MMG-VL: A Vision-Language Driven Approach for Multi-Person Motion Generation

Songyuan Yang*, Wanrong Huang, Yinuo Liu, Zhang Ke-Di, Xihuai He, Shaowu Yang*, Huibin Tan†

College of Computer Science and Technology, National University of Defense Technology
{yangsongyuan, tanhb_}@nudt.edu.cn

Abstract

Generating realistic and coordinated 3D human motion for multiple individuals within complex environments remains a significant challenge. Existing text-to-motion methods are often "blind" to the physical scene, leading to implausible motions, while scene-conditioned (HSI) approaches demand cumbersome full 3D data and largely neglect multi-person dynamics. To address these limitations, we introduce the **VL2Motion** paradigm and its embodiment, **MMG-VL**, a hierarchical framework that generates coordinated multi-person motions from the most accessible inputs: a single 2D image and natural language. MMG-VL first employs a **Scene-Aware Intent Planner (SAIP)** to interpret the visual context and decompose the user's command into a set of spatially-grounded, multi-person action blueprints. Subsequently, a **Coordinated Motion Synthesizer (CMS)** translates these blueprints into high-fidelity 3D motion sequences. The synergy between these stages is driven by two novel loss functions: a **Spatial-Semantic Grounding Loss** ($\mathcal{L}_{SSG}$) to ensure the planner's output is grounded in visual reality, and a **Coordinated Environmental Realism Loss** ($\mathcal{L}_{CER}$) that enforces physical constraints and coherent group dynamics during synthesis. To facilitate this research, we introduce **HumanVL**, the first large-scale dataset featuring multi-person activities in multi-room scenes, providing aligned images, text, blueprints, 3D motions, and scene geometry. Extensive experiments demonstrate that MMG-VL significantly outperforms existing methods in generating spatially coherent, physically realistic, and coordinated multi-person motions, paving the way for more scalable and intuitive creation of dynamic virtual worlds.

Introduction

Generating realistic 3D human motion from language has achieved significant breakthroughs [Tevet et al. 2023, Guo et al. 2023, Jiang et al. 2024a, Lan et al. 2018, 2023], unlocking potential in virtual reality, robotics, and digital content creation. However, a prevailing limitation of these text-driven methods is their inherent "blindness" to the physical world. Motions are often synthesized in a conceptual vacuum, disembodied from the spatial and physical constraints of a 3D environment. This disconnect leads to physically implausible behaviors, such as floating through furniture or ignoring walls, critically undermining their utility in embodied AI applications that demand robust scene interaction.

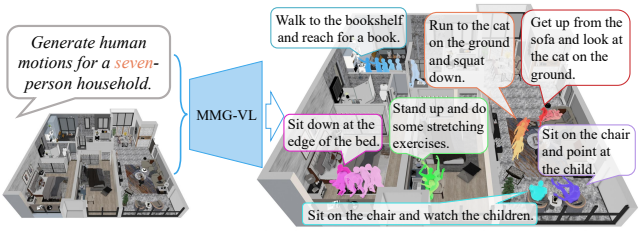

Figure 1: VL2Motion paradigm. Given an environmental image and a natural language description, MMG-VL can generate coordinated multi-person motions that aligns naturally with the multi-room environment.

Concurrently, an alternative paradigm of Human-Scene Interaction (HSI) [Jiang et al. 2024b,c, Li et al. 2025] generates high-fidelity, physically-grounded motions by conditioning on explicit 3D scene geometry. While effective, this approach is often cumbersome, relying on complete and costly 3D data (e.g., CAD models, dense point clouds) and non-intuitive control mechanisms like specifying 3D waypoints. This high barrier to entry restricts their use to scenarios where full 3D information is readily available and users possess specialized skills. More importantly, both paradigms largely overlook the complexity of multi-person scenarios, where coordinated, spatially-aware group behavior is not merely a collection of individual actions but a dynamically interconnected system.

To bridge this gap, we propose a new paradigm, VL2Motion, which synthesizes multi-person motion by leveraging the most accessible and intuitive data modalities: a 2D image of the environment and natural language instructions. This approach presents a formidable technical challenge: **How to infer rich 3D spatial semantics, ground complex motion trajectories, and orchestrate plausible multi-person coordination from sparse, multimodal inputs?** Our work tackles this challenge head-on by introducing MMG-VL, an end-to-end framework designed to generate spatially-aware and coordinated 3D human motions within complex, multi-room household environments.

*These authors contributed equally.
†Corresponding author.

MMG-VL operates on a hierarchical principle. First, a novel Scene-Aware Intent Planner (SAIP) module interprets the visual context from the 2D image and the user's textual instructions, which are enriched with key 3D coordinates. It decomposes the high-level goal into a set of spatially-grounded, coordinated action blueprints. Second, a Coordinated Motion Synthesizer (CMS) module translates these blueprints into high-fidelity, full-body 3D motion sequences. The efficacy of these modules is enabled by two key technical innovations. We introduce a Spatial-Semantic Grounding Loss ($\mathcal{L}_{SSG}$) that explicitly trains the SAIP to align textual semantics with visual and spatial information. Furthermore, we design a Coordinated Environmental Realism Loss ($\mathcal{L}_{CER}$) for the CMS, which jointly optimizes for individual motion realism, enforces physical non-penetration with the scene geometry, and models inter-agent spatial relationships to ensure coherent group dynamics.

To catalyze research within this new paradigm, we introduce HumanVL, a large-scale dataset featuring diverse multi-person activities in multi-room scenes. Each sample provides a complete quintuple: a scene image, a natural language instruction, the corresponding high-level action blueprint, the resulting 3D motion data, and the underlying 3D scene geometry. This resource is crucial for training and evaluating models on the nuanced task of vision-language-driven multi-person motion generation.

Our contributions are summarized as follows: 1) We propose the VL2Motion paradigm and its embodiment, MMG-VL, a framework that generates coordinated multi-person motions from only a 2D image and text. 2) We introduce a hierarchical architecture powered by two novel loss functions, $\mathcal{L}_{SSG}$ and $\mathcal{L}_{CER}$, which enable robust scene understanding and physically-plausible multi-agent synthesis. 3) We contribute HumanVL, the first large-scale dataset specifically designed for this task, providing a comprehensive benchmark for future research. 4) Finally, extensive experiments demonstrate that MMG-VL significantly outperforms existing approaches in generating spatially coherent, physically realistic, and coordinated multi-person motions, paving the way for more interactive and intelligent embodied systems.

Related Work

Paradigms in Motion Generation. Research in 3D human motion generation has largely progressed along two distinct axes. The dominant paradigm, Text-to-Motion synthesis [Tevet et al. 2023, Guo et al. 2023, Jiang et al. 2024a], has achieved impressive semantic control from linguistic inputs [Ma, Bai, and Zhou 2022, Zhang et al. 2023, Sun et al. 2024]. However, these methods typically operate in a spatial vacuum, their "blindness" to the physical environment leading to unrealistic motions that ignore scene constraints. In response, the Human-Scene Interaction (HSI) paradigm [Jiang et al. 2024b,c] generates physically plausible motions by directly conditioning on 3D scene geometry. While effective, this reliance on complete 3D models is cumbersome and limits applicability in scenarios where only 2D images are available. Crucially, both paradigms have predominantly focused on single-agent generation. Recent explorations into multi-person synthesis [Xu et al. 2023, Liang et al. 2024, Fan et al. 2024] often overlook the intricate, coordinated interplay between agents and their shared environment. Our work directly addresses this tripartite gap: bridging text and scene, moving from 3D to 2D visual inputs, and scaling from single- to multi-person coordinated generation.

Vision-Language Guided Synthesis. Our framework is built upon the synergy of two powerful technologies: Vision-Language Models (VLMs) and Diffusion Models. VLMs [Liu et al. 2023, Zhang et al. 2024b, OpenAI 2023, Wang et al. 2022, 2023a,b] have demonstrated profound capabilities in multimodal reasoning, making them ideal for scene perception and intent planning. In parallel, Denoising Diffusion Models [Ho, Jain, and Abbeel 2020, Rombach et al. 2021] have become the state of the art for high-fidelity generative tasks, including human motion [Tevet et al. 2023, Liang et al. 2024, Sun et al. 2024]. While VLMs excel at perception and DDMs at synthesis, each has inherent limitations for our task when used alone. An emerging trend in other domains, such as text-to-image generation [Li et al. 2024, Richardson et al. 2024] and embodied AI [Chen et al. 2024, Zeng et al. 2024], is to integrate their complementary strengths. MMG-VL advances this principle into the complex domain of 3D human animation. We propose a novel hierarchical framework that does not simply cascade these models but deeply integrates them through specialized loss functions ($\mathcal{L}_{SSG}$, $\mathcal{L}_{CER}$), enabling a seamless flow from visual perception and coordinated planning to realistic multi-person motion synthesis.

Methodology

The VL2Motion Paradigm

The VL2Motion paradigm addresses tasks of generating a set of P concurrent motion sequences $\mathcal{M} = \{\mathbf{M}_1, \ldots, \mathbf{M}_P\}$. Each sequence $\mathbf{M}_p \in R^{L \times D}$ represents the motion of person p over L frames, where $D = 22$ is the dimension of the per-frame pose representation [Tevet et al. 2023]. The generation is conditioned on a tuple $(\mathcal{I}, \mathcal{T}, \mathcal{G})$, where $\mathcal{I}$ is a 2D image of the environment, $\mathcal{T}$ is a high-level textual instruction (e.g., "Generate motions for a three-person household."), and $\mathcal{G}$ is the corresponding 3D scene geometry, used for grounding and training. The core objective is to learn a mapping $f : (\mathcal{I}, \mathcal{T}) \rightarrow \mathcal{M}$ such that the resulting motions are not only individually realistic but also collectively coherent and physically compliant with the scene geometry $\mathcal{G}$ implied by the image $\mathcal{I}$. This paradigm shifts the burden of explicit 3D specification from users to the model's learned visual and spatial reasoning capabilities.

MMG-VL: A Hierarchical Generation Framework

We introduce MMG-VL (Multi-person Motion Generation via Vision-Language), a novel framework designed to operationalize the VL2Motion paradigm. As illustrated in Figure 2, MMG-VL adopts a hierarchical approach, decomposing the generation task into two synergistic stages: intent planning and motion synthesis. The framework is trained in a staged manner to optimize each component's specialized function before an optional end-to-end fine-tuning.

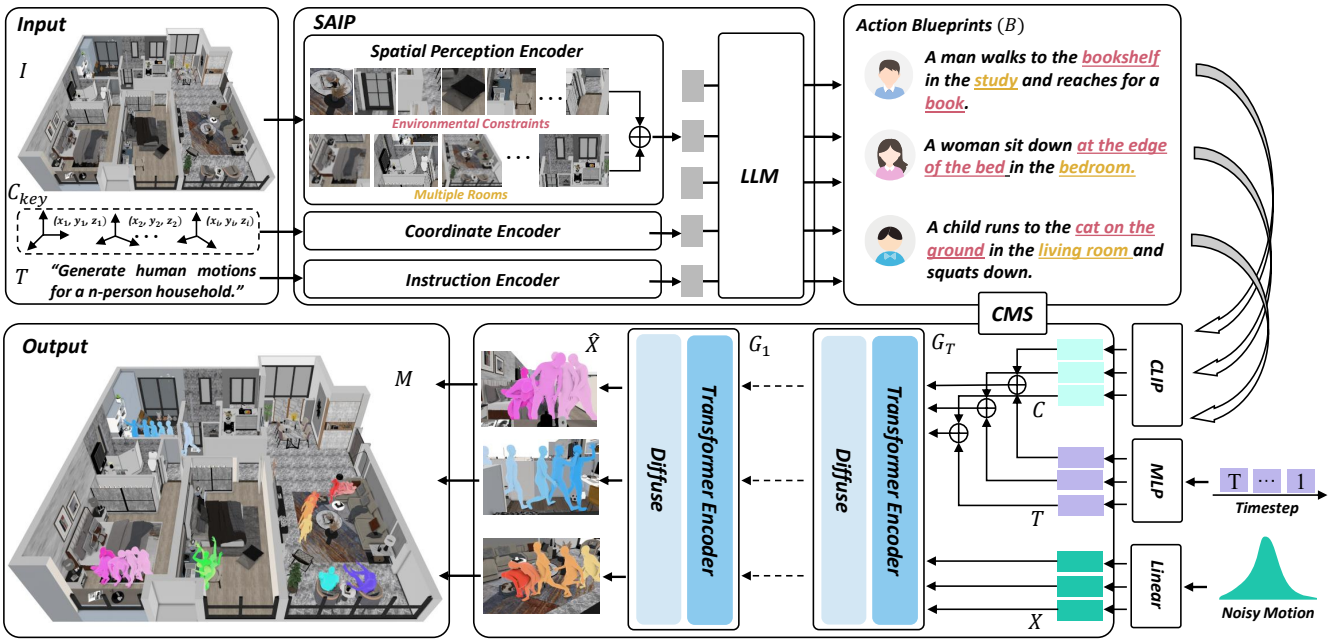

Figure 2: Our method follows a hierarchical approach. Given a 2D image ($\mathcal{I}$) and a high-level textual instruction ($\mathcal{T}$), the Scene-Aware Intent Planner (SAIP) first leverages a Vision-Language Model to analyze the scene and generate a set of spatially-grounded, multi-person Action Blueprints ($\mathcal{B}$). These blueprints serve as detailed, structured guidance for the subsequent stage. The Coordinated Motion Synthesizer (CMS), a diffusion-based model, then takes these blueprints and synthesizes the final, coordinated 3D motion sequences ($\mathcal{M}$) for all individuals. The entire process is optimized end-to-end, guided by our proposed $\mathcal{L}_{SSG}$ and $\mathcal{L}_{CER}$ loss functions, ensuring the motions are realistic, environmentally compliant, and socially coherent.

Stage 1: Scene-Aware Intent Planning and Grounding The first stage of our framework is the Scene-Aware Intent Planner (SAIP), which tackles the critical task of translating a user's high-level, often abstract, instruction into a concrete, executable plan. Its core function is to perceive the 3D environment through a 2D image and decompose the user's goal into a set of structured, spatially-aware instructions termed Action Blueprints.

Action Blueprints ($\mathcal{B}$) are detailed, per-person textual directives that form the bridge between high-level intent and low-level motion synthesis. For a scenario with P individuals, the SAIP generates a set of blueprints $\mathcal{B} = \{B_1, \ldots, B_P\}$. Each blueprint B_p specifies not just the action (e.g., "walks," "sits"), but also the crucial spatial context, including target objects, destination coordinates, and relational information with respect to other agents (e.g., "walks to the *bookshelf* at *(x,y,z)*", "sits on the *chair* facing person 2"). This structured representation provides unambiguous, fine-grained guidance for the subsequent motion generation stage.

Model and Input Formulation. The SAIP is instantiated using a powerful Vision-Language Model (VLM), adept at joint reasoning over visual and textual data. To equip the VLM with precise spatial knowledge, we devise a multi-modal prompt that synergizes three information sources: 1) The scene image $\mathcal{I}$, which provides rich visual context about the environment's layout, objects, and affordances. 2) The high-level instruction $\mathcal{T}$, which defines the overall goal (e.g., "Three people are preparing dinner"). 3) A set of key 3D coordinates $\mathcal{C}_{key} = \{(o_k, \mathbf{c}_k)\}_{k=1}^{K}$, where o_k is the name of a key object (e.g., "sofa") and $\mathbf{c}_k \in R^3$ is its 3D position in the scene's world frame. These coordinates are serialized into a string format and concatenated with the instruction $\mathcal{T}$ to form a unified textual prompt. This formulation allows the VLM to directly associate semantic object labels with their precise geometric locations. We report the prompt in the supplementary material.

The SAIP processes the image $\mathcal{I}$ and the combined textual prompt to autoregressively generate the set of action blueprints $\mathcal{B}$. This process is formally defined as learning a mapping that maximizes the conditional probability $P(\mathcal{B}|\mathcal{I}, \mathcal{T}, \mathcal{C}_{key}; \theta_{\text{SAIP}})$, where θ_{SAIP} represents the model's parameters.

Training with Spatial-Semantic Grounding. The SAIP is trained on our HumanVL dataset, which provides ground-truth tuples of $(\mathcal{I}, (\mathcal{T}, \mathcal{C}_{key}), \mathcal{B}_{gt})$. The training objective is a composite loss function designed to ensure both textual fluency and robust vision-language-geometry grounding:

$$\mathcal{L}_{\text{SAIP}} = \mathcal{L}_{\text{AR}} + \lambda_{ssg}\mathcal{L}_{SSG}. \quad (1)$$

The primary component, $\mathcal{L}_{\text{AR}}$, is a standard autoregressive cross-entropy loss that maximizes the likelihood of generating the ground-truth blueprint tokens $\mathcal{B}_{gt}$. To explicitly tackle the challenge of aligning 3D spatial information with 2D visual cues, we introduce the novel Spatial-Semantic Grounding Loss ($\mathcal{L}_{SSG}$). This loss enforces a consistent rep-

resentation between an object's visual appearance, its semantic label, and its spatial coordinates. Specifically, for each key object o_k mentioned in the input, we extract its visual feature embedding $\mathbf{v}_k$ from the VLM's vision encoder and its corresponding textual embedding $\mathbf{t}_k$ from the language encoder. $\mathcal{L}_{SSG}$ is a contrastive loss that pulls the representations of corresponding visual-textual pairs together while pushing apart non-matching pairs within a batch:

$$\mathcal{L}_{SSG} = -\frac{1}{K}\sum_{k=1}^{K}\log\frac{\exp(\text{sim}(\mathbf{v}_k,\mathbf{t}_k)/\tau)}{\sum_{j=1}^{K}\exp(\text{sim}(\mathbf{v}_k,\mathbf{t}_j)/\tau)}. \quad (2)$$

Here, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a learnable temperature parameter. By optimizing this loss, the SAIP learns to correctly associate the name "bookshelf" and its coordinate "(x,y,z)" with the visual pixels depicting the bookshelf in the image $\mathcal{I}$, thereby achieving a deep, multimodal understanding of the scene.

Stage 2: Coordinated Motion Synthesis and Realism The second stage, the Coordinated Motion Synthesizer (CMS), is tasked with translating the structured action blueprints $\mathcal{B}$ from the SAIP into high-fidelity, full-body 3D motion sequences $\mathcal{M}$. This stage moves from abstract planning to concrete physical embodiment, ensuring the final animations are realistic, environmentally aware, and socially coordinated.

Model and Conditioning Mechanism. The CMS is built upon a conditional denoising diffusion probabilistic model [Ho, Jain, and Abbeel 2020], renowned for its ability to generate diverse and high-quality data. The model learns to reverse a forward diffusion process that gradually adds Gaussian noise to a clean motion sequence $\mathbf{M}_p^0$ over T timesteps. The reverse process, starting from pure noise $\mathbf{M}_p^T \sim \mathcal{N}(0,\mathbf{I})$, iteratively refines the motion by predicting the noise ϵ to be removed at each step t.

A key innovation of our CMS lies in its holistic conditioning mechanism. Traditional text-to-motion models are typically conditioned on a single, isolated textual description. In contrast, our CMS is conditioned on the *entire set* of action blueprints $\mathcal{B} = \{B_1,\ldots,B_P\}$. This is achieved by first encoding each blueprint B_p into a latent representation using a text encoder. These representations are then aggregated and fed into the diffusion model's denoising network, $\epsilon_{\theta_{\text{CMS}}}$. This global context allows the model, while generating the motion for a single person p, to be simultaneously aware of the intentions, locations, and actions of all other agents $q \neq p$. This shared awareness is fundamental to preventing inter-personal collisions and enabling complex coordinated behaviors like yielding, gathering, or joint object attention.

Training with Coordinated Environmental Realism. The CMS is trained to produce motions that satisfy three critical criteria: kinetic realism, environmental compliance, and inter-agent coordination. To this end, we introduce the Coordinated Environmental Realism Loss ($\mathcal{L}_{CER}$), a comprehensive objective function that integrates these aspects. The CMS is trained to minimize the difference between the true noise ϵ and the predicted noise $\epsilon_{\theta_{\text{CMS}}}$, guided by this composite loss:

$$\mathcal{L}_{\text{CER}} = E_{\mathbf{M}^0,\mathcal{B},\epsilon,t}\left[w_{kin}\mathcal{L}_{kin} + w_{env}\mathcal{L}_{env} + w_{coord}\mathcal{L}_{coord}\right], \quad (3)$$

where $\mathbf{M}^0$ is the ground-truth motion from HumanVL, and $w_{(\cdot)}$ are scalar weights that balance the influence of each component:

(a) Kinetic Realism ($\mathcal{L}_{kin}$) ensures that the synthesized motion for each individual is physically plausible and adheres to the natural dynamics of human movement. We employ a standard and effective L1 reconstruction loss on the predicted noise for each person p:

$$\mathcal{L}_{kin} = \sum_{p=1}^{P}||\epsilon_p - \epsilon_{\theta_{\text{CMS}}}(\mathbf{M}_p^t, t, \mathcal{B})||_1. \quad (4)$$

(b) Environmental Realism ($\mathcal{L}_{env}$) enforces physical adherence to the static scene geometry $\mathcal{G}$, preventing unrealistic penetration of solid objects. Leveraging the 3D mesh of the scene from our HumanVL dataset, we pre-compute its Signed Distance Field (SDF) [Park et al. 2019]. An SDF is a continuous function that provides the shortest distance to the scene surface for any point in space, with the sign indicating whether the point is inside or outside. The loss penalizes any body vertex v of the predicted clean motion $\hat{\mathbf{M}}_p^0$ that enters a solid object (where SDF < 0):

$$\mathcal{L}_{env} = \sum_{p=1}^{P}\sum_{v\in\text{body}_p}\max(0, -\text{SDF}(\pi(\hat{\mathbf{M}}_p^0, v))), \quad (5)$$

where $\pi(\cdot, v)$ is a function that computes the world-space position of vertex v based on the predicted body pose. This loss term is crucial for grounding the generated motion in its physical environment.

(c) Coordination Realism ($\mathcal{L}_{coord}$) is the cornerstone of generating believable multi-person interactions. It moves beyond individual realism to model the agents as a cohesive group. This loss consists of two parts: a low-level collision penalty and a high-level semantic adherence term. The collision penalty, $\mathcal{L}_{coll}$, uses the SDF of each person's body to penalize inter-personal penetration. More importantly, the Blueprint Adherence Loss ($\mathcal{L}_{bp}$) enforces the high-level relational constraints specified in the action blueprints $\mathcal{B}$. For example, if a blueprint states "person i faces person j", $\mathcal{L}_{bp}$ introduces a term to minimize the angle between their forward-facing vectors. If an action specifies reaching for an object at a target coordinate $\mathbf{c}_{tgt}$, the loss penalizes the distance between the character's hand and $\mathbf{c}_{tgt}$.

The HumanVL Dataset

Training a model capable of sophisticated vision-language-driven motion generation requires a specialized, large-scale dataset. To this end, we construct HumanVL, the first benchmark tailored for multi-person motion synthesis in complex indoor environments from 2D images and text. As shown in Figure 3, HumanVL is structured as a collection of quintuples: $(\mathcal{I}, \mathcal{T}_{in}, \mathcal{B}_{gt}, \mathbf{M}_{gt}, \mathcal{G})$.

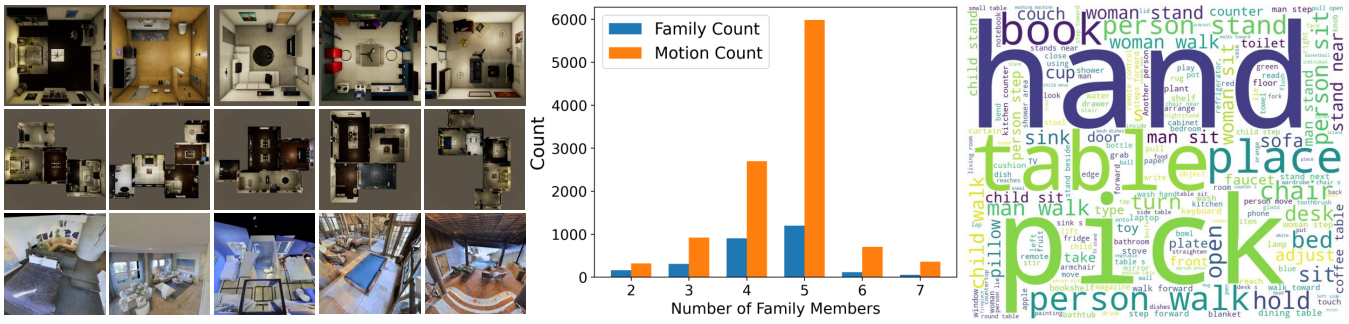

Figure 3: HumanVL Dataset. **(Left)** Sample top-down views of the multi-room household scenes. **(Middle)** Statistics showing the distribution of family sizes and the corresponding number of motion sequences. **(Right)** A word cloud of common verbs and nouns in the action blueprints, highlighting the diversity of tasks.

Dataset Curation and Statistics. HumanVL encompasses 648 unique multi-room household scenes sourced from simulators like iGibson [Li et al. 2021], Virtual-Home [Puig et al. 2018], Matterport3D [Chang et al. 2017] and AI2-THOR [Kolve et al. 2022], providing high-quality images $\mathcal{I}$ and 3D geometry $\mathcal{G}$. We designed over 2,000 diverse multi-person daily activity scenarios, from simple navigation to complex group interactions, involving 2 to 7 individuals. This resulted in a total of 35,622 coordinated motion sequences, amounting to over 20 hours of animation.

Motion and Blueprint Refinement. A core contribution of HumanVL lies in its data quality, achieved through a meticulous refinement process. We began by generating a vast pool of initial motion candidates using state-of-the-art synthesis techniques. However, these raw motions often lacked spatial coherence and environmental awareness. To rectify this, a team of 10 3D animation experts dedicated over 300 person-hours to a semi-automated, expert-in-the-loop refinement pipeline. This rigorous process included: **1) Trajectory Re-planning:** Manually adjusting motion paths to respect furniture layouts, navigate doorways, and avoid collisions. **2) Temporal and Spatial Coordination:** Synchronizing the start times, durations, and relative positioning of motions for different individuals to ensure logical and social coherence. **3) Interaction Polishing:** Fine-tuning postures for plausible object interaction, such as ensuring a character sits properly on a chair rather than hovering above it. Simultaneously, for each refined motion, we authored the corresponding high-information ground-truth Action Blueprint ($\mathcal{B}_{gt}$), carefully describing the spatial targets and relational dynamics. This intensive, expert-driven curation transforms HumanVL from a collection of raw data into a high-fidelity benchmark that directly supports the training of sophisticated, scene-aware generative models.

Experiments

Experimental Settings

Baselines. Our evaluation involves comparisons with two distinct classes of state-of-the-art methods. *1) Text2Motion Methods:* To assess the fundamental motion generation quality of our CMS module, we compare it against leading text-driven models: MDM [Tevet et al. 2023] and MotionDiffuse [Zhang et al. 2024a]. This comparison is conducted on the standard HumanML3D benchmark, isolating the motion synthesis capability from the vision-language planning stage. *2) Human-Scene Interaction Methods:* To evaluate performance on scene-aware motion generation, we compare MMG-VL with top-performing Human-Scene Interaction (HSI) models: LINGO [Jiang et al. 2024b] and TRUMANS [Jiang et al. 2024c]. Since these methods require full 3D scene inputs, we provide them with the ground-truth 3D geometry from our HumanVL test set to ensure a fair comparison. It is important to note that these baselines are limited to single-person generation, whereas our method supports multi-room multi-person scenarios. We also report the implementation details in the supplementary material.

Evaluation Datasets. Our experiments utilize two distinct datasets, each tailored to a specific evaluation aspect.

(a) HumanML3D [Guo et al. 2022] is the standard benchmark for general-purpose Text2Motion synthesis. It aggregates over 14,000 motion sequences from AMASS [Mahmood et al. 2019] and HumanAct12 [Guo et al. 2020], paired with 44,970 textual descriptions. We use this dataset exclusively for the quantitative comparison against Text2Motion baselines to measure motion quality and text-motion alignment on common actions.

(b) HumanVL is our new dataset created to evaluate vision-language-driven multi-person motion generation in complex environments. It provides comprehensive data for our primary task. The official test split, which we use for all HSI and multi-person evaluations, comprises 50 distinct household scenes and their corresponding multi-person activity instructions. To ensure robust evaluation, we generate 10 motion sequences for each test case.

Evaluation Metrics. We employ a diverse set of metrics to provide a holistic assessment of our model's performance, covering motion quality, text-motion consistency, and scene-interaction plausibility.

(a) Metrics for Text2Motion Evaluation. For the comparison on HumanML3D, we adopt standard metrics from the Text2Motion literature [Guo et al. 2023]. *1) R-Precision* and *Multimodal Distance (MMDist)* measure the alignment between the generated motions and the input text prompts;

Nums of Human	Formula-based Metrics					Perceptual-based metrics (max = 10.00)					
	$Pene_{mean}\downarrow$	$Pene_{max}\downarrow$	$Cont._{mean}\downarrow$	$Cont._{max}\downarrow$	Diversity↑	Single-person Quality↑	Spatial Distribution↑	Commonsense Constraints↑	Environmental Alignment↑	Multi-person Coordination↑	Multi-room Coverage↑
1	0.243	1.492	0.330	1.489	1.289	7.67	-	8.56	7.71	-	-
2	0.252	1.578	0.323	1.488	1.302	7.48	6.02	8.66	7.77	8.82	5.02
3	0.247	1.598	0.328	1.492	1.304	7.79	6.99	8.61	7.78	8.80	6.72
4	0.251	1.580	0.334	1.501	1.310	7.52	7.23	8.40	7.80	8.89	7.17
5	0.255	1.602	0.335	1.497	1.292	7.32	7.59	8.27	7.92	8.66	7.78
6	0.256	1.611	0.334	1.498	1.300	7.21	7.88	8.10	8.01	8.60	8.09
7	0.255	1.613	0.331	1.492	1.301	7.46	8.12	8.09	8.02	8.61	8.50

Table 1: Quantitative results for multi-person motion generation on the HumanVL dataset. We run all the evaluation 10 times. The perceptual-based evaluation was carried out by 20 PhD candidates, who rated each sample across six dimensions. Each perceptual-based dimension was scored on a scale from 0 to 10, with the final score being the average of all ratings.

we report Top-3 accuracy for R-Precision. *2) Fréchet Inception Distance (FID)* measures the distributional similarity between features of generated and ground-truth motions. *3) Diversity (Div.)* and *Multimodality (MModality)* assess the variety of motions generated from different and identical text prompts, respectively.

(b) Metrics for HSI and Multi-Person Evaluation. For evaluating performance on our HumanVL dataset, we use a combination of formula-based and perceptual metrics. **Formula-based metrics** adapted from [Zhao et al. 2023] include: *Penetration Mean/Max ($Pene_{mean/max}$)*, which measures the average/maximum body penetration into scene objects, and *Contact Mean/Max ($Cont_{mean/max}$)*, which measures the average/maximum distance of feet to the floor. **Perceptual-based metrics** are derived from a user study where human evaluators rate generated motions on a scale of 1-10 across six dimensions: *Single-person Quality (SQ)*, *Spatial Distribution (SD)*, *Commonsense Constraints (CC)*, *Environmental Alignment (EA)*, *Multi-person Coordination (MPC)*, and *Multi-room Coverage (MRC)*. These metrics provide crucial insights into the nuanced aspects of realism and coherence that are difficult to capture automatically.

Main Results

Multi-Person Motion Generation in Multi-Room Scenes We first evaluate MMG-VL's primary capability: generating coordinated motions for varying numbers of individuals in complex, multi-room environments. Since no existing work addresses this specific VL2Motion task, we conduct a comprehensive self-evaluation on our HumanVL test set.

Quantitative Analysis. Table 1 presents a detailed breakdown of performance across different group sizes. On formula-based metrics, MMG-VL consistently maintains low penetration rates ($Pene_{mean} < 0.26$) and natural foot-ground contact ($Cont_{mean} < 0.34$), demonstrating its robustness in handling scene geometry regardless of population density. The perceptual metrics reveal even more compelling insights. The model achieves high scores in Single-person Quality (SQ > 7.2), Environmental Alignment (EA > 7.7), and crucially, Multi-person Coordination (MPC > 8.6), confirming that the generated motions are not only individually plausible but also collectively coherent. As the number of people increases, the Multi-room Coverage (MRC) score steadily rises, indicating the model's ability to intelligently distribute agents across the available space, a direct outcome of the SAIP's planning capability.

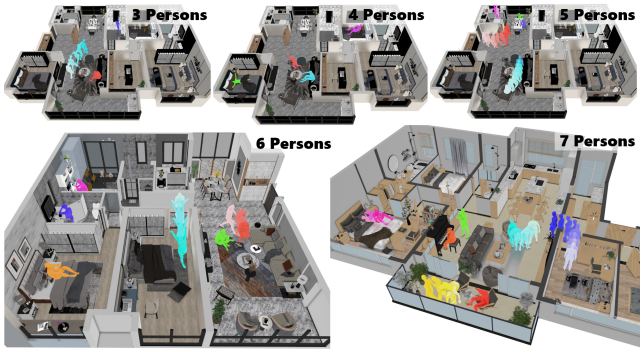

Figure 4: Qualitative results of multi-person motions generated by our MMG-VL in multi-room household scenes.

Qualitative Analysis. Figure 4 provides qualitative visualizations of generated multi-person motions. The results showcase MMG-VL's ability to produce complex, logically sound scenarios. Characters navigate different rooms, interact with furniture appropriately (e.g., sitting on chairs, lying on beds), and maintain natural interpersonal distances.

Method	$Pene_{mean}\downarrow$	$Pene_{max}\downarrow$	$Cont._{mean}\downarrow$	$Cont._{max}\downarrow$	Div.↑
TRUMANS	0.387	2.161	0.412	1.639	0.141
LINGO	0.398	2.306	0.398	1.611	0.143
MMG-VL	**0.283**	**1.492**	**0.330**	**1.389**	**1.289**

Table 2: Quantitative results for single-person motion generation with HSI baselines. We run all the evaluation 10 times.

Comparison with HSI Baselines We compare MMG-VL with leading HSI methods, LINGO [Jiang et al. 2024b] and TRUMANS [Jiang et al. 2024c], on the task of single-person, scene-aware motion generation. It is critical to note the input disparity: the HSI baselines require full 3D scene geometry as input, whereas MMG-VL only uses a single 2D image and text. We provide the baselines with the ground-truth 3D meshes from HumanVL to create the fairest possible comparison.

Quantitative Results. As shown in Table 2, MMG-VL significantly outperforms both baselines across all metrics. It achieves a 27% reduction in mean penetration ($Pene_{mean}$)

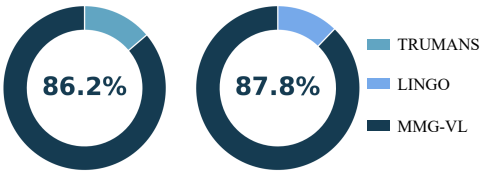

Figure 5: User study of 3D human motion generation using the two-alternative forced choice (2AFC) method, where participants show a strong preference for the generations by our approach, in comparison with all baselines.

compared to the best baseline, demonstrating superior environmental compliance. This is a direct result of our $\mathcal{L}_{env}$ loss and the SAIP's ability to plan feasible paths. Furthermore, MMG-VL generates motions with an order of magnitude higher diversity, showcasing its ability to produce varied and rich actions while respecting scene constraints.

User Preference Study. To assess perceptual quality, we conducted a two-alternative forced-choice (2AFC) user study with 200 participants. They were shown motion pairs generated by MMG-VL and a baseline under identical conditions and asked to choose the more realistic and instruction-compliant result. As depicted in Figure 5, participants overwhelmingly preferred MMG-VL's generations, with a preference rate of 86.2% against TRUMANS and 87.8% against LINGO. This strong preference highlights our model's superior ability to generate motions that are not just technically non-penetrating but also perceptually aligned with the environment and user intent.

Comparison with Text2Motion Baselines To benchmark the fundamental generative quality of our CMS module against standard text-to-motion models, we conduct a comparison with MDM [Tevet et al. 2023] and MotionDiffuse [Zhang et al. 2024a] on the HumanML3D dataset. For this experiment, MMG-VL's CMS is conditioned directly on the ground-truth text annotations, bypassing the SAIP planner, to ensure a direct, apples-to-apples comparison of the synthesis modules.

Method	R-Precision (Top 3)↑	FID↓	MMDist↓	MModality↑
MDM	0.611	0.544	5.566	2.799
MotionDiffuse	0.782	0.630	3.113	1.553
MMG-VL	**0.790**	**0.501**	**2.956**	**2.816**

Table 3: Quantitative results for single-person motion generation with Text2Motion baselines.

As shown in Table 3, our CMS achieves state-of-the-art or competitive performance across all standard metrics. Notably, it obtains the best FID and MMDist scores, indicating that the distribution of our generated motions is closest to the real data and that the motions are highly aligned with the textual descriptions. It also achieves high R-Precision and Multimodality, confirming strong text-motion consistency and generative diversity. These results validate that our framework's design, particularly the realism-focused $\mathcal{L}_{CER}$ loss, not only excels in scene-aware tasks but also enhances core motion synthesis quality, providing a solid foundation for our hierarchical approach.

Ablation Study

To validate the effectiveness of our key technical contributions, we conduct a series of ablation studies on the HumanVL dataset. We systematically remove or replace components of our proposed loss functions, $\mathcal{L}_{\text{SAIP}}$ and $\mathcal{L}_{\text{CER}}$, and evaluate the impact on generation quality. The results for a 4-person generation task are presented in Table 4.

Configuration	EA↑	MPC↑	$\text{Pene}_{\text{mean}}$↓	Div.↑
MMG-VL (full)	**7.80**	**8.89**	**0.251**	**1.310**
w/o SAIP losses:				
(a) w/o $\mathcal{L}_{SSG}$	6.95	8.75	0.259	1.305
w/o CMS losses:				
(b) w/o $\mathcal{L}_{env}$	5.82	8.51	0.873	1.321
(c) w/o $\mathcal{L}_{coord}$	7.55	6.23	0.288	1.315
(d) w/o both	5.11	5.95	0.912	1.328

Table 4: Ablation study of our proposed loss functions. We report key perceptual and formula-based metrics for a 4-person generation task on the HumanVL test set.

Removing the Spatial-Semantic Grounding Loss ($\mathcal{L}_{SSG}$) significantly degrades Environmental Alignment (EA), confirming its crucial role in forcing the SAIP to generate spatially and visually consistent action blueprints. Ablating the CMS losses reveals their distinct functions. Without the Environmental Realism Loss ($\mathcal{L}_{env}$), scene penetration ($Pene_{mean}$) increases by over 240%, highlighting its necessity for physical plausibility. Removing the Coordination Realism Loss ($\mathcal{L}_{coord}$) causes a dramatic drop in the Multi-person Coordination (MPC) score from 8.89 to 6.23. This demonstrates that $\mathcal{L}_{coord}$ is the key component that transforms independent agents into a coherent group. The complete ablation fails on all interaction metrics. These results unequivocally validate that our proposed losses are essential for generating high-quality, scene-aware, and coordinated multi-person motions. Moreover, we conduct ablation study among different input combinations, which is reported in the supplementary material.

Conclusion

This paper introduces the VL2Motion paradigm to overcome the environmental blindness of text-driven motion synthesis and the data bottleneck of HSI methods. We present its embodiment, MMG-VL, a hierarchical framework that generates coordinated, multi-person 3D motions from only a 2D image and text. MMG-VL leverages a Scene-Aware Intent Planner (SAIP) and a Coordinated Motion Synthesizer (CMS), optimized with our novel spatial grounding ($\mathcal{L}_{SSG}$) and coordinated realism ($\mathcal{L}_{CER}$) losses. To train and evaluate this new task, we built the large-scale HumanVL dataset. Experiments demonstrate that MMG-VL significantly outperforms state-of-the-art baselines and generalizes effectively to unseen environments.

Acknowledgments

This work is supported by the Young Scientists Fund of the Hunan Natural Science Foundation (Grant No.2024JJ6474), the Youth Independent Innovation Science Fund Project of NUDT (Grant No.ZK24-08).

References

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*.

Chen, W.; Xiao, C.; Gao, G.; Sun, F.; Zhang, C.; and Zhang, J. 2024. DreamArrangement: Learning Language-conditioned Robotic Rearrangement of Objects via Denoising Diffusion and VLM Planner. *Authorea Preprints*.

Fan, K.; Tang, J.; Cao, W.; Yi, R.; Li, M.; Gong, J.; Zhang, J.; Wang, Y.; Wang, C.; and Ma, L. 2024. FreeMotion: A Unified Framework for Number-free Text-to-Motion Synthesis. arXiv:2405.15763.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. arXiv:2312.00063.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

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

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

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2024a. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36.

Jiang, N.; He, Z.; Wang, Z.; Li, H.; Chen, Y.; Huang, S.; and Zhu, Y. 2024b. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

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

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; Kembhavi, A.; Gupta, A.; and Farhadi, A. 2022. AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv:1712.05474.

Lan, L.; Teng, X.; Zhang, J.; Zhang, X.; and Tao, D. 2023. Learning to purification for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 32: 3338–3353.

Lan, L.; Wang, X.; Zhang, S.; Tao, D.; Gao, W.; and Huang, T. S. 2018. Interacting tracklets for multi-object tracking. *IEEE Transactions on Image Processing*, 27(9): 4585–4597.

Li, C.; Xia, F.; Martín-Martín, R.; Lingelbach, M.; Srivastava, S.; Shen, B.; Vainio, K.; Gokmen, C.; Dharan, G.; Jain, T.; Kurenkov, A.; Liu, C. K.; Gweon, H.; Wu, J.; Fei-Fei, L.; and Savarese, S. 2021. iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks. arXiv:2108.03272.

Li, H.; Yu, H.-X.; Li, J.; and Wu, J. 2025. ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation. arXiv:2412.18600.

Li, S.; Wang, R.; Hsieh, C.-J.; Cheng, M.; and Zhou, T. 2024. MuLan: Multimodal-LLM Agent for Progressive and Interactive Multi-Object Diffusion. arXiv:2402.12741.

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

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Ma, J.; Bai, S.; and Zhou, C. 2022. Pretrained Diffusion Models for Unified Human Motion Synthesis. *arXiv preprint arXiv:2212.02837*.

Mahmood, N.; Ghorbani, N.; F. Troje, N.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *The IEEE International Conference on Computer Vision (ICCV)*.

OpenAI. 2023. Gpt-4v(ision) System Card. https://openai.com/index/gpt-4v-system-card.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Puig, X.; Ra, K.; Boben, M.; Li, J.; Wang, T.; Fidler, S.; and Torralba, A. 2018. VirtualHome: Simulating Household Activities Via Programs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8494–8502.

Richardson, E.; Goldberg, K.; Alaluf, Y.; and Cohen-Or, D. 2024. ConceptLab: Creative Concept Generation using VLM-Guided Diffusion Prior Constraints. *ACM Trans. Graph.*, 43(3).

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Sun, H.; Zheng, R.; Huang, H.; Ma, C.; Huang, H.; and Hu, R. 2024. LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model. In *ACM SIGGRAPH 2024 Conference Papers*, 1–9.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Wang, H.; Kuang, K.; Chi, H.; Yang, L.; Geng, M.; Huang, W.; and Yang, W. 2023a. Treatment effect estimation with

adjustment feature selection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2290–2301.

Wang, H.; Kuang, K.; Lan, L.; Wang, Z.; Huang, W.; Wu, F.; and Yang, W. 2023b. Out-of-distribution generalization with causal feature separation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4): 1758–1772.

Wang, H.; Yang, W.; Yang, L.; Wu, A.; Xu, L.; Ren, J.; Wu, F.; and Kuang, K. 2022. Estimating individualized causal effect with confounded instruments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1857–1867.

Xu, L.; Song, Z.; Wang, D.; Su, J.; Fang, Z.; Ding, C.; Gan, W.; Yan, Y.; Jin, X.; Yang, X.; et al. 2023. ActFormer: A GAN-based Transformer towards General Action-Conditioned 3D Human Motion Generation. *ICCV*.

Zeng, Y.; Wu, M.; Yang, L.; Zhang, J.; Ding, H.; Cheng, H.; and Dong, H. 2024. LVDiffusor: Distilling Functional Rearrangement Priors From Large Models Into Diffusor. *IEEE Robotics and Automation Letters*, 9(10): 8258–8265.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024a. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6): 4115–4128.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *arXiv preprint arXiv:2304.01116*.

Zhang, P.; Dong, X.; Zang, Y.; Cao, Y.; Qian, R.; Chen, L.; Guo, Q.; Duan, H.; Wang, B.; Ouyang, L.; Zhang, S.; Zhang, W.; Li, Y.; Gao, Y.; Sun, P.; Zhang, X.; Li, W.; Li, J.; Wang, W.; Yan, H.; He, C.; Zhang, X.; Chen, K.; Dai, J.; Qiao, Y.; Lin, D.; and Wang, J. 2024b. InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output. *arXiv preprint arXiv:2407.03320*.

Zhao, K.; Zhang, Y.; Wang, S.; Beeler, T.; ; and Tang, S. 2023. Synthesizing Diverse Human Motions in 3D Indoor Scenes. In *International conference on computer vision (ICCV)*.