Motion Generation. Vector-Quantized Variational Autoencoders (VQ-VAEs) have been used to create quantized motion latent spaces [13], [14], [108], but struggle with complex, diverse motion generation due to limitations like small codebooks, as increasing the codebook size for complex datasets causes codebook collapse, as we observed while generalizing [14] to multi-human settings; smaller codebooks even in single-human settings result in less diverse motion.Hierarchical architectures [15] and diffusion models [14], [16] have shown promise in modeling complex human motion. Extending these ideas, our approach incorporates a hierarchical motion-specific VQ-VAE architecture and a diffusion model guided by LLM-generated plans to effectively generate collaborative human-object-human interactions.

III. Methodology

A. Hierarchical VQ-VAE with Description Cues

Modeling complex human-object interactions necessitates capturing motion dynamics at multiple levels of abstraction, from high-level trajectories and interaction types to low-level limb movements and object manipulations. To achieve this, we propose a hierarchical Vector Quantized Variational Autoencoder (VQ-VAE) that incorporates description cues provided by a Language Model (LLM) at each level of abstraction. This architecture enables the model to learn disentangled motion representations corresponding to different semantic concepts guided by hierarchical textual cues.

Our hierarchical VQ-VAE architecture captures motion dynamics at multiple levels of abstraction, as shown in 2. The encoders at each level map the inputs to latent representations, which are then quantized using codebooks. The decoders reconstruct the original data from the quantized latent representations. At each level l, the encoder for human i computes $Z_H^{i,(l)} = E_H^{(l)}\left(Z_H^{i,(l-1)}; \theta_H^{(l)}\right)$, where $E_H^{(l)}$ is a neural network with parameters $\theta_H^{(l)}$, and $Z_H^{i,(0)} = X^i$ is the input sequence for human i. Similarly, the object encoder computes $Z_O^{(l)} = E_O^{(l)}\left(Z_O^{(l-1)}; \theta_O^{(l)}\right)$, with $Z_O^{(0)} = Y$. We incorporate description cues $\mathbf{e}^{(l)}$ provided by an LLM at each level l, which are integrated into the encoder by augmenting the latent representations: $\tilde{Z}_H^{i,(l)} = \text{Concat}\left(Z_H^{i,(l)}, \mathbf{e}_H^{(l)}\right), \tilde{Z}_O^{(l)} = \text{Concat}\left(Z_O^{(l)}, \mathbf{e}_O^{(l)}\right)$, where $\mathbf{e}_H^{(l)}$ and $\mathbf{e}_O^{(l)}$ are the description embeddings for humans and objects at level l, respectively.

Multi-head attention mechanisms are employed to capture interactions between all pairs of entities, including n humans and m objects. We compute the attention for each entity i at level l across all pairs involving entity i, $A_{total}^{i,(l)} = \sum_{j \neq i} \text{MultiHeadAttention}\left(Q^{i,(l)}, K^{j,(l)}, V^{j,(l)}\right)$, where $Q^{i,(l)}$, $K^{j,(l)}$, and $V^{j,(l)}$ are the query, key, and value matrices respectively, and the summation extends over all other entities j. The updated latent vector for entity i after applying attention and layer normalization, $\hat{Z}_H^{i,(l)} = \text{LayerNorm}\left(\tilde{Z}_H^{i,(l)} + A_{total}^{i,(l)}\right)$. Vector quantiza-

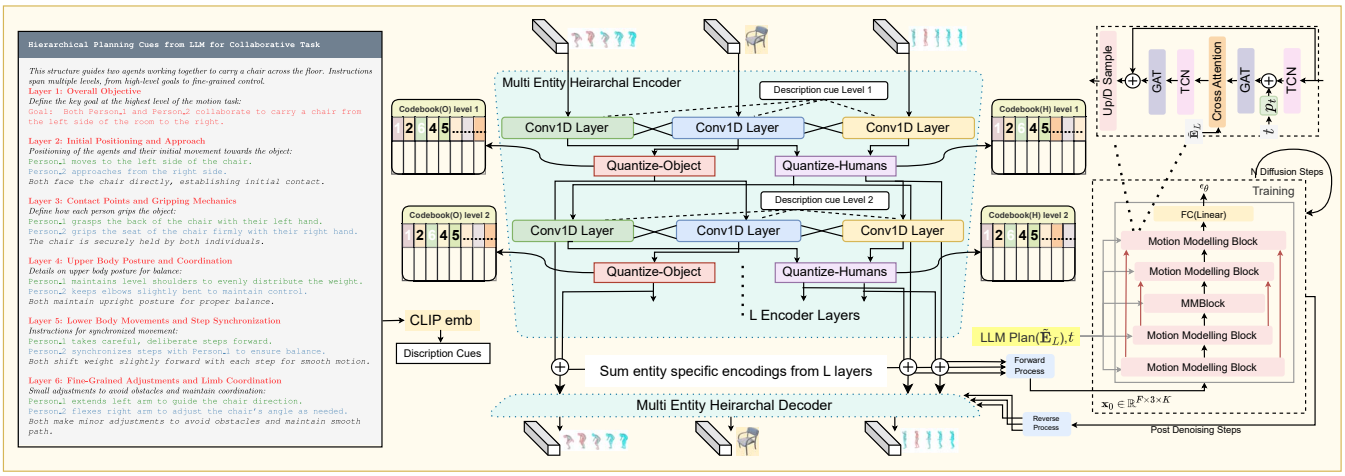


Fig. 2: *Overview of the proposed **COLLAGE** framework for collaborative human-object interaction generation. The hierarchical VQ-VAE encoder captures motion-specific characteristics at different levels of abstraction. The latent diffusion model operates in the learned latent space and incorporates LLM-generated motion planning cues to guide the denoising process, enabling the generation of prompt-specific interactions with enhanced control and diversity as in Fig 1.*

tion is performed using codebooks $\mathcal{C}_H^{(l)}$ and $\mathcal{C}_O^{(l)}$, mapping each latent vector to its nearest codeword, $\bar{Z}_H^{i,(l)} = \text{Quantize}\left(\hat{Z}_H^{i,(l)}, \mathcal{C}_H^{(l)}\right)$, $\bar{Z}_O^{(l)} = \text{Quantize}\left(\hat{Z}_O^{(l)}, \mathcal{C}_O^{(l)}\right)$. The decoder reconstructs the inputs from the quantized latent representations, proceeding hierarchically. For human i, the decoder function is $\hat{X}^i = D_H\left(\sum_{l=1}^{L} \bar{Z}_H{}^{i,(l)}; \phi_H\right)$, where D_H is a neural network with parameters ϕ_H. Similarly, for the object, $\hat{Y} = D_O\left(\sum_{l=1}^{L} \bar{Z}_O^{(l)}; \phi_O\right)$. It is worth noting that our approach differs from other models like Priority-based VQ-VAE [14] and T2M-GPT [13] in terms of the latent representation used by the decoder. In our model, we aggregate the latent codes from all layers of the VQ-VAE before passing them to the decoder. This allows our decoder to work with non-discrete, continuous representations. As a result, our model can directly utilize the continuous representations generated by the diffusion model without the need for discrete mapping. In contrast, models like Priority-based VQ-VAE and T2M-GPT learn distributions over discrete latent codes, requiring an additional step to map the continuous diffusion outputs to discrete codes. Our training objective combines several loss terms, including reconstruction loss, commitment loss, codebook loss, alignment loss, hierarchical disentanglement loss, velocity smoothing loss, penetration loss, and contact loss. The reconstruction loss measures the discrepancy between inputs and reconstructions, $\mathcal{L}_{\text{recon}} = \sum_{i=1}^{n} \left|X^i - \hat{X}^i\right|_2^2 + \left|Y - \hat{Y}\right|_2^2$. The commitment and codebook losses encourage alignment between encoder outputs and codebook embeddings:

$$\mathcal{L}_{\text{commit}}^{(l)} = \sum_{i=1}^{n} \left|\hat{Z}_H^{i,(l)} - \text{sg}\left(\bar{Z}_H^{i,(l)}\right)\right|_2^2 + \left|\hat{Z}_O^{(l)} - \text{sg}\left(\bar{Z}_O^{(l)}\right)\right|_2^2,$$

$$\mathcal{L}_{\text{codebook}}^{(l)} = \sum_{i=1}^{n} \left|\text{sg}\left(\hat{Z}_H^{i,(l)}\right) - \bar{Z}_H^{i,(l)}\right|_2^2 + \left|\text{sg}\left(\hat{Z}_O^{(l)}\right) - \bar{Z}_O^{(l)}\right|_2^2,$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The alignment loss with description cues ensures latent representations align with semantic embeddings, $\mathcal{L}_{\text{align}}^{(l)} = \sum_{i=1}^{n} \left|\bar{Z}_H^{i,(l)} - \mathbf{e}_H^{(l)}\right|_2^2 + \left|\bar{Z}_O^{(l)} - \mathbf{e}_O^{(l)}\right|_2^2$. The hierarchical disentanglement loss encourages different levels to capture distinct features:

$$\mathcal{L}_{\text{disent}} = \sum_{l=1}^{L-1} \sum_{i=1}^{n} \left|\text{Cov}\left(\bar{Z}_H^{i,(l)}, \bar{Z}_H^{i,(l+1)}\right)\right|_F^2 + \left|\text{Cov}\left(\bar{Z}_O^{(l)}, \bar{Z}_O^{(l+1)}\right)\right|_F^2,$$

where $\text{Cov}(\cdot)$ denotes covariance, and $|\cdot|_F$ is the Frobenius norm. The penetration loss penalizes interpenetration between humans and objects, $\mathcal{L}_{\text{penetration}} = \sum_{i=1}^{n} \sum_{t=1}^{T} \max(0, -d(\hat{X}_t^i, \hat{Y}_t))$, where $d(\cdot,\cdot)$ computes the signed distance between human and object meshes. The contact loss encourages plausible human-object contacts: $\mathcal{L}_{\text{contact}} = \sum_{i=1}^{n} \sum_{t=1}^{T} \left|C(\hat{X}_t^i, \hat{Y}_t) - C(X_t^i, Y_t)\right|_2^2$, where $C(\cdot,\cdot)$ computes the contact map between human and object meshes. The velocity smoothing loss encourages smooth motion transitions:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{n} \sum_{t=1}^{T-1} \left|\hat{X}_t^i - \hat{X}_{t-1}^i\right|_2^2 + \sum_{t=1}^{T-1} \left|\hat{Y}_t - \hat{Y}_{t-1}\right|_2^2.$$

The overall objective is:

$$\begin{aligned}\mathcal{L} =& \lambda_{\text{disent}}\mathcal{L}_{\text{disent}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{penetration}}\mathcal{L}_{\text{penetration}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} \\ &+ \mathcal{L}_{\text{recon}} + \sum_{l=1}^{L}\left(\lambda_{\text{commit}}^{(l)}\mathcal{L}_{\text{commit}}^{(l)} + \lambda_{\text{codebook}}^{(l)}\mathcal{L}_{\text{codebook}}^{(l)} + \lambda_{\text{align}}^{(l)}\mathcal{L}_{\text{align}}^{(l)}\right)\end{aligned}$$

with weighting coefficients λ. The hierarchical architecture enables the model to capture motion dynamics at multiple levels of abstraction, as described in Figure 1. The LLM provides hierarchical description cues for each level's abstraction, guiding the model to associate latent representations with appropriate semantic concepts. Attention mechanisms across entities capture dependencies and interactions essential for understanding coordinated actions. The hierarchical disentanglement loss encourages different levels to focus on different features, preventing redundancy and promoting specialization, leading to more meaningful and interpretable representations. Codebooks are updated using exponential moving averages, and the straight-through estimator is employed to allow gradients to flow through quantization operations. Our hierarchical VQ-VAE with description cues effectively captures the multi-scale nature of human-object interactions, learning disentangled representations at different abstraction levels and aligning them with hierarchical semantic cues to enhance interpretability and facilitate advanced control in the motion latent space.

B. *Latent Diffusion with LLM Guidance*

Our goal is to generate realistic motion involving multiple humans and objects, guided by hierarchical planning cues from a Large Language Model (LLM). We propose to utilise a denoising diffusion probabilistic model [110], [111] operating on hierarchical latents learned by a VQ-VAE (Fig. 2). The model integrates reasoning cues at multiple diffusion stages, generating motions aligned with semantic intent from the LLM. Given a dataset $\mathcal{D} = \{(X_i, \mathbf{e}_i)\}_{i=1}^{N}$, where each motion sequence X_i consists of

the trajectories of n humans $H = H^1, \ldots, H^n$ and m objects $O = O^1, \ldots, O^m$ over K time steps, and $\mathbf{e}_i = [\mathbf{e}_i^{(1)}, \ldots, \mathbf{e}_i^{(L)}]$ are the LLM-provided planning cues at L reasoning steps, our diffusion model learns to generate motion sequences conditioned on these cues. To enhance the integration of planning cues $\mathbf{E}_L = [\mathbf{e}_1, \ldots, \mathbf{e}_L]$ into the diffusion model, we associate each cue $\mathbf{e}_l$ with relevant latent codes from the VQ-VAE codebook $\mathcal{C}^{(l)}$ at level l. After training the VQ-VAE, we compute associations between latent codes $c \in \mathcal{C}^{(l)}$ and planning cues $\mathbf{e}_l$ by learning embedding functions $\phi_c^{(l)}(c)$ and $\phi_e^{(l)}(\mathbf{e}_l)$ that map codes and cues into a shared semantic space. We optimize a contrastive loss to ensure associated pairs are close in the embedding space:

$$\mathcal{L}_{\text{assoc}}^{(l)} = -\sum_{(c,\mathbf{e}_l)} \log \frac{\exp\left(\cos\left(\phi_c^{(l)}(c),\ \phi_e^{(l)}(\mathbf{e}_l)\right)/\tau\right)}{\sum_{c' \in \mathcal{C}^{(l)}} \exp\left(\cos\left(\phi_c^{(l)}(c'),\ \phi_e^{(l)}(\mathbf{e}_l)\right)/\tau\right)},$$

where τ is a temperature parameter. For each planning cue $\mathbf{e}_l$, we select the top u latent codes $\{c_{l,1}, \ldots, c_{l,u}\}$ most associated with $\mathbf{e}_l$ based on cosine similarity. We augment the cue by concatenating the embeddings of these codes, $\tilde{\mathbf{e}}_l = \left[\phi_e^{(l)}(\mathbf{e}_l);\ \phi_c^{(l)}(c_{l,1});\ \ldots;\ \phi_c^{(l)}(c_{l,u})\right]$.

These augmented cues $\tilde{\mathbf{E}}_L = [\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_L]$ are used in the denoising network, enhancing the model's capacity to generate motion sequences aligned with the planning cues by incorporating both semantic and structural information. This method leverages learned associations between planning cues and latent codes, improving the diffusion model's performance during the denoising process.

To prepare the input for the diffusion model, we aggregate the hierarchical latent codes from the VQ-VAE to form the initial latent representation $\mathbf{x}_0$. This is achieved by summing the latent codes across all levels, $\mathbf{x}_0 = \sum_{l=1}^{L} \left[z_H^{(l)},\ z_O^{(l)}\right] \in \mathbb{R}^{F \times V \times K}$, where $z_H^{(l)} \in \mathbb{R}^{F \times n \times K}$ are the latent codes for humans at level l, $z_O^{(l)} \in \mathbb{R}^{F \times m \times K}$ are object latent codes, F is the feature dimension, $V = n + m$ is the total number of nodes, and K is the number of time steps. This gives us a fully connected graph $\mathcal{G} = (\mathcal{V})$.

Our denoising network extends the U-Net architecture to handle spatio-temporal graph data. It incorporates downsampling and upsampling paths with residual connections, allowing the network to capture multi-scale temporal dependencies. We develop Motion Modeling Blocks (MM-Blocks) to process the data at different resolutions, effectively modeling the complex dynamics of motion sequences. Like previous diffusion-based models [111], we use positional encodings of the diffusion step $t \in 1, \ldots, T$ and process it using a transformer positional embedding. This embedding, denoted as $\mathbf{p}_t$, is added between the temporal layers in each MM-Block, allowing the network to condition on the noise level and adapt its denoising strategy accordingly.

In the encoding path (downsampling), at each MM-Block i, we first apply Temporal Convolutional Networks (TCNs) [112] to capture temporal dependencies at multiple scales. The TCN at layer i is defined as, $\overline{\mathbf{H}}_i = \text{TCN}(\mathbf{H}_{i-1}) + \mathbf{p}_t$, where $\mathbf{H}_{i-1} \in \mathbb{R}^{C_{i-1} \times V \times K_{i-1}}$ is the input from the previous layer, C_{i-1} is the feature dimension, K_{i-1} is the temporal length at layer $i-1$, and $\mathbf{p}_t \in \mathbb{R}^{Ci-1 \times 1 \times 1}$ is the positional embedding of the diffusion step t, broadcasted to match the dimensions. Adding the positional embedding allows the network to be aware of the current noise level, which is crucial for effective denoising, as observed by [111]. Next, we apply Graph Attention Networks (GATs) [113] to model spatial dependencies, $\mathbf{H}_i = \text{GAT}(\overline{\mathbf{H}}_i, \mathbf{A})$, where $\mathbf{A}$ is the adjacency matrix denoting fully connected graph. The GAT allows the network to focus on important interactions between humans and objects by computing attention weights for the edges in the graph.

To incorporate the reasoning cues $\tilde{\mathbf{E}}_L = [\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_L]$, we perform cross-attention between the node features $\mathbf{H}_i$ and the reasoning cues at each MM-Block, $\mathbf{H}_i^l = \text{CrossAttn}(\mathbf{H}_i, \gamma_l(t) \cdot \tilde{\mathbf{E}}_L)$, for each $l \in 1, \ldots, L$, where $\gamma_l(t)$ is a time-dependent modulation function. The time-dependent modulation function $\gamma_l(t)$ dynamically adjusts the influence of each reasoning cue $\tilde{\mathbf{e}}_L$ over the diffusion steps t, emphasizing high-level planning cues at early steps and fine-grained details later. Specifically, for reasoning cue level l, we define $\gamma_l(t) = \lambda_l \exp(-k_l t/T)$ for high-level cues (l small), where λ_l is a scaling factor, k_l controls the rate of decay, and T is the total number of diffusion steps. This function decreases over time, giving high-level cues more influence when the data is noisy. For low-level cues(l large), we use $\gamma_l(t) = \lambda_l[1 - \exp(-k_l t/T)]$, which increases over time, allowing fine-grained details to impact later steps when refining the motion. The rate of decay k_l can be a learnable parameter, enabling the model to adaptively determine the optimal influence schedule for each cue level. This modulation ensures the network focuses on appropriate aspects of the reasoning cues at each stage, effectively aligning the generated motion sequences with the LLM provided hierarchical plan. We then again pass the output through TCN and GAT layers. Finally, the outputs $\{\mathbf{H}_i^l\}_{l=1}^L$ are concatenated with $\mathbf{H}_i$ and passed to the next layer, ensuring that the semantic guidance is integrated throughout the network. We then perform downsampling to reduce the temporal dimension, $\mathbf{H}_i = \text{Downsample}(\mathbf{H}_i)$, which enables the network to capture long-range temporal dependencies. We repeat modelling spatial and temporal motion dynamics in alternate fashion, with downsampling steps.

In the decoding path (upsampling), we mirror the operations of the encoding path. Finally, we project the features back to the original latent space dimension to obtain the predicted noise:

$$\hat{\epsilon} = \text{Linear}(\mathbf{H}_0) \in \mathbb{R}^{F \times V \times K}. \tag{1}$$

The forward diffusion process [110], [111], [114] gradually adds Gaussian noise to the data, $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t;\ \sqrt{1-\beta_t}\mathbf{x}_{t-1},\ \beta_t\mathbf{I})$, where ${\beta_t}_{t=1}^T$ is a predefined noise schedule. The reverse diffusion process aims to recover the original data from the noisy observations. The denoising network learns to predict the added noise ϵ at each diffusion step t, conditioned on the current noisy data $\mathbf{x}_t$, the graph structure $\mathcal{G}$, and the reasoning cues $\mathbf{E}_L$. The training objective is to minimize the expected L2 loss between the true noise ϵ and the predicted noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}\mathbf{x}_0, \epsilon, t\left[|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathcal{G}, \mathbf{E}_L)|^2\right], \tag{2}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\ \epsilon$, with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\bar{\alpha}t = \prod s = 1^t(1-\beta_s)$. By minimizing this loss, the network learns to denoise the data effectively. Incorporating the reasoning cues $\mathbf{E}_L$ during diffusion allows the network to generate motion sequences that fulfill the intended actions and interactions. The time-dependent modulation function $\gamma_l(t)$ can be designed to emphasize high-level planning cues at early diffusion steps and fine-grained details at later steps, enabling the network to focus on different aspects of the reasoning at appropriate times.

During inference, we sample Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ conditioned on $(\mathcal{G}, \mathbf{E}_L)$ and iteratively denoise using the DDIM update [115]. Our U-shaped denoising network, augmented with Motion Modeling Blocks (MM-Blocks) and hierarchical conditioning, synthesizes motion sequences guided by planning cues derived from the LLM. The architecture leverages multi-scale temporal modeling to capture complex motion dynamics and employs Graph Attention Networks (GATs) to explicitly encode spatial interactions between humans and objects. Through dynamic modulation of cross-attention with hierarchical reasoning cues across diffusion steps, the proposed approach generates semantically coherent and physically plausible motions closely aligned with the hierarchical planning from the LLM.

IV. Experimentation and Results

*a) **Implementation Details**:* Our model consists of a hierarchical VQ-VAE with $L = 6$ levels, each with a codebook size of 512×512 (latent dimension 512) and two Conv1D blocks per level per entity (kernel size 3, residual connections), similar to T2M-GPT [13]. Vector quantization is performed using the straight-through estimator, and hierarchical planning cues are generated via GPT-4 [7] and embedded using CLIP ViT-B/32 [119], associated

Methods(CORE-4D)	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑	Methods(InterHuman)	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3						Top 1	Top 2	Top 3				
Real	$0.312^{\pm0.007}$	$0.587^{\pm0.006}$	$0.673^{\pm0.006}$	$0.005^{\pm0.0005}$	$4.124^{\pm0.019}$	$8.151^{\pm0.091}$	-	Real	$0.452^{\pm0.008}$	$0.610^{\pm0.009}$	$0.701^{\pm0.008}$	$0.273^{\pm0.007}$	$3.755^{\pm0.008}$	$7.948^{\pm0.064}$	-
TEMOS [45]	$0.065^{\pm0.006}$	$0.179^{\pm0.006}$	$0.211^{\pm0.005}$	$9.214^{\pm0.0758}$	$8.536^{\pm0.019}$	$4.671^{\pm0.091}$	$0.510^{\pm0.052}$	TEMOS [45]	$0.224^{\pm0.010}$	$0.316^{\pm0.013}$	$0.450^{\pm0.018}$	$17.375^{\pm0.043}$	$6.342^{\pm0.015}$	$6.939^{\pm0.071}$	$0.535^{\pm0.014}$
T2M [116]	$0.195^{\pm0.003}$	$0.141^{\pm0.003}$	$0.267^{\pm0.002}$	$11.258^{\pm0.0694}$	$5.867^{\pm0.013}$	$2.738^{\pm0.078}$	$1.672^{\pm0.041}$	T2M [116]	$0.238^{\pm0.012}$	$0.325^{\pm0.012}$	$0.464^{\pm0.014}$	$13.769^{\pm0.072}$	$5.731^{\pm0.018}$	$7.046^{\pm0.022}$	$1.387^{\pm0.076}$
MDM [117]	$0.163^{\pm0.013}$	$0.257^{\pm0.010}$	$0.348^{\pm0.008}$	$9.671^{\pm0.0629}$	$10.219^{\pm0.020}$	$\mathbf{7.395^{\pm0.090}}$	$\mathbf{3.526^{\pm0.074}}$	MDM [117]	$0.153^{\pm0.009}$	$0.260^{\pm0.011}$	$0.339^{\pm0.012}$	$9.167^{\pm0.056}$	$7.125^{\pm0.018}$	$\mathbf{7.602^{\pm0.045}}$	$2.355^{\pm0.080}$
MDM(GRU) [117]	$0.168^{\pm0.009}$	$0.279^{\pm0.008}$	$0.361^{\pm0.010}$	$9.587^{\pm0.1382}$	$10.228^{\pm0.025}$	$6.951^{\pm0.151}$	$3.170^{\pm0.046}$	MDM(GRU) [117]	$0.179^{\pm0.006}$	$0.299^{\pm0.005}$	$0.387^{\pm0.007}$	$32.617^{\pm0.1221}$	$9.557^{\pm0.019}$	$7.003^{\pm0.134}$	$\mathbf{3.430^{\pm0.035}}$
ComMDM [118]	$0.187^{\pm0.005}$	$0.256^{\pm0.007}$	$0.301^{\pm0.007}$	$9.217^{\pm0.0727}$	$7.541^{\pm0.023}$	$5.367^{\pm0.080}$	$0.721^{\pm0.065}$	ComMDM [118]	$0.223^{\pm0.010}$	$0.334^{\pm0.008}$	$0.466^{\pm0.012}$	$7.069^{\pm0.054}$	$6.212^{\pm0.021}$	$7.244^{\pm0.038}$	$1.822^{\pm0.052}$
InterGen [18]	$0.206^{\pm0.007}$	$0.312^{\pm0.008}$	$0.401^{\pm0.008}$	$7.217^{\pm0.2321}$	$10.251^{\pm0.017}$	$6.162^{\pm0.225}$	$3.402^{\pm0.063}$	InterGen [18]	$0.371^{\pm0.010}$	$0.515^{\pm0.012}$	$0.624^{\pm0.010}$	$5.918^{\pm0.079}$	$5.108^{\pm0.014}$	$7.387^{\pm0.029}$	$2.141^{\pm0.063}$
COLLAGE	$\mathbf{0.229^{\pm0.008}}$	$\mathbf{0.332^{\pm0.009}}$	$\mathbf{0.435^{\pm0.009}}$	$\mathbf{6.890^{\pm0.2198}}$	$\mathbf{5.526^{\pm0.016}}$	$7.373^{\pm0.237}$	$3.589^{\pm0.066}$	**COLLAGE**	$\mathbf{0.383^{\pm0.005}}$	$\mathbf{0.547^{\pm0.006}}$	$\mathbf{0.657^{\pm0.006}}$	$\mathbf{4.987^{\pm0.2061}}$	$\mathbf{4.992^{\pm0.012}}$	$7.515^{\pm0.214}$	$2.872^{\pm0.057}$
w/o Hierarchy	$0.201^{\pm0.007}$	$0.309^{\pm0.008}$	$0.411^{\pm0.008}$	$7.452^{\pm0.2381}$	$5.582^{\pm0.018}$	$6.995^{\pm0.224}$	$3.209^{\pm0.058}$	w/o Hierarchy	$0.355^{\pm0.009}$	$0.521^{\pm0.010}$	$0.632^{\pm0.009}$	$5.543^{\pm0.2154}$	$5.048^{\pm0.015}$	$7.137^{\pm0.201}$	$2.492^{\pm0.061}$
w/o LLM	$0.208^{\pm0.007}$	$0.315^{\pm0.008}$	$0.419^{\pm0.008}$	$7.235^{\pm0.2305}$	$5.561^{\pm0.017}$	$6.549^{\pm0.230}$	$3.152^{\pm0.063}$	w/o LLM	$0.362^{\pm0.008}$	$0.528^{\pm0.009}$	$0.639^{\pm0.008}$	$5.326^{\pm0.2087}$	$5.027^{\pm0.014}$	$6.691^{\pm0.207}$	$2.435^{\pm0.059}$
w/o Time Modulation	$0.218^{\pm0.008}$	$0.317^{\pm0.009}$	$0.420^{\pm0.009}$	$7.071^{\pm0.2251}$	$5.556^{\pm0.017}$	$7.263^{\pm0.234}$	$3.474^{\pm0.065}$	w/o Time Modulation	$0.372^{\pm0.007}$	$0.536^{\pm0.008}$	$0.647^{\pm0.007}$	$5.162^{\pm0.2129}$	$5.021^{\pm0.013}$	$7.405^{\pm0.211}$	$2.767^{\pm0.062}$

TABLE I: Experimental results and Ablation studies for text-conditioned interaction generation on the CORE-4D and InterHuman datasets, where $\pm$ indicates 95% confidence interval and $\rightarrow$ means the closer the better. **Bold** indicates best results.

Test Set	Method	$RR.J_e$ (mm, ↓)	$RR.V_e$ (mm, ↓)	C_{acc} (%, ↑)	FID (↓)
S1	MDM [117]	138.0 (± 0.3)	194.6 (± 0.2)	76.9 (± 0.5)	7.7 (± 0.2)
	OMOMO [17]	137.8 (± 0.2)	196.7 (± 0.3)	78.2 (± 0.5)	8.3 (± 0.6)
	COLLAGE	**131.2 (± 0.2)**	**185.1 (± 0.2)**	**80.5 (± 0.4)**	**7.2 (± 0.2)**
S2	MDM [117]	145.9 (± 0.2)	208.2 (± 0.2)	76.7 (± 0.1)	7.7 (± 0.2)
	OMOMO [17]	145.2 (± 0.6)	209.9 (± 1.0)	77.8 (± 0.3)	8.3 (± 1.0)
	COLLAGE	**138.5 (± 0.5)**	**198.7 (± 0.8)**	**79.9 (± 0.2)**	**7.3 (± 0.8)**

TABLE II: Quantitative results on object-conditioned interaction synthesis on CORE-4D.

with the VQ-VAE codebooks through contrastive learning (temperature $\tau = 0.07$, top $u = 8$ latent codes per level). The latent diffusion model is based on a U-Net architecture with $M = 4$ Motion Modeling Blocks (MM-Blocks), each consisting of Temporal Convolutional Networks (TCNs) with kernel sizes $\{3, 5, 7\}$ [112] and Graph Attention Networks (GATs) with 8 attention heads [113], capturing spatio-temporal dependencies.For training, we use the Adam optimizer for VQ-VAE with a learning rate of 1×10^{-4} and AdamW [120] for the diffusion model with a learning rate of 2×10^{-4}, both with cosine annealing, gradient clipping (max norm 1.0), and weight decay of 1×10^{-5}. For CORE-4D [17], we train for 50K iterations with a learning rate of 2×10^{-4} and an additional 30K iterations with a reduced learning rate of 1×10^{-5}. For InterHuman [18], we train for 200K iterations at 2×10^{-4} and 100K iterations at 1×10^{-5}. We use a batch size of 256 for both datasets and apply the Adam optimizer with $[\beta_1, \beta_2] = [0.9, 0.99]$ and an exponential moving constant $\lambda = 0.99$. Loss terms include $\lambda_{\text{recon}} = 1.0$, $\lambda^{(l)}_{\text{commit}} = 0.25$ per level, $\lambda^{(l)}_{\text{codebook}} = 0.25$ per level, $\lambda^{(l)}_{\text{align}} = 0.5$ per level, $\lambda_{\text{smooth}} = 0.1$ [48], $\lambda_{\text{penetration}} = 10.0$ [69], and $\lambda_{\text{contact}} = 5.0$. The hierarchical disentanglement loss is weighted by $\lambda_{\text{disent}} = 1.0$. The diffusion model uses 1000 diffusion steps and we test for 5, 15, 55, 100 DDIM [121] sampling steps during inference. The hierarchical cue modulation function applies exponential decay for high-level cues and increasing influence for low-level cues across diffusion steps

We train COLLAGE on the CORE-4D dataset [17], which contains 998 motion sequences of human-object-human interactions spanning 5 object categories. We annotates the motion sequences with textual descriptions, the annotated text-motion dataset has an average length of 8.54 words, totaling 8,542 words. We split the dataset into training, validation, and test sets with a ratio of 0.8, 0.05, and 0.15, respectively. We also evaluate our model on the InterHuman dataset [18] for multi-human generation, which includes 6,022 motions with 16,756 unique descriptions. We use same train/test formulation as [18]. We additionally also train our model for single human motion generation on KIT-ML [122] and HumanML3D [123] Dataset, the visualisations and comparisons are available on the paper website.

*b) **Evaluation Metrics**:* For text-conditioned generation on CORE-4D and InterHuman, we adopt the metrics from InterGen [18]: (1) *FID*, (2) *R-Precision*, (3) *Diversity*, (4) *Multimodality (MModality)*, and (5) *MM Dist*. For additional tasks on CORE-4D, we follow their own metrics [17]: (1) *RR.J$_e$*, (2) *RR.V$_e$*, and (3) *C$_{acc}$*. All evaluations are run 20 times (except MModality, 5 times) with average results reported with a 95% confidence interval. For detailed descriptions of these metrics, we refer readers to [17], [18].

*c) **Baselines**:* For text-conditioned generation on the CORE-4D dataset, we compare COLLAGE against state-of-the-art methods, including TEMOS [45], T2M [116], MDM [117], MDM-GRU [117], ComMDM [118], and InterGen [18]. We modify these models to handle two-person interactions and train them on the CORE-4D dataset. For the additional tasks on the CORE-4D dataset, we compare against MDM [117], a one-stage motion diffusion model, and OMOMO [17], a two-stage approach for object-conditioned human motion generation.

A. Results

*1) **Text-Conditioned Generation**:*

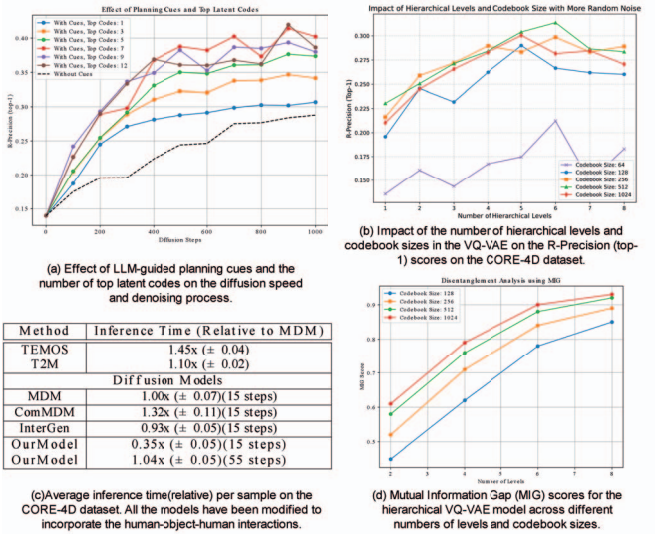
a) Results on CORE-4D: Table I (left) presents the results of text-conditioned generation on the CORE-4D dataset. COLLAGE outperforms all baselines across most metrics, achieving the highest R-Precision scores, lowest FID, and best diversity. The hierarchical VQ-VAE effectively captures multi-scale motion dynamics, while the LLM-guided diffusion model generates motions that align well with the textual descriptions. The incorporation of hierarchical planning cues enables COLLAGE to generate more coherent and diverse interactions compared to the baselines.

b) Results on InterHuman: We further evaluate COLLAGE on the InterHuman dataset for multi-human generation. Table I (right) shows the comparison with state-of-the-art methods. COLLAGE achieves superior performance across nearly all metrics, demonstrating its effectiveness in generating diverse and realistic multi-human interactions. The hierarchical modeling of motion dynamics and the incorporation of LLM-guided planning enable COLLAGE to better capture the complexities of human-human interactions compared to the baselines.

2) Object-Conditioned Generation on CORE-4D: We evaluate COLLAGE on the task of object-conditioned human motion generation on the CORE-4D dataset. Given an object geometry sequence, the goal is to generate two-person collaboration motions using the SMPL-X model [124]. Table II presents the quantitative results, comparing COLLAGE with MDM [117] and OMOMO [17]. COLLAGE achieves the lowest joint and vertex position errors, highest contact accuracy , and best motion quality (FID) on both test sets (S1 and S2). The hierarchical modeling and LLM guidance enable COLLAGE to generate more precise and realistic human-object interactions compared to the baselines.

3) Ablation Studies: We conduct ablations on the CORE-4D dataset to validate the effectiveness of the proposed components in COLLAGE. Table I (bottom) presents the results. Removing the hierarchical structure in the VQ-VAE (w/o Hierarchy) significantly drops performance across metrics, highlighting the importance of modeling motion dynamics at multiple scales. Removing LLM guidance (w/o LLM) also decreases performance, demonstrating the effectiveness of incorporating hierarchical planning cues. Replacing time-dependent modulation with fixed weighting (w/o Time Modulation) degrades performance, indicating the benefit of adaptively adjusting the influence of planning cues over diffusion steps. These studies confirm that hierarchical VQ-VAE, LLM guidance, and time-dependent modulation are essential components of COLLAGE, contributing to its superior performance in generating collaborative human-object-human interactions.

a) Impact of Hierarchical Levels and Codebook Size: We evaluate COLLAGE's performance with different numbers of hierarchical levels and codebook sizes in the VQ-VAE architecture. Figure 3 shows the R-Precision (top-1) scores on the CORE-4D dataset as we vary the number of levels from 1 to 8 and the



(a) Effect of LLM-guidad planning cues and the number of top latent codes on the diffusion speed and denoising process.

(b) Impact of the number of hierarchical levels and codebook sizes in the VQ-VAE on the R-Precision (top-1) scores on the CORE-4D dataset.

Method	Inference Time (Relative to MDM)
TEMOS	1.45x (± 0.04)
T2M	1.10x (± 0.02)
Diffusion Models	
MDM	1.00x (± 0.07)(15 steps)
ComMDM	1.32x (± 0.11)(15 steps)
InterGen	0.93x (± 0.05)(15 steps)
OurModel	0.35x (± 0.05)(15 steps)
OurModel	1.04x (± 0.05)(55 steps)

(c)Average inference time(relative) per sample on the CORE-4D dataset. All the models have been modified to incorporate the human-object-human interactions.

(d) Mutual Information Gap (MIG) scores for the hierarchical VQ-VAE model across different numbers of levels and codebook sizes.

Fig. 3: Ablation Studies

codebook sizes from 128 to 1024. Increasing the number of levels initially improves performance, with the best results achieved at 6 levels for all codebook sizes, indicating that the hierarchical structure effectively captures motion dynamics at multiple scales. However, further increasing levels beyond 6 slightly degrades performance, likely due to overfitting and increased model complexity. We find that codebook sizes of 256 and 512 strike a good balance between expressiveness and efficiency, with 512 yielding the best overall performance across different levels.

b) Effect of Planning Cues and Top Latent Codes: We analyze the effect of LLM-guided planning cues and the number of top latent codes on diffusion speed, i.e., the number of steps required to generate high-quality motions. Figure 3 compares the R-Precision (top-1) scores of COLLAGE with and without planning cues at different diffusion steps and the impact of using different numbers of top latent codes. Incorporating planning cues significantly accelerates the diffusion process, allowing the model to generate high-quality motions in fewer steps. Increasing the number of top latent codes further improves performance, with diminishing returns beyond 7-8 codes, demonstrating the effectiveness of cues and top latent codes in guiding the denoising process and highlighting the efficiency gains achieved by our approach.

c) Impact of Hierarchical Structure and Codebook Size on Latent Disentanglement: Our ablation study investigates the influence of the hierarchical structure and codebook size on the disentanglement of latent representations in our VQ-VAE model. Figure 3 reveals that increasing the number of levels consistently improves disentanglement, as evidenced by higher MIG [125] scores, suggesting that the hierarchical structure effectively captures disentangled representations at different levels of abstraction, with higher levels focusing on more abstract patterns. Similarly, larger codebook sizes lead to better disentanglement for a given number of levels, indicating that a larger codebook enables more expressive and disentangled representations. However, the diminishing gaps between the lines at higher levels and codebook sizes imply that the benefits of increasing these hyperparameters saturate beyond certain thresholds, highlighting the importance of balancing model complexity and computational efficiency when designing hierarchical VQ-VAE architectures for learning disentangled representations of human motion data. These ablation studies provide insights into the functioning and effectiveness of COLLAGE. The evaluation of different hierarchical levels and codebook sizes highlights the importance of finding the optimal balance between model complexity and expressiveness. The latent space disentanglement analysis demonstrates the hierarchical VQ-VAE's ability to capture distinct and independent features at different levels of abstraction. The analysis of diffusion speed showcases the efficiency gains achieved by incorporating planning cues and utilizing top latent codes, enabling faster generation of high-quality motions.

B. Qualitative Analysis

Attached video presents qualitative examples of generated collaborative human-object-human interactions by COLLAGE and the baselines on the CORE-4D dataset. COLLAGE generates more realistic and coherent interactions compared to the baselines, accurately capturing the coordination between the two humans and their interactions with the object. The generated motions align well with the input text descriptions, demonstrating the effectiveness of the LLM-guided planning cues in controlling the generation process. In contrast, the baselines struggle to generate precise and coordinated interactions, often resulting in unrealistic or inconsistent motions.

a) Runtime Analysis: We compare the inference time of COLLAGE with the baselines on the CORE-4D dataset. Table 3 presents the average relative inference time per sample for each method with respect to the MDM [117] runtime. COLLAGE achieves significantly faster inference. Furthermore, we tested our performance for different DDIM sampling steps (5, 15, 55, and 100 in the main model). As expected, with an increase in the number of steps, the generation quality improves. However, the improvement in generation quality from 55 to 100 steps is minor, while the generation time nearly doubles. Notably, we observe that with just 15 steps, we achieve relatively better generation quality than InterGen [18], and our model with 15 steps is faster than the MDM model (15 DDIM steps) by 65%. Additionally, our near-best performance (55 DDIM steps) has a runtime similar to MDM. Thus, our hierarchical VQ-VAE enables efficient compression and decompression, while the LLM-guided cues and codebook associations provide curated motion priors, allowing the diffusion model to denoise in fewer steps and generate smoother motion faster. In contrast, other baselines require longer inference times due to their complex architectures and the need to generate motions in the original high-dimensional space.

V. Discussion and Limitations

The experimental results demonstrate the effectiveness of COLLAGE in generating realistic and diverse collaborative human-object-human interactions. The hierarchical VQ-VAE architecture captures motion dynamics at multiple scales, while the LLM-guided diffusion model generates motions aligned with textual descriptions and planning cues. Incorporating hierarchical planning cues from the LLM enables more coherent and controllable generation, evidenced by COLLAGE's superior performance across various metrics and datasets. However, there are some limitations to our approach. First, COLLAGE does not explicitly model physical interactions between humans and objects, relying on learned motion priors to generate plausible interactions. Incorporating explicit physics modeling could further improve the realism and consistency of generated motions. Second, the current approach generates motions from scratch based on input text and object geometry but does not allow fine-grained editing or control over specific motion aspects. Extending the model to support motion editing and user-guided refinement could enhance its practical utility. **Despite** these limitations, COLLAGE represents a significant step towards generating realistic and diverse collaborative human-object-human interactions. Our approach can be seamlessly extended to collaborative interactions between humanoid robots and objects. By training our model on humanoid robot motion data, we can generate realistic and diverse interactions that mimic human-like behaviors. This extension has significant implications for deploying humanoid robots in various real-world scenarios, where they are expected to collaborate with objects and other agents in a human-like manner. **The** proposed approach opens new possibilities for applications in robotics, virtual reality, and computer graphics, where generating plausible and coordinated multi-agent interactions is crucial. Future work could explore incorporating explicit physics modeling, supporting motion editing and user control, and extending the approach to handle a larger variety of objects and interaction scenarios.

Acknowledgement

This material is based upon work supported in part by the DEVCOM Army Research Laboratory under cooperative agreement W911NF2020221.

REFERENCES

[1] C. Diller and A. Dai, "Cg-hoi: Contact-guided 3d human-object interaction generation," *arXiv*, 2023.

[2] J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, and C. K. Liu, "Controllable human-object interaction synthesis," *arXiv*, 2023.

[3] X. Peng, Y. Xie, Z. Wu, V. Jampani, and H. Jiang, "Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models," *arXiv*, 2023.

[4] Q. Wu, Y. Shi, X. Huang, J. Yu, L. Xu, and J. Wang, "THOR: Text to human-object interaction diffusion via relation intervention," *arXiv*, 2024.

[5] S. Xu, Z. Wang, Y.-X. Wang, and L.-Y. Gui, "Interdreamer: Zero-shot text to 3d dynamic human-object interaction," *arXiv*, 2024.

[6] J. Zhang, J. Zhang, Z. Song, Z. Shi, C. Zhao, Y. Shi, J. Yu, L. Xu, and J. Wang, "Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment," in *CVPR*, 2024, pp. 516–526.

[7] OpenAI, "ChatGPT," https://chat.openai.com/, 2023.

[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv*, 2023.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *NeurIPS*, 2022.

[10] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *ICML*, 2022.

[11] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *CoRL*, 2023.

[12] J. Sun, Q. Zhang, Y. Duan, X. Jiang, C. Cheng, and R. Xu, "Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning," in *ICRA*, 2024.

[13] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2M-GPT: Generating human motion from textual descriptions with discrete representations," in *CVPR*, 2023.

[14] H. Kong, K. Gong, D. Lian, M. B. Mi, and X. Wang, "Priority-centric human motion generation in discrete latent space," in *ICCV*, 2023.

[15] J. Li, R. Villegas, D. Ceylan, J. Yang, Z. Kuang, H. Li, and Y. Zhao, "Task-generic hierarchical human motion prior using vaes," in *3DV*. IEEE, 2021, pp. 771–781.

[16] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *CVPR*, 2023, pp. 18 000–18 010.

[17] C. Zhang, Y. Liu, R. Xing, B. Tang, and L. Yi, "Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement," *arXiv*, 2024.

[18] H. Liang, W. Zhang, W. Li, J. Yu, and L. Xu, "InterGen: Diffusion-based multi-human motion generation under complex interactions," *arXiv*, 2023.

[19] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *ACMMM*, 2020.

[20] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *ICCV*, 2021.

[21] T. Lee, G. Moon, and K. M. Lee, "Multiact: Long-term 3d human motion generation from multiple action labels," in *AAAI*, 2023.

[22] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "Teach: Temporal action composition for 3d humans," in *3DV*, 2022.

[23] Y. Yuan and K. Kitani, "DLow: Diversifying latent flows for diverse human motion prediction," in *ECCV*, 2020.

[24] G. Barquero, S. Escalera, and C. Palmero, "BeLFusion: Latent diffusion for behavior-driven human motion prediction," in *ICCV*, 2023.

[25] L.-H. Chen, J. Zhang, Y. Li, Y. Pang, X. Xia, and T. Liu, "HumanMAC: Masked motion completion for human motion prediction," in *ICCV*, 2023.

[26] S. Xu, Y. Wang, and L. Gui, "Diverse human motion prediction guided by multi-level spatial-temporal anchors," in *ECCV*, 2022.

[27] S. Xu, Y.-X. Wang, and L. Gui, "Stochastic multi-person 3d motion forecasting," in *ICLR*, 2023.

[28] M. Kaufmann, E. Aksan, J. S, F. R, R. Z, and O. Hilliges, "Convolutional autoencoders for human motion infilling," in *3DV*, 2020.

[29] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "GMD: Controllable human motion synthesis via guided diffusion models," in *ICCV*, 2023.

[30] D. Rempe, Z. Luo, X. Bin P, Y. Yuan, K. Kitani, K. Kreis, S. Fidler, and O. Litany, "Trace and pace: Controllable pedestrian animation via guided trajectory diffusion," in *CVPR*, 2023.

[31] Y. Xie, V. Jampani, L. Zhong, D. Sun, and H. Jiang, "OmniControl: Control any joint at any time for human motion generation," *arXiv*, 2023.

[32] W. Wan, Z. Dou, T. Komura, W. Wang, D. Jayaraman, and L. Liu, "Tlcontrol: Trajectory and language control for human motion synthesis," *arXiv*, 2023.

[33] Z. Cao, H. Gao, K. Mangalam, Q. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *ECCV*, 2020.

[34] M. Hassan, P. Ghosh, D. Tzionas, and M. J. Black, "Populating 3d scenes by learning human-scene interaction," in *CVPR*, 2021.

[35] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang, "Synthesizing long-term 3d human motion and interaction in 3d scenes," in *CVPR*, 2021.

[36] J. Wang, S. Yan, B. Dai, and D. Lin, "Scene-aware generative network for human motion synthesis," in *CVPR*, 2021.

[37] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai, "Towards diverse and natural scene-aware 3d human motion synthesis," in *CVPR*, 2022.

[38] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *CVPR*, 2023.

[39] K. Zhao, S. Wang, Y. Zhang, and S. Tang, "Compositional human-scene interaction synthesis with semantic control," in *ECCV*, 2022.

[40] K. Zhao, Y. Zhang, S. Wang, T. Beeler, and S. Tang, "Synthesizing diverse human motions in 3d indoor scenes," in *ICCV*, 2023.

[41] P. Tendulkar, D. Surís, and C. Vondrick, "FLEX: Full-body grasping without full-body grasps," in *CVPR*, 2023.

[42] W. Zhang, R. Dabral, T. Leimkühler, V. Golyanik, M. Habermann, and C. Theobalt, "ROAM: Robust and object-aware motion generation using neural pose descriptors," *arXiv*, 2023.

[43] M. Petrovich, M. J. Black, and G. Varol, "TMR: Text-to-motion retrieval using contrastive 3d human motion synthesis," in *ICCV*, 2023.

[44] C. Guo, X. Zuo, S. Wang, and L. Cheng, "Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts," in *ECCV*, 2022.

[45] M. Petrovich, M. J. Black, and G. Varol, "TEMOS: Generating diverse human motions from textual descriptions," in *ECCV*, 2022.

[46] X. Chen, B. Jiang, W. Liu, Z. H, B. Fu, T. C, and G. Yu, "Executing your commands via motion diffusion in latent space," in *CVPR*, 2023.

[47] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "MotionDiffuse: Text-driven human motion generation with diffusion model," *arXiv*, 2022.

[48] Y. Zhang, D. Huang, B. Liu, S. Tang, Y. Lu, L. Chen, L. Bai, Q. Chu, N. Yu, and W. Ouyang, "Motiongpt: Finetuned llms are general-purpose motion generators," *arXiv*, 2023.

[49] 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 *CVPR*, 2023.

[50] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *ECCV*, 2022.

[51] C. Ahuja and L. Morency, "Language2pose: Natural language grounded pose forecasting," in *3DV*, 2019.

[52] 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 *CVPR*, 2022.

[53] J. Kim, J. Kim, and S. Choi, "Flame: Free-form language-based motion synthesis & editing," in *AAAI*, 2023.

[54] S. Lu, L.-H. Chen, A. Zeng, J. Lin, R. Zhang, L. Zhang, and H.-Y. Shum, "HumanTOMATO: Text-aligned whole-body motion generation," *arXiv*, 2023.

[55] D. Raab, T. Leibovitch, G. Tevet, M. Arar, A. H. Bermano, and D. Cohen-Or, "Single motion diffusion," *arXiv*, 2023.

[56] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," *arXiv*, 2023.

[57] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt, "MoFusion: A framework for denoising-diffusion-based motion synthesis," in *CVPR*, 2023.

[58] D. Wei, X. Sun, H. Sun, B. Li, S. Hu, W. Li, and J. Lu, "Understanding text-driven motion synthesis with keyframe collaboration via diffusion models," *arXiv*, 2023.

[59] Z. Zhang, R. Liu, K. Aberman, and R. Hanocka, "TEDi: Temporally-entangled diffusion for long-term motion synthesis," *arXiv*, 2023.

[60] P. J. Yazdian, E. Liu, L. Cheng, and A. Lim, "MotionScript: Natural language descriptions for expressive 3d human motions," *arXiv*, 2023.

[61] G. Barquero, S. Escalera, and C. Palmero, "Seamless human motion composition with blended positional encodings," in *CVPR*, 2024.

[62] W. Zhou, Z. Dou, Z. Cao, Z. Liao, J. Wang, W. Wang, Y. Liu, T. Komura, W. Wang, and L. Liu, "EMDM: Efficient motion diffusion model for fast, high-quality motion generation," *arXiv*, 2023.

[63] S. Ma, Q. Cao, J. Zhang, and D. Tao, "Contact-aware human motion generation from textual descriptions," *arXiv*, 2024.

[64] Y. Liu, C. Chen, and L. Yi, "Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting," *arXiv*, 2023.

[65] Z. Wang, J. Wang, D. Lin, and B. Dai, "InterControl: Generate human motion interactions by controlling every joint," *arXiv*, 2023.

[66] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "ReMoS: Reactive 3d motion synthesis for two-person interactions," *arXiv*, 2023.

[67] 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 *CVPR*, 2024.

[68] P. Cong, Z. W. Dou, Y. Ren, W. Yin, K. Cheng, Y. Sun, X. Long, X. Zhu, and Y. Ma, "LaserHuman: Language-guided scene-aware human motion generation in free environment," *arXiv*, 2024.

[69] Q. Li, J. Wang, C. C. Loy, and B. Dai, "Task-oriented human-object interactions generation with implicit neural representations," *arXiv*, 2023.

[70] Y. Ye, X. Li, A. Gupta, S. De Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu, "Affordance diffusion: Synthesizing hand-object interactions," in *CVPR*, 2023.

[71] J. Zheng, G. Zheng, L. Fang, Y. Liu, and L. Yi, "CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis," in *CVPR*, 2023.

[72] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll, "Toch: Spatio-temporal object-to-hand correspondence for motion refinement," in *ECCV*, 2022.

[73] H. Zhang, S. Christen, Z. Fan, L. Zheng, J. Hwangbo, J. Song, and O. Hilliges, "ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation," *arXiv*, 2023.

[74] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "Chore: Contact, human and object reconstruction from a single rgb image," in *ECCV*, 2022.

[75] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa, "Perceiving 3d human-object spatial arrangements from a single image in the wild," in *ECCV*, 2020.

[76] X. Wang, G. Li, Y. Kuo, M. Kocabas, E. Aksan, and O. Hilliges, "Reconstructing action-conditioned human-object interactions using commonsense knowledge priors," in *3DV*, 2022.

[77] I. A. Petrov, R. Marin, J. Chibane, and G. Pons-Moll, "Object pop-up: Can we infer 3d objects and their poses from human interactions alone?" in *CVPR*, 2023.

[78] Z. Hou, B. Yu, and D. Tao, "Compositional 3d human-object neural animation," *arXiv*, 2023.

[79] T. Kim, S. S, and H. J, "NCHO: Unsupervised learning for neural 3d composition of humans and objects," in *ICCV*, 2023.

[80] L. Li and A. Dai, "GenZI: Zero-shot 3d human-scene interaction generation," in *CVPR*, 2024.

[81] Y. Yang, W. Zhai, and Z.-J. Zha, "LEMON: Learning 3d human-object interaction relation from 2d images," in *CVPR*, 2024.

[82] H. Kim, S. Han, P. Kwon, and H. Joo, "Zero-shot learning for the primitives of 3d affordance in general objects," *arXiv*, 2024.

[83] S. Starke, H. Zhang, T. K, and J. Saito, "Neural state machine for character-scene interactions." *ACM Trans. Graph.*, vol. 38, 2019.

[84] O. Taheri, V. Choutas, and D. Tzionas, "GOAL: Generating 4d whole-body motion for hand-object grasping," in *CVPR*, 2022.

[85] Y. Wu, J. Wang, F. Yu, and S. Tang, "SAGA: Stochastic whole-body grasping with contact," in *ECCV*, 2022.

[86] X. Zhang, B. L. Bhatnagar, S. Starke, V. G, and G. PM, "COUCH: Towards controllable human-chair interactions," in *ECCV*, 2022.

[87] J. Lee and H. Joo, "Locomotion-Action-Manipulation: Synthesizing human-scene interactions in complex 3d environments," in *ICCV*, 2023.

[88] X. Xu, H. Joo, G. Mori, and M. Savva, "D3D-HOI: Dynamic 3d human-object interactions from videos," *arXiv*, 2021.

[89] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "IMoS: Intent-driven full-body motion synthesis for human-object interactions," *arXiv*, 2022.

[90] W. Wan, L. Yang, T. Komura, and W. Wang, "Learn to predict how humans manipulate large-sized objects from interactive motions," *IEEE RA-L*, 2022.

[91] H. Razali and Y. Demiris, "Action-conditioned generation of bimanual object manipulation sequences," in *AAAI*, 2023.

[92] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The kit bimanual manipulation dataset," in *Humanoids*, 2021.

[93] S. Xu, Z. Li, and L.-Y. Gui, "InterDiff: Generating 3d human-object interactions with physics-informed diffusion," in *ICCV*, 2023.

[94] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," *ACM TOG*, vol. 42, no. 6, pp. 1–11, 2023.

[95] Y.-W. Chao, J. Yang, W. Chen, and J. Deng, "Learning to sit: Synthesizing human-chair interactions via hierarchical control," in *AAAI*, 2021.

[96] J. Merel, S. Tunyasuvunakool, and N. Heess, "Catch & carry: reusable neural controllers for vision-guided whole-body tasks," *ACM TOG*, vol. 39, no. 4, pp. 39–1, 2020.

[97] M. Hassan, Y. G, T. Wang, M. Black, S. Fidler, and X. B. Peng, "Synthesizing physical character-scene interactions," in *SIGGRAPH*, 2023.

[98] J. Bae, J. Won, D. Lim, C.-H. Min, and Y. M. Kim, "Pmp: Learning to physically interact with environments using part-wise motion priors," in *SIGGRAPH*, 2023.

[99] Z. Yang, K. Yin, and L. Liu, "Learning to use chopsticks in diverse gripping styles," *ACM TOG*, vol. 41, no. 4, pp. 1–17, 2022.

[100] Z. Xie and M. Panne, "Learning soccer juggling skills with layer-wise mixture-of-experts," in *SIGGRAPH*, 2022.

[101] Z. Xie, J. Tseng, M. Panne, and C. K. Liu, "Hierarchical planning and control for box loco-manipulation," *arXiv*, 2023.

[102] L. Pan, J. Wang, B. Huang, J. Zhang, H. Wang, X. Tang, and Y. Wang, "Synthesizing physically plausible human motions in 3d scenes," *arXiv*, 2023.

[103] J. Braun, S. Christen, M. Kocabas, E. Aksan, and O. Hilliges, "Physically plausible full-body hand-object interaction synthesis," *arXiv*, 2023.

[104] Y. Wang, J. Lin, A. Zeng, Z. Luo, J. Zhang, and L. Zhang, "PhysHOI: Physics-based imitation of dynamic human-object interaction," *arXiv*, 2023.

[105] J. Cui, T. Liu, N. Liu, Y. Yang, Y. Zhu, and S. Huang, "AnySkill: Learning open-vocabulary physical skill for interactive agents," in *CVPR*, 2024.

[106] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol, "SINC: Spatial composition of 3d human motions for simultaneous action generation," in *ICCV*, 2023.

[107] H. Yao, Z. Song, Y. Zhou, T. Ao, B. Chen, and L. Liu, "MoConVQ: Unified physics-based motion control via scalable discrete representations," *arXiv*, 2023.

[108] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "MotionGPT: Human motion as a foreign language," in *NeurIPS*, 2023.

[109] Z. Xiao, T. Wang, J. Wang, J. Cao, W. Zhang, B. Dai, D. Lin, and J. Pang, "Unified human-scene interaction via prompted chain-of-contacts," *arXiv*, 2023.

[110] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[111] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *ICML*. PMLR, 2021.

[112] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv*, 2018.

[113] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv*, 2017.

[114] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "Csdi: Conditional score-based diffusion models for probabilistic time series imputation," *NeurIPS*, 2021.

[115] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv*, 2020.

[116] 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 *CVPR*, 2022, pp. 5152–5161.

[117] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[118] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," *arXiv*, 2023.

[119] A. Radford, J. W. Kim, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[120] I. Loshchilov, "Decoupled weight decay regularization," *arXiv*, 2017.

[121] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv*, 2020.

[122] M. Plappert and T. Asfour, "The kit motion-language dataset," *Big data*, vol. 4, no. 4, pp. 236–252, 2016.

[123] 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 *CVPR*, 2022.

[124] G. Pavlakos, V. Choutas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019.

[125] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *NeurIPS*, vol. 31, 2018.