SIMS: Simulating Stylized Human-Scene Interactions with Retrieval-Augmented Script Generation

Wenjia Wang[1] Liang Pan[1,2] Zhiyang Dou[1] Jidong Mei[1] Zhouyingcheng Liao[1]
Yuke Lou[1] Yifan Wu[1] Lei Yang[2] Jingbo Wang[2†] Taku Komura[1†]

[1] The University of Hong Kong [2] Shanghai AI Laboratory

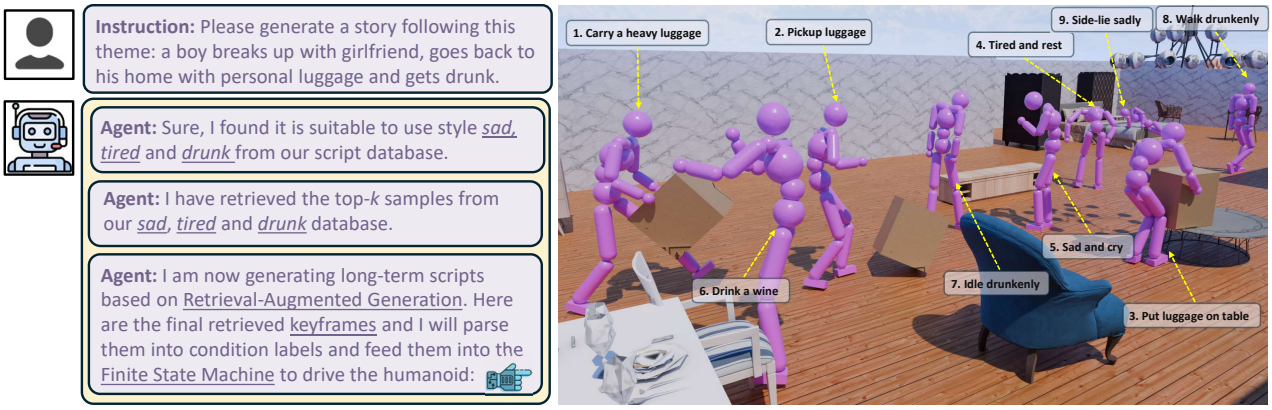


Figure 1. SIMS enables physically simulated characters to perform diverse skills within complex 3D scenes given long-term daily narratives and scene inputs. Our character could perform versatile skills, including Locomotions, Human Scene Interactions and Dynamic Object Interactions with diverse styles while accomplishing physically plausible contacts and obstacle avoidance. Left: a dialogue-based retrieval-augmented script generation process. Right: a skillful humanoid performing diverse stylized interactions in a 3D scene.

Abstract

Simulating stylized human-scene interactions (HSI) in physical environments is a challenging yet fascinating task. Prior works emphasize long-term execution but fall short in achieving both diverse style and physical plausibility. To tackle this challenge, we introduce a novel hierarchical framework named SIMS that seamlessly bridges high-level script-driven intent with a low-level control policy, enabling more expressive and diverse human-scene interactions. Specifically, we employ Large Language Models with Retrieval-Augmented Generation (RAG) to generate coherent and diverse long-form scripts, providing a rich foundation for motion planning. A versatile multi-condition physics-based control policy is also developed, which leverages text embeddings from the generated scripts to encode stylistic cues, simultaneously perceiving environmental geometries and accomplishing task goals. By integrating the retrieval-augmented script generation with the multi-condition controller, our approach provides a unified solution for generating stylized HSI motions. We further introduce a comprehensive planning dataset produced by RAG and a stylized motion dataset featuring diverse locomotions and interactions. Extensive experiments demonstrate SIMS's effectiveness in executing various tasks and generalizing across different scenarios, significantly outperforming previous methods. Project page: https://wenjiawang0312.github.io/projects/sims/.

1. Introduction

Developing skillful characters with a broad repertoire of motor skills, such as walking, sitting, and reaching—while facilitating rich interactions with their environments has long been a desirable goal for animation, robotics, and VR/AR applications. In particular, achieving *long-term, stylized, and physically plausible* interactions with diverse styles and intricate details is crucial for bringing characters and narratives to life.

Motion generation has been a popular research topic recent years [32, 49, 52, 60]. Concerning human-scene interactions, previous works [12, 35, 43, 46, 58, 59] have explored long-term motion generation for kinematics-based

*†: equal advising.

Method	Physical-Plausibe	Planner		Controller			Incorporated Skills						
		Automatic	Style-Diversity	Text-Aware	Scene-Aware	Skill-Scalability	Walk	Sit	Lie	GetUp	Reach	Idle	Carry
NSM[35]	✗	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	✓	✓
SAMP[12]	✗	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	✗	✗
Humanise[46]	✗	✗	✗	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗
AffordMotion[47]	✗	✗	✗	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗
TesMo[54]	✗	✗	✗	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗
InterScene[26]	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗
UniHSI[50]	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	✗
SIMS (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparision of Kinematics-Based(upper 5) and Physics-Based(lower 3) Long-term Human Scene Interaction methods.

human-scene interactions. However, they typically suffer from severe physical artifacts such as penetration and foot skating. To address these issues, recent studies [13, 17, 26, 27, 36, 50, 51, 55] have started incorporating physics simulators, i.e., [23] to produce more physically plausible motions. Despite these advancements, the frameworks are limited to a small number of specific skills and task objectives, lacking diversity. Moreover, their planning results are often simplistic by following chronological lists [26, 51] or focusing solely on contacts [50]. This stands in contrast to real-world situations where body language in human motion and interactions directly convey a large number of ***emotional or stylized*** states. For example, a person sitting on a chair with their head down and supporting it with their hands often conveys a sense of depression.

To address the aforementioned challenges, we propose a novel framework terms SIMS, (**S**multating styl**I**zed hu**M**an **S**cene interactions). Specifically, SIMS utilizes an LLM [1] as a powerful high-level motion planner and physical policies as low-level controllers equipped with diverse motor skills. Inspired by Retrieval-Augmented Generation [18], to generate semantically rich scripts, we develop a method of first creating a short script database and then retrieving and generating longer scripts. Each short script includes several keyframes detailing stylized interactions that the low-level control policy can effectively execute. We then retrieve the top-k short scripts via the CLIP [33] similarity between short script summaries and the user-provided story themes. Finally, we prompt the LLM to retrieve and generate stylized long-term scripts based on the short script inputs. Given the planned keyframes, a low-level control policy is employed to obtain the detailed body motions in the physical simulator, producing natural, diverse, and high-quality interactions. To ensure stylized motions are adaptable to various furniture shapes within a complex indoor environment, we propose a multi-condition control policy that is attuned to scene geometries, task goal observations, and text embeddings from the CLIP model [33] for high-fidelity motion generation. Our multi-condition design not only facilitates effective scene perception but also captures fine-grained body movements, enabling a better grasp of stylized motor skills, i.e., the policy learns to perform more skills during imitation learning. Compared to previous policies [13, 26] that lack style control and UniHSI [50], which relies on accurate references, our approach supports flexible multi-condition control while mitigating mode collapse in AMP-based methods [30]. We incorporate a finite state machine (FSM) to manage multiple policies guided by specified keyframes, enabling the synthesis of physics-based animation that aligns with real-world distributions while improving scalability. To address the scarcity of motion data in the field of stylized motion generation, we collected and annotated captions and style labels from five existing motion capture datasets. Additionally, we capture a new dataset named ViconStyle to supplement the limitations in both the categories and quantity of stylized motion data.

We conduct an extensive evaluation of our method to validate its effectiveness. To provide a more comprehensive overview, we compare five SOTA kinematics-based [12, 35, 46, 47, 54] and two physics-based [26, 50] long-term HSI methods with SIMS to explain our task setting in Tab. 1. Our method, SIMS, surpasses existing approaches with a fully automatic framework that integrates style diversity, text awareness, scene awareness, and physics plausibility for realistic human-scene interactions. Unlike prior methods, it supports easy extension, ensuring scalability and adaptability. SIMS also achieves the most comprehensive skill coverage, making it a state-of-the-art solution for versatile and controllable motion synthesis.

In summary, our contributions are threefold:

1. We propose a framework for physically simulated characters to perform stylized 3D interactions using RAG-based script generation and a multi-condition control policy that encodes style from text while adapting to the environment, featuring: *(a) Stylized Control*: A script planner for coherent storytelling and a text-conditioned controller for expressive, style-consistent motion. *(b) Automatic Generation*: A planner that generates executable keyframes from theme descriptions. *(c) Scalability*: New skills and styles can be integrated by updating the script database and training a new policy.
2. We provide a comprehensive dataset of restructured motion clips with captions, emotional labels, and a short script database for stylized interactions.
3. Our method outperforms previous approaches across multiple metrics, achieving high-quality, diverse, and

physically plausible long-term motion generation.

2. Related Works

Kinematic-based Human Scene Interaction Synthesizing realistic human behavior has been a long-standing challenge. While most methods enhance the quality and diversity of humanoid movements [3, 15, 20, 40–42, 56, 60], they often overlook scene interactions. Recently, there's been growing interest in integrating human-scene interactions, crucial for applications like embodied AI and virtual reality. Many previous approaches [3, 12, 16, 35, 43, 46, 47, 54, 57, 58] rely on data-driven kinematic models [6, 9, 37, 44, 45] for static or dynamic interactions. However, these often lack physical plausibility, resulting in artifacts like penetration, floating, and sliding, and require additional post-processing, limiting real-time use.

Physics-based Human-Scene Interaction While previous physics-based animation approaches mainly focused on human motion alone [5, 14, 29–31]. InterPhys [13] presents a framework extending AMP to include character and object dynamics, using a scene-conditioned discriminator for superior performance compared to previous methods. Additionally, InterScene [26] effectively synthesizes physically plausible long-term human motions in complex 3D scenes by decomposing interactions into Interacting and Navigating processes. This method uses reusable controllers trained in simple environments to generalize across diverse scenarios. With the development of LLMs, UniHSI [50] introduces a unified framework for human-object interaction via language commands, featuring an LLM Planner and Unified Controller, which reduces training labor with LLM-generated plans. The effectiveness of this approach is evaluated using the ScenePlan dataset.

Comparison with Previous HSI Methods We compare five kinematics-based SOTA and two physics-based long-term HSI methods with SIMS to explain our task setting in Tab. 1. NSM [35] and SAMP [12] use goal positions for planning. Humanise [46], AffordMotion [47], and TeSMo [54] utilize text-based control for human motion, with the latter two leveraging textual annotations from datasets like HumanML3D [11], enabling some details in motion expression. All five kinematics-based methods rely on continuous keyframe control, requiring frequent user input updates. In contrast, InterScene [26] automates control by setting long-term keyframes for FSM to switch skills, and UniHSI [50] applies long-term keyframes of body-object contacts. Our planning uses RAG to generate long-term scripts, and enable automation and diversity. For HSI skills, we focus on 2 locomotion skills: walk and idle, 4 common human scene interaction skills: sit, lie, get up, and touch, and 1 dynamic object interaction skill: carry. Regarding control extensibility, only InterScene and our approach allow training solely for new skills without retraining the entire controller. In Supp.Mat, we demonstrate how to easily involve new interaction skills with specific styles into our framework.

3. Method

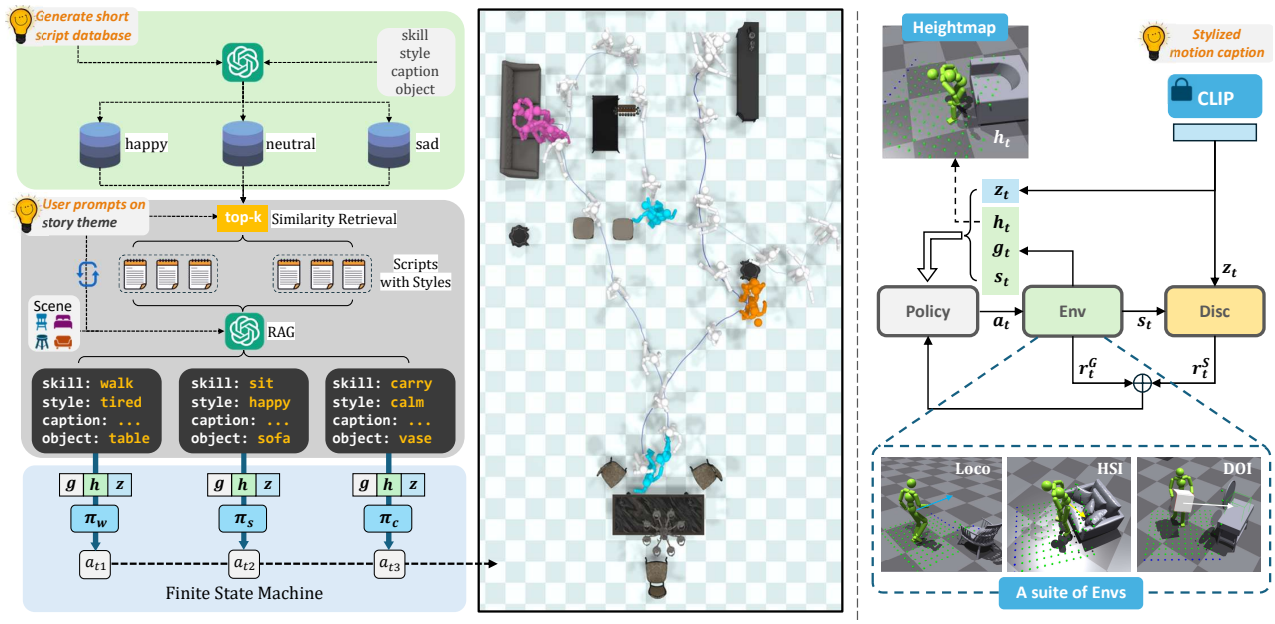
We present SIMS as a hierarchical character animation system that leverages LLMs for high-level long-term script planning, multi-condition policies for low-level character control, and a finite state machine to bridge two levels. In Sec. 3.1, we first describe the construction of short script databases. Sec. 3.2 then describes the generation of stylized long-term scripts using Retrieval-Augmented Script Generation (RASG). Finally, Sec. 3.3 explains the training of multi-condition policies and their scheduling through the finite state machine based on key frames. The supplementary material demonstrates our system's extensibility in adding new scene interaction skills.

3.1. Short Script Database Construction

A short script p consists of a sequence of key frames $\{f_0, f_1, ..., f_N\}$. Each key frame $f = (s, o, c, e)$ specifies (1) a skill s to execute, (2) a target object o to interact with, (3) captions c that describes motion attributes, and (4) the emotion or style e the motion expresses. Inspired by filmmaking, the short script uses only a few key frames to represent a short daily human-scene interaction segment. We add a concise one-sentence summary u that encapsulates the core style or emotion and interaction events of the short script. We further separate the style or emotion keyword as a distinctive label d, as a conclusion of the keyframe style labels. hus, the final formation of the short script is $p = [\{f_0, f_1, ..., f_N\}, u, d]$, serving as the foundational building block in the database. We prompt a Large Language Model (LLM) [1] to generate a wide range of short scripts by providing it with the available skills, text captions, specific styles, and available objects. The LLM is tasked not only with creating coherent and lifelike key frame sequences but also with generating matching summaries u. These short scripts are further categorized based on their distinct emotion or style labels for better modular organization. To enable retrieval, we employ CLIP [33] to extract embeddings from the summaries of the short scripts. The extracted embeddings act as keys for efficient and precise retrieval within the database.

3.2. Retrieval Augmented Script Generation

Long-term script generation with LLMs faces challenges such as redundancy, lack of diversity, and insufficient guidance in maintaining coherent narratives. Previous works, such as [50], focus on generating limited keyframes with minimal diversity, which constrains their ability to create engaging and robust long-term stories. Inspired by



(a) Inference Pipeline **(b) Environment** **(c) Control Policy Training**

Figure 2. (a) Our main pipeline. We prompt LLMs to generate new short scripts following their emotion and interaction logic. The retrieval process includes 2 stages. We first retrieve the top-k short script with semantics similarity, then ask LLM to retrieve useful samples from the short scripts and concatenate them as a fluent long-term story. In the Finite State Machine. We parse skills, captions, and scene geometry from each keyframe into task goals, language embeddings, and heightmap conditions to drive the low-level physical control policy. (c) The multi-condition physics policy. We divide common skills into 3 categories: Lococmotion, HSI, and DOI. Skills in the same category share similar task observations and reward computations.

Retrieval-Augmented Generation (RAG) [18], we propose a novel Retrieval-Augmented Script Generation (RASG) method to address these issues.

To enhance long-term script generation, the LLM retrieves and builds upon the pre-generated short scripts based on user themes in the following steps:

1) The LLM identifies M styles most relevant to the theme, narrowing down the potential scope of retrieval. 2) Semantic Similarity Retrieval: The user-provided theme sentence is extracted as a CLIP feature, which serves as the retrieval query. By computing the cosine distance between query and keys, the LLM retrieves top-k of short scripts for each style. Resulting in $M \times k$ summaries being retrieved for further processing. 3) Summary Filtering and Long Script Creation: The retrieved summaries are passed to the LLM. Then, based on the given scene layout, the LLM selects and combines suitable summaries into a cohesive narrative by logically concatenating keyframes.

To ensure executable permutations, we structure skills into tuples, such as *(sit, getup)*, *(lie, getup)*, *(idle)*, *(walk, carry)*, *(walk, reach)*, etc. Notably, the *walk* skill can serve as a transition motion between any skill tuples, enabling seamless connections across sequences. We use this rule to process the generated keyframes and add transitions for interaction skills.

3.3. Multi-Condition Controller

Overview Once a long-term script generated, our goal is to direct a simulated character to perform the key frame sequence in complex 3D scenes. To train characters to complete tasks in a lifelike and stylized manner, we adopt a goal-conditioned RL framework with a text-conditioned discriminator [30]. At each time step t, the policy $\pi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{h}_t, \mathbf{g}_t, \mathbf{z})$ receives the humanoid proprioception $\mathbf{s}_t \in \mathcal{S}$, an egocentric heightmap $\mathbf{h}_t \in \mathcal{H}$, a task-specific goal state $\mathbf{g}_t \in \mathcal{G}$, and a language embedding $\mathbf{z} \in \mathcal{Z}$. The goal $\mathbf{g}_t$ specifies high-level task objectives that the character should achieve, such as contacting with a certain furniture or moving an object to a certain coordinate. The $\mathbf{h}_t$ is the egocentric heightmap around the character, representing the surrounding geometries. The language embedding $\mathbf{z}$ specifies the style that the character should use to achieve the desired task, such as walking excitedly or sitting with legs crossed. The policy π then samples an action $\mathbf{a}_t \in \mathcal{A}$. Applying the action $\mathbf{a}_t$, the environment performs state transition and the policy receives a reward r_t. The objective is to learn a policy that maximizes the expected discounted return $J(\pi) = \mathbb{E}_{p(\tau|\pi)}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$, where T is the horizontal length and $\gamma \in [0, 1]$ defines the discount factor. In order to train the policy π to perform the task using diverse motion styles, we utilize a reward function consisting of two components: $r_t = \lambda^{\text{style}} r_t^{\text{style}} + \lambda^{\text{task}} r_t^{\text{task}}$, where r_t^{style} is a style reward modeled by the text-conditioned motion

discriminator, and r_t^{task} is a task-specific reward with coefficient λ^{task}.

Finite State Machine As illustrated in Fig 2, our framework integrates several reusable policies, serving as low-level controllers. We have trained 7 policies: the Walk policy π_w, Idle policy π_i, Sit policy π_s, Lie policy π_l, Reach policy π_r, GetUp policy π_g and Carry policy π_c.

Based on these control policies, our FSM provides users with 7 reusable skills Following [26], the FSM determines when to transition between skills. For instance, it initiates the next skill when the overlap time between the character's root and its target position exceeds a specific threshold. This simple rule-based FSM allows users to achieve desired long-term human motions in complex 3D scenes. Compared to the recent work InterScene [26], our FSM contains egocentric heightmaps by frame and text embedding by skill, which could ensure scene understanding and semantic control.

Language Condition To control policy language constraints, we build an embedding space where motion representations are aligned with natural language descriptions. Given a motion clip $\hat{\mathbf{m}} = (\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_n)$, the motion encoder $\mathbf{z} = \text{Enc}_m(\hat{\mathbf{m}})$ maps the motion to a unit sphere embedding $\|\mathbf{z}\| = 1$, while corresponding text captions are processed through a pre-trained CLIP [33] encoder Enc_l and use fully connected layers to match the latent dimensionality. The training combines reconstruction and alignment losses to ensure that motion and text embeddings effectively correspond to each other. For further details on the network architecture and training losses, please refer to the Supp. Mat.

Scene Condition To enhance the humanoid's navigation and interaction capabilities, it is crucial to maintain environmental awareness to prevent collisions. We draw inspiration from methods such as [35, 43, 48, 50], which utilize environmental sampling for humanoid observations. A square, ego-centric heightmap is generated to capture the elevation of surrounding objects. See in Fig. 2. Consistent with UniHSI [50], we pre-generate pointclouds for each scene. However, creating detailed pointclouds while preserving surface intricacies is computationally intensive. To enhance the humanoid's understanding of complex surfaces for sitting or lying, we pre-generate scene pointclouds by voxelizing the objects within the bounding box range. The egocentric heightmap is updated by calculating the nearest object's pointclouds only when the object is sufficiently close to the humanoid's root position. The heightmap is a 12×12 grid with an adjacent distance of 0.15 meters. We flatten the heightmap grid to a vector and concatenate it into the observation.

Universal Goal Condition We consider 7 distinct scene interaction skills. To reduce the development overhead of diverse task-specific configurations, we implement all interaction tasks based on 3 task templates: Loco (Walk and Idle), HSI (Sit Lie, Reach and GetUp) and DOI (Carry). The implementation details are as follows:

- **Loco tasks** require the humanoid to position its pelvis at a target 2D location $\mathbf{g} \in \mathbb{R}^2$. For Walk, the location is set $\geq 1m$ from the humanoid's initial position, whereas the location of Idle is identical to the humanoid's current position, encouraging pacing in place.
- **HSI tasks** require a specific body joint to contact with the surface of a target object. We constrain the pelvis joint in Sit, Lie, and GetUp, and use either the left or right hand for Reach. The target location $\mathbf{g} \in \mathbb{R}^3$ is determined by the nearest 3D point on the object's interactable surface.
- **DOI tasks** no longer constrain body joints, but encourage the character to move the dynamic object's root to a target 3D location. We use the bounding box coordinates of the object $\mathbf{g}^{bbox} \in \mathbb{R}^{3\times8}$ and the target location $\mathbf{g}^{tar} \in \mathbb{R}^3$ as the goal condition $\mathbf{g} = \{\mathbf{g}^{bbox}, \mathbf{g}^{tar}\}$.

Using sparse goal conditions can effectively train policies to perform scene interaction tasks [8, 13, 26]. However, we cannot control motion styles via these conditions. Tracking-based methods [21, 39, 50, 53] enable fine-grained control of each frame but require accurate stylized reference motions as dense input conditions. We employ a conditional discriminator [5, 38] to inject text-based style control into policies. Unlike motion [38] or one-hot [5] conditions, language is a more intuitive interface for LLMs and users.

Policy Training We train 7 task-specific policies: (1) Walk, (2) Idle, (3) Sit, (4) Lie, (5) Reach, (6) GetUp, and (7) Carry. We provide Walk, Idle, Sit, Lie, Carry policies with text conditions since these behaviors contain diverse interaction styles that represent vivid emotions. For Reach and GetUp, we do not use text conditions.

- **Initialization.** Following UniHSI [50], we create the environment by randomly sampling objects from 3DFront [7]. For HSI skills, we initialize characters using reference state initialization [28] and default pose initialization with a random global rotation and location[26, 50] nearby the object. For locomotion skills, we randomly sampled on the whole ground plane while calculating the collision with the objects. For DOI skills, we randomly sample target position on the whole ground plane, and initialize objects in the humanoid's hands from reference object motion. Notebly, we add Walk motion data to the initiate reference state data during the training of all the skills because we use Walk as the transition between different interactions.
- **Rewards.** See the detailed reward function in Supp.Mat.
- **Reset and early termination conditions.** Following [30], we use a fixed episode length and fall detection as early termination triggers. We also use early termination when the task is accomplished for a certain time [26] or

Datasets	Loco		HSI				DOI
	Walk	Idle	Sit	Lie	Getup	Reach	Carry
SAMP [12]	20.6	-	35.2	14.8	11.2	-	-
COUCH [58]	-	-	36.4	-	23.4	-	-
Circles [2]	-	-	-	-	-	3.6	-
100Style [24]	203.1	-	-	-	-	-	-
AMASS [22]	8.2	-	-	-	-	-	3.4
ViconStyle	-	12.0	-	21.9	11.7	-	26.0

Table 2. Mixture of collected stylized motion datasets.

the contact forces are extremely large [50].

4. Experiments

4.1. Dataset

We show our collected mixture of 6 motion dataset in Tab. 2. We show the skill for training and the motion duration in minutes. The number with black bounding-box like 20.6, means the 20.6 minutes of motion in this dataset do not have style diversity, only counted as *neutral*. ViconStyle is our captured dataset, which supplements for the quantity and the category of stylized motions. See details in Supp.Mat. We annotate all the motion clips with captions and style labels. For each caption, we provide 5 synonymous sentences with the help of LLM [1]. Besides neutral, we categorize the emotion or style of the remaining motions into 8 categories: *happy, angry, hurried, tired, sad, stressed, drunk, and relaxed.* We left-right-flip all the motions so we get double the amount, and the captions are flipped concerning body joint symmetry as well.

For 3D objects, we use the furniture and scene layouts from the 3DFront [7] dataset for training. Since 3DFront does not provide segmentation information, we voxelize the object meshes and segment the point clouds based on normal vectors to get the affordance surface.

4.2. Motion Metrics

To evaluate motion diversity, we use two metrics from the previous papers: Fréchet Inception Distance (FID) [5, 41] and Average Pairwise Distance (APD) [5, 43]. FID measures the similarity between the distributions of generated and real data in a feature space, reflecting the realism and quality of the generated motions. Lower FID values indicate closer alignment with real data. APD, on the other hand, quantifies the diversity within the generated motions by calculating the average pairwise distance between samples. Higher APD values indicate greater diversity in the generated motions. We calculate FID and APD on joint rotations and positions.

We follow [12, 50] that uses *Success Rate* and *Contact Error* as the main metrics to measure the quality of interactions quantitatively. Success Rate records the percentage of trials that humanoids successfully complete the contact within a certain threshold. We follow [13, 26, 50] to set the threshold of Sit as 20cm, Reach as 20cm, Lie as 30cm, Carry as 20cm.

To evaluate the generation quality of long-term scripts, we also involve user study and SBERT [34] Model, please see the metrics in the corresponding part.

Methods	Success Rate (%) ↑				Contact Error ↓			
	Sit	Lie	Reach	Carry	Sit	Lie	Reach	Carry
InterPhys [13]	93.7	80.0	-	94.3	0.09	0.30	-	**0.08**
InterScene [26]	97.8	-	-	-	0.04	-	-	-
UniHSI[50]	94.3	81.5	**97.5**	-	0.032	0.061	**0.016**	-
SIMS	98.1	87.6	95.2	92.9	**0.028**	0.049	0.026	0.099
SIMS (+data)	**98.4**	**89.6**	-	**96.4**	0.033	**0.048**	-	0.085

Table 3. Comparision on Baseline Models. For fair comparison, our Sit, Lie, and Reach policies are only trained on SAMP [12] here. While our Carry policy is trained on the small amount of carry motions from AMASS [22]. (+data) here represents our results trained on available motions from the mixture of 6 datasets.

4.3. Comparison with SOTA methods

4.3.1. Physical Performance for Different Skills

Our method achieves better or comparable results across various metrics in Tab. 3. Unlike previous physics-based methods [13, 26, 50] which only care about contact but not styles, our result is achieved on 4096 random text conditions sampled from the datasets. The previous methods could be viewed as just a specific situation of our model. Under this background, we can see from Tab. 3 that our results are only slightly lower than the best methods in Reach and Carry skill. Since Interphys [13] have not released their code and carry motion data, we only train on the small amount of carry motion in AMASS [22] for Tab. 3.

4.3.2. Motion Diversity for Different Skills

We compare motion diversity in the Sit and Lie skills with UniHSI [50] and our re-implemented Interphys [13]. All experiments are conducted on a single RTX 4090 GPU, running 1024 sequences and aggregating the results over 10 trials. For each sequence, the text condition is randomly sampled from the dataset. To test UniHSI [50], we randomly sample contact pairs from the provided chain of contacts from the generated ScenePlan dataset. We measure the FID between the generated motions and that of reference motions from SAMP [12]. The APD measures the diversity among the generated motion sequences. As shown in Tab. 4, our results significantly outperform UniHSI in both FID and APD metrics. Our method achieves lower FID, indicating motions produced from ours are closer to the distribution of reference motions. Notably, the APD results highlight that the motions generated by UniHSI are nearly identical, demonstrating a lack of diversity. Our method also surpass the re-implemented InterPhys [13].

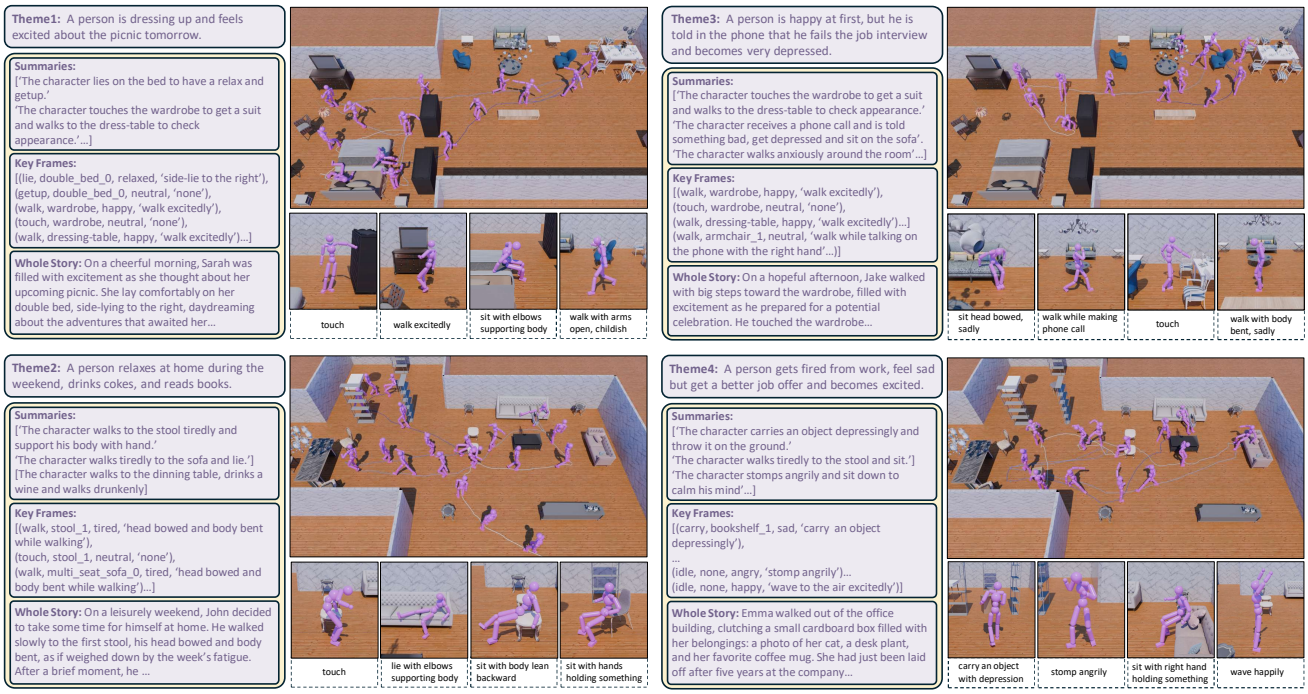


Figure 3. Long-term scripts with detailed keyframes and vivid final stories in two complex 3D scenes generated by our complete system. Upper: character in the bedroom and living room. Lower: character in the living room, dining room, and study room. We briefly demonstrate the retrieved summaries, key frames and part of the final long stories.

Method	FID↓			APD↑		
	Sit	Lie	Carry	Sit	Lie	Carry
InterPhys* [13]	-	-	81.0	-	-	12.41±0.19
UniHSI [50]	153.84	211.22	-	1.14±0.01	1.35±0.02	-
SIMS	**125.66**	**171.24**	**65.14**	**16.55±0.54**	**16.40±0.94**	**14.36±0.12**

Table 4. Motion diversity results. InterPhys [13] is not released, so we report our re-implemented version here. For fair comparison, our Sit, Lie, and Reach policies are only trained on SAMP [12] here. While the Carry policy and the re-implemented InterPhys are both trained on the carry motions from ViconStyle.

Metrics		UniHSI	SIMS
Motion	Physical Realism ↑	2.6	**3.4**
	Motion Diversity ↑	2.9	**3.6**
Script	Plot Engagement ↑	2.4	**3.0**
	Emotional Resonace ↑	3.0	**3.8**

Table 5. User Study on SOTA long-term HSI methods. SIMS outperforms the SOTA method UniHSI by a significant margin.

4.3.3. User Study on Different Long-Term HSI Methods

To further evaluate the control capabilities of the long-term scripts, we conducted a user study on the rendered videos generated from different methods. We use the same category of interactions to drive the characters in the scenes. 30 participants were asked to rate the physical realism, motion diversity, split engagement and emotion resonace of the videos produced by each method on a scale from 1 (poor) to 5 (excellent). In Tab. 5, the results indicate that our approach significantly outperformed UniHSI, demonstrating its effectiveness in both body motion superiority and script superiority in the generated animations.

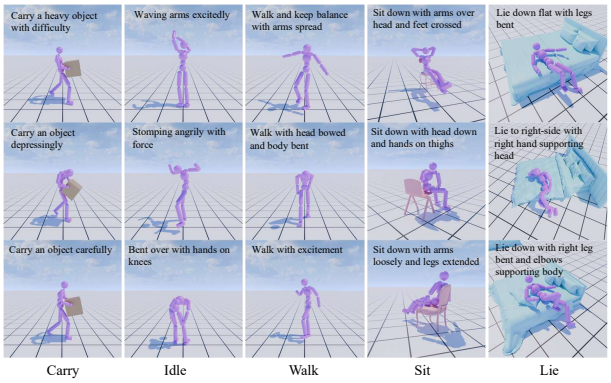


Figure 4. Qualitative results for skills with different text conditions.

4.3.4. User Study on Different HSI Skills

Tab. 6 presents a user study where 21 participants rated stylistic diversity and physical realism of skills learned by existing methods from 1 (poor) to 5 (good). Results show that SIMS outperforms most methods in both aspects. Although InterPhys achieves slightly higher physical realism on the *Sit* skill, it lags behind SIMS in diversity.

Method	Stylistic Diversity↑				Physical Realism↑			
	Walk	Sit	Lie	Carry	Walk	Sit	Lie	Carry
InterPhys [12]	3.9	3.9	3.8	3.4	4.2	**4.1**	3.8	3.8
InterScene [25]	3.4	3.4	3.3	-	3.5	3.8	3.9	-
UniHSI [46]	3.5	3.3	3.3	-	3.1	3.6	2.7	-
SIMS	**4.3**	**4.2**	**4.3**	**4.2**	**4.2**	4.0	**3.9**	**4.3**

Table 6. User Study on SOTA HSI Skills.

4.4. Ablation Study

We show more ablation studies in Supp.Mat.

4.4.1. Generalizability on Different Text Conditions

The style of each skill is controlled by the input text condition. To evaluate text generalization, we prompt GPT-4 [1] to generate 3 unseen datasets with varying levels of semantic similarity to the *original* training dataset: *high*, only keywords are replaced with synonyms; *medium*, more words are replaced; *low*, the entire sentence is rephrased. In Tab. 7, as the similarity decreases, our policy still performs tasks with comparable FID and success rates. This indicates that the style control does not overfit the training texts and could generalize to new captions (See Fig. 5). This ability is primarily attributed to the continuous latent space of our implemented MotionCLIP.

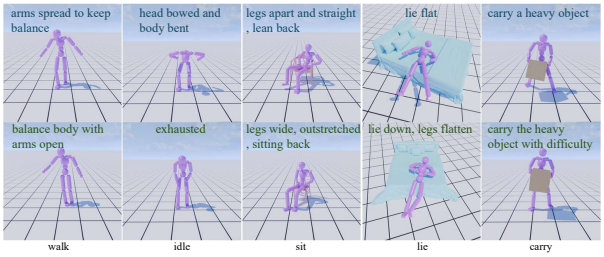


Figure 5. Demo on seen text conditions(1st row) and unseen text conditions(2nd row).

Similarity	FID↓				Success Rate (%)↑			
	Walk	Sit	Lie	Carry	Walk	Sit	Lie	Carry
low	172.75	132.69	182.84	68.78	95.6	91.8	86.5	92.3
medium	164.68	127.13	198.97	70.84	97.8	94.1	87.5	92.5
high	**163.38**	**122.48**	188.99	70.16	**98.4**	95.7	89.1	92.8
origin	170.03	125.66	**171.24**	**65.14**	98.3	**98.4**	**89.6**	**96.4**

Table 7. Diversity on Different Text Conditions.

4.4.2. Scale Up on New Motion Datasets

To prove the reliable of the proposed datasets, and the generality of our text-conditioned policy, we report the Success Rate and APD for Walk, Carry, Sit, and Lie skills in Tab. 8, Tab. 9, and Tab. 10. From the tables, we could find that with more data, Walk achieves a higher success rate mainly because AMASS provides stable neutral walking and running motions. The APD changes little because 100Style also contains neutral walking styles. For carry skill, since ViconStyle is the first dataset containing stylized carrying motion, both metrics increase by a large margin. For HSI skills, sit and lie both become slightly better with the introduction of COUCH and ViconStyle dataset. Couch provides more stylized sitting motions and ViconStyle provides more stylized lying motions.

Datasets	Success Rate(%)↑ Walk	APD↑ Walk
100S	92.6	14.83±0.35
A+100S	95.1	14.88±0.29

Table 8. Dataset ablation on Walk Skill. 100S: 100Style , A: AMASS.

Datasets	Success Rate(%)↑ carry	APD↑ carry
A	92.9	14.36±0.12
A+VS	96.4	14.92±0.23

Table 9. Dataset ablation on Carry Skill. A: AMASS. VS: ViconStyle.

Datasets	Success Rate(%)↑		Contact error↓		APD↑	
	Sit	Lie	Sit	Lie	Sit	Lie
S	95.5	86.9	0.040	0.055	16.43±0.90	16.40±0.94
S+C	96.9	-	0.014	-	16.52±0.47	-
S+C+VS	-	89.7	-	0.030	-	16.84±1.28

Table 10. Dataset ablation on HSI Skills. S: SAMP [12], C: Couch [58], VS: ViconStyle

4.5. Qualitative Results

We show 4 generated long narratives executed by our policies in two large indoor scenes. The details can be viewed in Fig. 3. In Fig. 4, we also showed some qualitative samples for 5 skills: Carry, Idle, Walk, Sit, and Lie. We suggest the readers to refer to the demonstration videos for a better knowledge of our ability to generate long-term stylized motions.

5. Conclusion

In this paper, we analyze and compare the current advancements in long-term human-scene interaction tasks, highlighting the lack of generating animations that are both physically plausible and stylistically expressive. To address this, we propose a novel framework for synthesizing long-term human-scene interactions by leveraging Retrieval-Augmented Generation as high-level planners and a multi-condition control policy as the low-level controller. By incorporating both stylized script generation and a stylized control policy, our approach facilitates the creation of diverse, expressive, and physically coherent long-term animations. Furthermore, the processed datasets open up new possibilities and directions for future research in this field.

6. Furture Work

In the future, it will be essential to collect more human motion data that captures realistic emotions and diverse styles. Additionally, exploring humanoid models with articulated fingers presents a promising avenue for research. Introducing multi-agent in HSI could also broaden the possibilities for physical animations.

7. Acknowledge

This work is partly supported by the Research Grant Council of Hong Kong (Ref: 17210222) and the Innovation and Technology Commission of Hong Kong under the ITSP-Platform grants (Ref: ITS/335/23FP). Part of the research was conducted in the JC STEM Lab of Robotics for Soft Materials, funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 6, 8, 1

[2] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *CVPR*, 2023. 6

[3] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024. 3

[4] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[5] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH 2023*, 2023. 3, 5, 6

[6] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. In *ICCV*, 2023. 3

[7] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021. 5, 6, 1

[8] Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. Coohoi: Learning cooperative human-object interaction with manipulated object dynamics. *Advances in Neural Information Processing Systems*, 37, 2024. 5

[9] Yongtao Ge, Wenjia Wang, Yongfan Chen, Hao Chen, and Chunhua Shen. 3d human reconstruction in the wild with synthetic data using generative models. *arXiv preprint arXiv:2403.11111*, 2024. 3

[10] Nima Ghorbani and Michael J. Black. Soma: Solving optical marker-based mocap automatically. In *ICCV*, 2021. 3

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 3

[12] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 1, 2, 3, 6, 7, 8

[13] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH 2023*, 2023. 2, 3, 5, 6, 7

[14] Yiming Huang, Zhiyang Dou, and Lingjie Liu. Modskill: Physical character skill modularization. *arXiv preprint arXiv:2502.14140*, 2025. 3

[15] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeuraIPS*, 36, 2023. 3

[16] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024. 3

[17] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeuraIPS*, 33, 2020. 2, 4

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *TOG*, 34(6), 2015. 3

[20] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. *arXiv preprint arXiv:2412.14559*, 2024. 3

[21] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023. 5

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 6

[23] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 2

[24] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5 (1), 2022. 6

[25] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019. 1

[26] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *3DV*, 2024. 2, 3, 5, 6

[27] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *CVPR*, 2025. 2

[28] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *TOG*, 37(4), 2018. 5

[29] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *TOG*, 37(4), 2018. 3

[30] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *TOG*, 2021. 2, 4, 5, 3

[31] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *TOG*, 41 (4), 2022. 3

[32] Huaijin Pi, Zhi Cen, Zhiyang Dou, and Taku Komura. Coda: Coordinated diffusion noise optimization for whole-body manipulation of articulated objects. *arXiv preprint arXiv:2505.21437*, 2025. 1

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 5

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 6, 1

[35] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *TOG*, 38(6), 2019. 1, 2, 3, 5

[36] Jingkai Sun, Qiang Zhang, Yiqun Duan, Xiaoyang Jiang, Chong Cheng, and Renjing Xu. Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning. In *ICRA*, 2024. 2

[37] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *CVPR*, 2024. 3

[38] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *SIGGRAPH 2023*, 2023. 5

[39] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *TOG*, 43 (6), 2024. 5

[40] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 2022. 3, 2

[41] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 6

[42] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 3

[43] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, 2022. 1, 3, 5, 6

[44] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Cheng Lin, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions. *arXiv preprint arXiv:2312.05295*, 2023. 3

[45] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *ICCV*, 2023. 3

[46] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *NeuraIPS*, 35, 2022. 1, 2, 3

[47] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *CVPR*, 2024. 2, 3

[48] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *TOG*, 41(4), 2022. 5

[49] Yifan Wu, Zhiyang Dou, Yuko Ishiwaka, Shun Ogawa, Yuke Lou, Wenping Wang, Lingjie Liu, and Taku Komura. Cbil: collective behavior imitation learning for fish from real videos. *ACM Transactions on Graphics (TOG)*, 43(6):1–17, 2024. 1

[50] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024. 2, 3, 5, 6, 7, 1

[51] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. Hierarchical planning and control for box loco-manipulation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6 (3), 2023. 2

[52] Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, et al. Mospa: Human motion generation driven by spatial audio. *arXiv preprint arXiv:2507.11949*, 2025. 1

[53] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. *arXiv preprint arXiv:2502.20390*, 2025. 5

[54] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. *ECCV*, 2024. 2, 3

[55] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 2

[56] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[57] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. In *3DV*, 2024. 3

[58] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *ECCV*, 2022. 1, 3, 6, 8

[59] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *ICCV*, 2023. 1

[60] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *ECCV*, 2024. 1, 3