Move-in-2D: 2D-Conditioned Human Motion Generation

Hsin-Ping Huang[1,2] Yang Zhou[1] Jui-Hsien Wang[1] Difan Liu[1]
Feng Liu[1] Ming-Hsuan Yang[2] Zhan Xu[1]

[1]Adobe Research [2]University of California, Merced

https://hhsinping.github.io/Move-in-2D

Abstract

Generating realistic human videos remains a challenging task, with the most effective methods currently relying on a human motion sequence as a control signal. Existing approaches often use existing motion extracted from other videos, which restricts applications to specific motion types and global scene matching. We propose Move-in-2D, a novel approach to generate human motion sequences conditioned on a scene image, allowing for diverse motion that adapts to different scenes. Our approach utilizes a diffusion model that accepts both a scene image and text prompt as inputs, producing a motion sequence tailored to the scene. To train this model, we collect a large-scale video dataset featuring single-human activities, annotating each video with the corresponding human motion as the target output. Experiments demonstrate that our method effectively predicts human motion that aligns with the scene image after projection. Furthermore, we show that the generated motion sequence improves human motion quality in video synthesis tasks.

1. Introduction

With the advancement of diffusion models, video generation has made significant progress. However, generating realistic human motion in a scene remains a nontrivial task due to the complexity of human movement. The human body is highly structured, and realistic motion generation requires models to learn and preserve articulations throughout the video. Many works [23, 28, 48, 51, 60] have improved the quality of human videos by incorporating human-specific priors, specifically by adopting motion sequences as control signals during the generation process. These driving motion sequences are typically extracted from another video of the same class, with poses mostly aligned with the target human and minimal global motion. Consequently, although these approaches enhance the quality of generated human videos, they are still limited to specific motion domains (such as dancing) and no locomotion.

In this work, we propose to **generate** a motion sequence based on a 2D background rather than relying on pre-existing driving sequences. Formally, we define 2D-conditioned human motion generation as follows: given an image representing the target scene and a text prompt describing the desired motion, we generate a motion sequence that aligns with the text description and can be projected naturally onto the scene image. This approach enables a two-pass human video generation pipeline, as shown in Fig. 1. In the first pass, human poses are positioned using a template prior, preserving body articulation and generating a plausible motion sequence. This generated motion then serves as the control signal for the subsequent video generation. Compared to methods that rely on external motion sequences, 2D-conditioned motion generation can produce sequences that consistently align with the target background and text prompt description, without being constrained by specific motion types or minimal global movement.

Although human motion generation has been extensively studied, no existing approach directly addresses this novel setting. Some methods condition motion generation solely on text prompts [1, 8, 45], which, while straightforward, may not produce motion that seamlessly integrates into specific target environments, requiring further adjustments for scene compatibility. Other methods [24, 49] generate human motion based on 3D representations of the scene, such as 3D meshes or scanned point clouds. While these methods ensure scene affordance, obtaining 3D scenes is time-consuming and demands specialized equipment and manual effort. As a result, 3D scene-aware motion generation approaches are often limited to simple motion types (walk, sit, etc.) and indoor scenes.

Our 2D-conditioned approach introduces a new modality for human motion generation by incorporating affordance awareness through the input *2D scene images*. This greatly expands the scope of existing approaches. A single 2D scene image provides semantic and spatial layout information about the target environment from a 2D per-

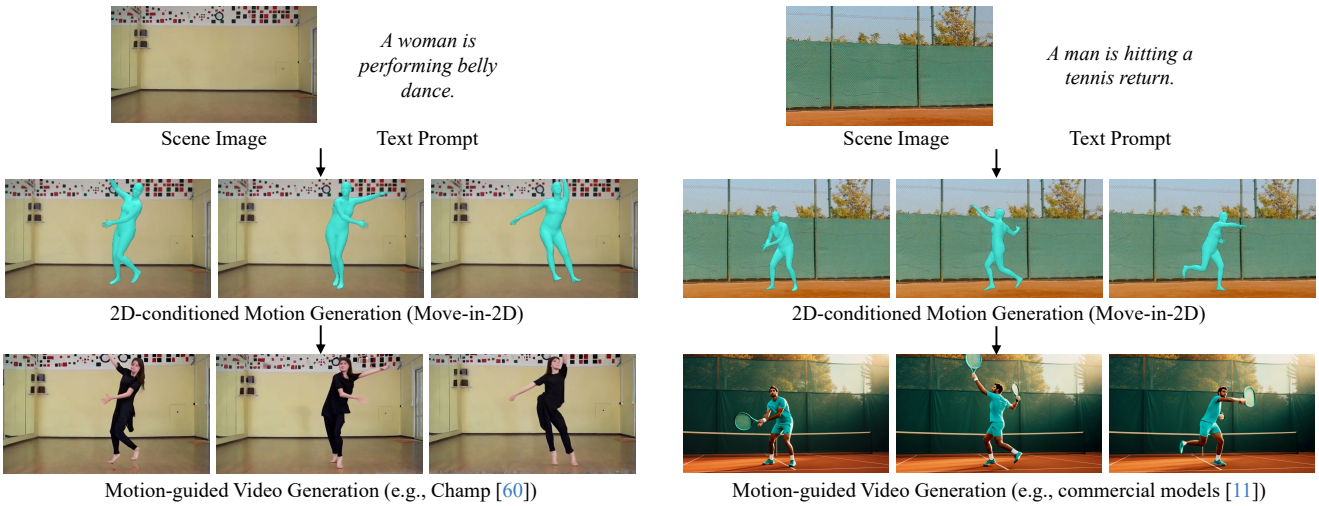

Figure 1. **2D-conditioned human motion generation.** Given an image representing the target scene and a text prompt describing the desired motion, we generate a motion sequence that aligns with the text description and projects naturally onto the scene image. This generated motion then serves as the control signal for the subsequent video generation tasks.

spective, enabling the generation of affordant human motion without the need for 3D scene reconstruction, especially for cases, e.g., video generation, when the motion is intended to be finally projected back onto a 2D plane. Furthermore, conditioning on 2D images allows for greater diversity in available scenes, as numerous online videos contain human activity in various environments. For example, outdoor scenes, which are hard to be used by 3D-aware motion generation networks, can be easily represented as 2D images and be consumed by the proposed approach.

On the other side, this novel setup also introduces several key challenges. First, training the model requires a dataset containing human motion sequences, text prompts describing the motion, and images representing the background scene. However, no existing dataset meets these requirements. Second, it remains unclear how to effectively condition the network on both text and scene image inputs. To address these challenges, we collect a large video dataset from internal data sources of open-domain internet videos. We filter the videos to ensure a static background, so that any selected frame can reliably represent the scene throughout the motion sequence. We further annotate the human motion using a state-of-the-art 3D pose estimation approach [14]. Leveraging this large-scale human motion dataset, we train a conditional diffusion model that generates human motion based on a single scene image and a text prompt. Inspired by in-context learning in large language models (LLMs) [2, 39, 50], we employ a similar strategy to convert scene and text inputs into a shared token space, integrating them within a transformer-based diffusion model for the output.

Our contributions can be summarized as follows:

- We introduce a novel task of generating human motion with a 2D image and text as conditions. It provides a more accessible way to motion generation by incorporating scene conditions without requiring 3D reconstruction.
- We collect a large human video dataset with annotated 3D human motion. The dataset significantly increases the scale of existing scene-aware motion generation datasets.
- We propose a diffusion-based network conditioned on both text and input scene images. We also show that the output motion is able to improve the quality of human motion when generating videos.

2. Related Work

Human video generation. Human-centric video generation typically leverages ControlNet to condition the generation process on motion guidance signals, such as OpenPose [23, 48], DensePose [28, 51] keypoints, or SMPL mesh sequences [60]. While these approaches deliver visually plausible results, they depend on predefined motion sequences as guidance, thus restricting their ability to generate diverse motions. In contrast, we address an orthogonal problem: generating motion guidance sequences conditioned on text prompts and a scene image. Our generated motion can then be used as guidance within human video generation frameworks.

Human motion datasets. Several datasets have been proposed to facilitate research on human motion understanding and generation. Datasets such as CMU Mocap [9], Human3.6M [25], and MoVi [12] capture human motion but lack textual descriptions of the actions. The KIT Motion Language Dataset [42] provides both motion sequences and textual prompts, containing approximately 3.9K motion sequences. HumanML3D [16] expands this number to 14.6K by sourcing motion data from HumanAct12 [15] and AMASS [34]. The Motion-X dataset [32] scales up fur-

ther to 81K motion sequences, and includes not only body movement but also facial expressions and hand poses. Despite the increasing scale of these datasets, none provide scene context aligned with the motion sequences.

Some datasets provide captured 3D scenes alongside motion sequences. Works such as [44, 52] include SMPL models with global positions in 3D scenes. The PROX dataset [18] utilizes optimization techniques with RGB-D data to reconstruct human motions, while others [3, 46] compile synthesized data of human-scene interactions. Additionally, datasets like [49, 58] incorporate both scene context and language descriptions for specific actions. However, these datasets primarily focus on indoor environments due to the challenges of 3D scene representation. Furthermore, some are oriented toward global motion prediction, resulting in limited scene diversity and a lack of detailed textual annotations.

Text-driven human motion generation. With the availability of motion datasets, human motion generation has made notable progress. Early approaches, such as Text2Action [1], utilized recurrent neural networks to capture temporal dependencies within motion sequences. Later, transformer-based architectures were introduced in works like TM2T [17] and TEACH [4], enabling improved control and the generation of longer, more coherent sequences. Building on these advancements, models such as MotionGPT [27] leverage large language model (LLM) pretraining for text-driven motion synthesis.

More recently, many works have applied diffusion models in motion generation tasks. MotionDiffuse [55] and MDM [45] generate motion sequences aligned with text prompts from an initial random noise. ReMoDiffuse [56] introduces a retrieval-augmented model, where knowledge from retrieved samples enhances motion synthesis. MLD [8] performs motion diffusion in a latent space using a variational autoencoder. EMDM [59] further reduces the number of sampling steps required during the denoising process. Although motion can be generated with only a text prompt, these methods often lack contextual alignment with specific virtual environments, limiting their direct applicability as control signals in video generation.

Scene-aware human motion generation. Given a 3D indoor scene, prior works [19, 20, 46, 47, 57, 57] have demonstrated the generation of physically plausible human poses or motion sequences. HUMANISE [49] aligns captured human motion sequences with various 3D indoor scenes, using text prompts as conditioning inputs. LaserHuman [10] incorporates real human motions within 3D environments, supporting open-form descriptions, and spanning both indoor and outdoor settings. However, due to the challenges associated with obtaining 3D scenes, these methods are often trained on limited datasets, which constrains their generalizability to more diverse, in-the-wild backgrounds.

Table 1. **Dataset statistics.** HiC-Motion is the largest dataset comprising motions, text, and diverse indoor and outdoor scenes.

Dataset	Motions	Texts	Scenes	Scene Representation	Scene Type
KIT [42]	3.9k	6.2k	No	No	Indoor
HumanML3D [16]	14.6k	44.9k	No	No	Indoor
HUMANISE [49]	19.6k	19.6k	643	RGBD	Indoor
PROX [18]	28k	No	12	RGBD	Indoor
LaserHuman [10]	3.5k	12.3k	11	RGBD	Indoor/Outdoor
Motion-X [32]	81.1k	81.1k	81.1k	Video	Indoor/Outdoor
HiC-Motion	300k	300k	300k	Video	Indoor/Outdoor

3. Humans-in-Context Motion Dataset

To advance human motion generation in 2D scenes, a large-scale video dataset capturing open-domain human motions in diverse scenes is crucial. Human-centric video datasets, such as HiC [6, 31, 38], provide millions of video clips but lack motion and text annotations. Additionally, these datasets are limited by short sequence lengths and low spatial resolutions. See Sec. 2 and Tab. 1 for more discussions. Inspired by HiC, we collect **Humans-in-Context Motion (HiC-Motion)**, a large-scale dataset of human motions capturing rich background scenes and natural language captions. Next, we describe our data collection and preprocessing pipeline.

Data collection. While action recognition datasets [30, 35] inherently include human actions, they are generally limited to short sequences and close-up shots, often omitting full-body views and background context. Our dataset is sourced from an internal dataset containing 30M open-domain internet videos. Despite the large pool, a significant portion of these videos lack human subjects. We filter for videos containing a single human moving within the scene using keypoint-based models, including Keypoint R-CNN for person detection [13] and OpenPose for keypoint prediction [7], following [6]. We retain videos with motion sequences exceeding 256 frames, resulting in a curated set of 300k videos—approximately 1% of the initial dataset. Our dataset includes high-quality, real-world videos with a range of indoor and outdoor scenes and diverse human activities, spanning daily tasks (e.g., drinking coffee, using a laptop) and sports (e.g., playing tennis, performing lunges), across more than 1k categories.

Data preprocessing. To obtain human motion annotations from the selected videos, we use the off-the-shelf method 4D-Humans [14] to extract pseudo ground-truth motions in SMPL format, ensuring high quality and frame-to-frame consistency. Since our objective is to condition human motion on scene backgrounds, we utilize Mask R-CNN to detect person masks and apply a basic inpainting model [26] to remove humans from the video frames. During training, we randomly select an inpainted frame from each video to serve as the background image. To enhance the model's generalization to unseen scenes, we apply color adjustments to simulate diverse lighting conditions in the background

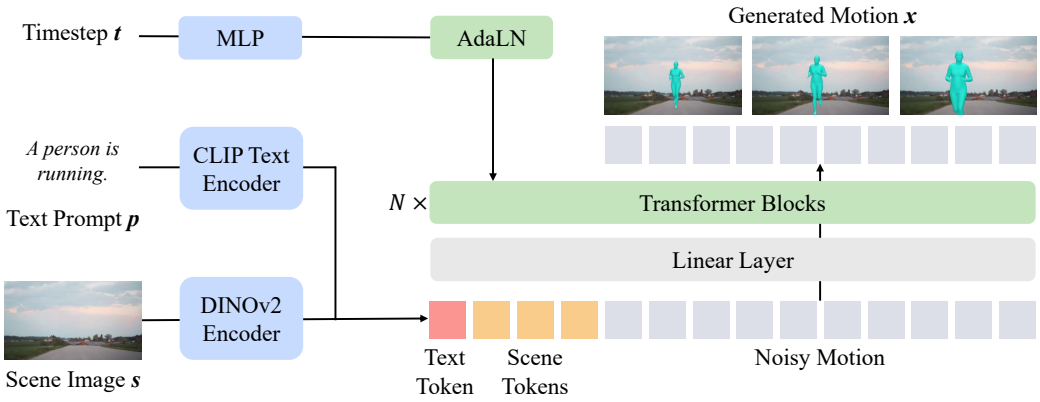

Figure 2. **Overview.** The text prompt and background scene image are encoded by the CLIP and DINO encoders, and incorporated into the model via in-context conditioning. The AdaLN layer receives the diffusion timestep as input. Our multi-conditional transformer model then generates a human motion sequence through a diffusion denoising process, aligning the generated motion with both input conditions.

images [29] as well as random cutout augmentations.

4. Approach

Given a text prompt and a background scene image, our objective is to generate a human motion sequence that aligns with the action described in the text prompt while maintaining physical compatibility with the background scene. We begin with a preliminary overview of diffusion model in Sec. 4.1. In Sec. 4.2, we propose a conditional motion diffusion model. We then introduce a multi-conditional transformer in Sec. 4.3. Finally, in Sec. 4.4, we present our training strategy. Fig. 2 provides an overview of our approach.

4.1. Preliminaries on Diffusion Models

Diffusion models like DDPM [22, 45] approximate the data distribution via a forward and backward process. In the forward process, Gaussian noise is added to the sample $\mathbf{x}_0$, resulting in $\mathbf{x}_t$. The model $\mathcal{M}$ learns to reverse this process by denoising $\mathbf{x}_t$ conditioned on timestep t and context c. The training minimizes the MSE loss between the predicted clean sample $\hat{\mathbf{x}}_0 = \mathcal{M}(\mathbf{x}_t|t, c)$ and the ground truth $\mathbf{x}_0$, i.e., $\mathcal{L}_{\text{mse}} = \mathbb{E}_{\mathbf{x}_0,t} \|\mathbf{x}_0 - \mathcal{M}(\mathbf{x}_t|t, c)\|^2$. During sampling, the model iteratively predicts $\hat{\mathbf{x}}_0$ at each timestep t over T steps to recover $\mathbf{x}_0$. Classifier-free guidance (CFG) [21] is applied to enhance alignment with conditions.

4.2. Conditional Motion Diffusion

Given input conditions, including a text prompt p and a background scene image s, we train a conditional motion diffusion model to generate the target human motion $\mathbf{x}$. The target human motion is represented as a sequence of N human poses, each with a dimensionality of D. Each pose is parameterized by body pose parameters $\theta_b \in \mathbb{R}^{23\times6}$, which capture 6D rotations for 23 SMPL joints [33], along with a global orientation parameter $\theta_g \in \mathbb{R}^6$ that defines the overall human body orientation. Unlike previous motion generation approaches [8, 45], we aim to generate human motion that projects naturally onto a 2D background scene image. Therefore, our model predicts an additional camera translation parameter $\pi \in \mathbb{R}^3$, assuming a perspective camera with fixed focal length and intrinsics to project SMPL space points onto the image plane. We train the conditional diffusion model $\mathcal{M}$ by randomly dropping the text prompt p and background scene s conditions by a probability of q where $q = 0.1$ in our experiments. During sampling, we apply CFG to both the text and scene conditions jointly to enhance the alignment between the generated motion and the input conditions, where g is the guidance scale.

$$\mathcal{M}_{\text{cfg}} = \mathcal{M}(\mathbf{x}_t|t) + g\left(\mathcal{M}(\mathbf{x}_t|t, p, s) - \mathcal{M}(\mathbf{x}_t|t)\right). \quad (1)$$

4.3. Multi-Conditional Transformer

We inject the text prompt and scene conditions into a diffusion transformer model to generate a motion sequence that aligns semantically with the input description and is physically compatible with the scene after projection to 2D. Our model architecture follows the diffusion transformer [39, 45]. The motion sequence $\mathbf{x} \in \mathbb{R}^{T\times D}$ is projected into the transformer's hidden size, with positional embeddings added to the tokens before feeding them into a series of transformer blocks. The output tokens are then linearly projected back to obtain the motion prediction $\hat{\mathbf{x}}$.

We now describe the condition encoding and injection process. We encode the diffusion timestep t by a positional embedding layer, the text prompt p by a CLIP encoder [43], and the background image condition s by a DINO encoder [37] which preserves the spatial relationships across patches. Each condition is then projected to the transformer dimension. To guide the diffusion process, we employ simple yet effective methods [39] for injecting the conditions into the transformer blocks:

- **In-context conditioning.** The conditions are concatenated to the motion sequence as additional tokens, which

are removed from the output sequence without being calculated the loss.

- **Adaptive layer normalization (AdaLN) [40].** A linear layer predicts the scale and shift parameters from the condition tokens, which are then applied to the motion sequence.
- **Cross-attention layer.** A cross-attention layer is inserted after the self-attention layer to take the input conditions.

We observe that in-context learning for both the text and scene modalities improves the model's ability to capture interactions between the inputs by converting them into a shared token space, leading to better alignment with both conditions. Using AdaLN for the diffusion timestep condition enhances the temporal smoothness of the generated motions. We thus adopt this configuration as our main framework (see Sec. 5.2).

4.4. Training Strategy

We adopt a two-stage training strategy to generate human motion aligned with text prompts and backgrounds. It first learns to generate diverse motion sequences and then to disentangle human motion from camera effects.

Selection of fine-tuning set. Our motion sequences are extracted from internet videos, which include both human and camera motion. For example, camera movement to the right can cause the pose sequence to shift left. However, our goal is to represent the scene using a single image, independent of any camera motion. To achieve this, we calculate the optical flow of the raw videos and use the median flow value to select videos with minimal background motion. Additionally, to address data distribution biases where many daily activities involve limited human movement (e.g., sitting), we select a subset of videos with significant human motion (exceeding 200 pixels of movement) to encourage the generation of sequences with large motion dynamics.

Two-stage training. We adopt a two-stage training strategy to generate human motion aligned with text prompts and backgrounds. In the first stage, the model is trained on the full 300k video dataset for 600k iterations to learn scene semantics and generate diverse motion sequences based on text prompts. In the second stage, we fine-tune the model on a mixed dataset of 150k videos, consisting of 60% large-motion videos and 40% fixed-background videos, for an additional 600k iterations. This fine-tuning enhances the model's ability to generate human motion that decouples camera-induced movement and improves the generation of large-motion dynamics.

5. Experiments

Evaluation data. To address the lack of benchmarks for evaluating human motion generation conditioned on 2D scenes, we construct a test set comprising text prompts, scene images, and ground-truth motion sequences for comprehensive evaluation. Our test set is sampled from a held-out portion of the HiC-Motion dataset. We first curate 100 high-frequency verb phrases from the data to serve as text prompts (e.g., "A person drinks coffee"). For each text prompt, we sample 10 videos and randomly select one frame, where the human is removed, as the corresponding scene image. This process results in a total of 957 test samples.

Evaluated methods. To assess the effectiveness of the proposed models in generating human motions that align with text prompts and are compatible with scene images, we compare them against state-of-the-art motion generation models conditioned on single or multiple modalities. Specifically, *MDM* [45] and *MLD* [8] generate motion conditioned solely on text prompts. While no existing methods generate motion conditioned on 2D scene images, we include models that utilize 3D point clouds to produce affordance-aware motion. We first employ a pretrained depth prediction model [53] to estimate scene depth, then back-project our 2D scene images into 3D point clouds as the input conditions to baselines. We compare to the following scene-conditioned approaches: *SceneDiff* [24], which uses 3D point clouds as input, and *HUMANISE* [49], the closest approach to ours, which conditions on both text prompts and 3D point clouds. Additionally, we extend MDM by training it on our HiC-Motion dataset, denoted as *MDM+*. We also evaluate two variants of our model: *Ours*, conditioned on both text prompts and scene images. *Ours-scene*, conditioned solely on scene images.

Evaluation metrics. To evaluate the quality and diversity of the generated human motions, previous works [15, 41] utilize a pre-trained human motion classifier to extract motion features for evaluation. However, due to the lack of motion feature extractors trained on open-domain videos, we train our classifier based on STGCN [41, 54] using motion sequences of length 256, with each pose represented by 21 SMPL joints in 6D rotation format. To standardize outputs across models, we ignore global orientation and translation. We evaluate the models using four metrics:

- **FID** assesses the overall quality of the generated motions by computing the distance between the feature distributions of generated and real motions.
- **Accuracy** evaluates the alignment between generated motions and input prompts by calculating the recognition accuracy of the generated motions.
- **Diversity** quantifies the variation across generated motions by calculating the distance between two randomly sampled subsets of generated motions from all prompts.
- **Multimodality** measures variation within identical prompts by computing the distance between two subsets of motions generated from the same prompts.

Implementation details. Our model generates motion sequences of length $N = 256$ and feature dimension $D =$

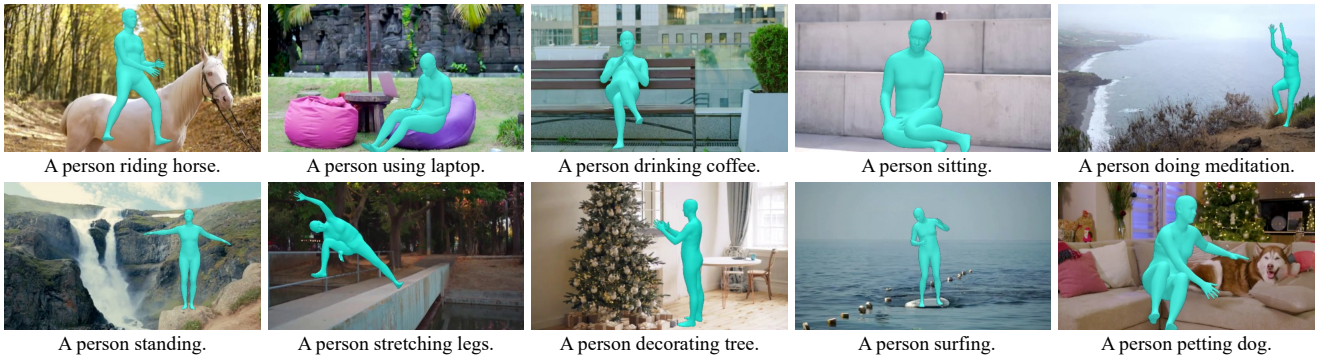

Figure 3. **Affordance-aware human generation.** Our model generates human poses consistent with both text prompts and scene context, such as standing on a cliff. It also supports complex human-scene interactions, including activities like petting a dog.

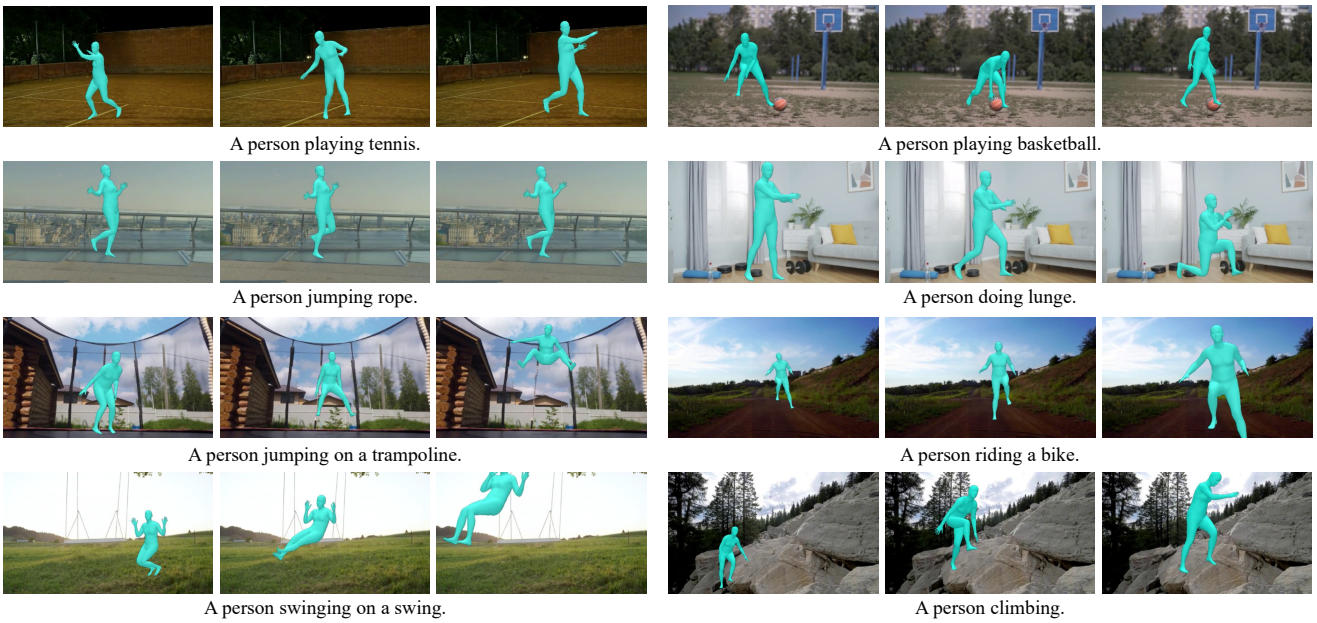

Figure 4. **Motion generation with large dynamics.** Our results show motion sequences that are accurately placed and move within scenes, such as playing tennis, enabling the generation of complex human activities that are challenging for video generation models.

147, using an architecture with 8 transformer blocks, 512 hidden units, feedforward layers of size 2048, and 4 attention heads. It is trained with the Adam optimizer with a learning rate of 0.0002, batch size 128 for 1.2M iterations, and 1000 diffusion steps with a cosine noise schedule. Scene images with resolution 168×280 are encoded into 240 tokens by DINO-B [37], and text prompts into a single token using CLIP-B [43], resulting in a sequence of 497 tokens.

5.1. Qualitative Results

Affordance-aware human generation. Fig. 3 shows that our model generates human poses that are consistent with both the text prompts and the scene context, such as standing at the edge of a cliff, sitting on a chair, and surfing on a board. In addition, our model is capable of generating complex human-scene interactions, including activities such as riding a horse, decorating a tree, and petting a dog.

Motion generation with large dynamics. In Fig. 4, we present examples with larger motion dynamics. Our results show strong scene compatibility, with the human sequence correctly placed and moving in environments like jumping on a trampoline. Our model generates complex human activities with detailed pose sequences aligned with text prompts, such as playing tennis, which is challenging for video generation models and is effectively handled by our approach.

Comparison to state-of-the-art. As shown in Fig. 5, the text-conditioned method *MDM* fails to generate plausible poses. *MLD* correctly generates the running action, but the motion is not compatible with the scene, especially since it fails to generate the person moving toward the camera. The

Figure 5. **Comparison to state-of-the-art.** *MDM* and *SceneDiff* produces implausible poses, *MLD* generates mismatched motion with the scene, and *HUMANISE* generates static poses. Our method generates coherent motion aligned with both the scene and text prompts.

Table 2. **Quantitative results.** Our method achieves better quality and diversity scores compared to state-of-the-art text-conditioned, scene-conditioned, and multimodal motion generation models.

Methods	FID (↓)	Accuracy (↑)	Diversity (↑)	Multimodality (↑)
MDM [45]	164.595	0.325	24.758	18.924
MLD [8]	85.913	0.322	25.119	19.464
SceneDiff [24]	543.769	0.203	4.217	3.861
HUMANISE [49]	159.935	0.225	23.287	19.956
MDM+ [45]	46.035	0.620	23.002	17.627
Ours-scene	46.458	0.482	24.968	**21.320**
Ours	**44.639**	**0.661**	**26.027**	20.130

Table 3. **Automated evaluation.** We report average VLM scores (0-5) for generated motions, assessing alignment with scene, text, and pose quality. Our method outperforms all evaluated baselines.

Methods	Scene-Align (↑)	Text-Align (↑)	Quality (↑)	Total (↑)
MDM [45]	2.25	1.35	1.50	5.10
MLD [8]	2.85	1.95	1.90	6.70
SceneDiff [24]	2.05	1.20	1.20	4.45
HUMANISE [49]	2.20	1.45	1.30	4.95
MDM+ [45]	2.57	1.73	1.94	6.24
Ours-scene	2.90	2.00	1.95	6.85
Ours	**3.55**	**2.70**	**2.85**	**9.10**

scene-conditioned method *SceneDiff* struggles to generate accurate human poses, while *HUMANISE* produces static poses throughout the sequence. These methods, trained on limited synthetic point cloud data, have difficulty adapting to real-world scene conditions. In contrast, our method generates motion that is both coherent within the scene and aligned with the text prompts.

5.2. Quantitative Results

The evaluation results are shown in Tab. 2. We observe that the scene-conditioned motion generation models, *HUMANISE* and *SceneDiff*, achieve a higher FID and lower recognition accuracy compared to our methods and the text-conditioned baselines, *MDM* and *MLD*. Since *HUMANISE* and *SceneDiff* are trained on limited synthetic 3D point clouds (e.g., 643 indoor scenes in ScanNet), these models struggle to generalize to real-world point clouds constructed from single images in diverse indoor and outdoor scenes, leading to lower motion quality. Compared to the models conditioned on text alone, the advanced model *MLD* achieves better metrics than *MDM*. By training on our large-scale human motion dataset with 300k sequences, *MDM+* achieved a 72% lower FID and a 90% higher accuracy compared to *MDM*, which is trained on the HumanML3D dataset with only 14k motion sequences. This result highlights the significant improvement in human motion generation enabled by training on our large-scale HiC-Motion dataset extracted from real-world videos.

Among models trained on the same backbone and dataset but with different input conditions (i.e., *MDM+*, *Ours-scene*, *Ours*), *Ours* achieves the lowest FID score, the highest accuracy and diversity. *Ours* achieves 37% higher accuracy compared to *Ours-scene*, indicating that the in-context conditioning method effectively enables the model to generate actions aligned with specific prompts. On the other hand, *Ours-scene* achieves a higher multimodality score, which measures diversity within the same prompt. As *Ours-scene* lacks the text constraint, it exhibits greater variation in outputs for identical prompts.

Automated evaluation. Since there is currently no established metric to assess the compatibility between generated motion sequences and 2D background images, we employ the vision-language model (VLM) ChatGPT-4o [36] for automated evaluation. Given the generated SMPL pose rendered on the background image alongside the input text prompt, the VLM provides scores on a 0-5 scale for the following criteria: 1) alignment of the pose with the background, 2) alignment of the pose with the text prompt, and 3) overall quality of the generated pose. For consistency, we use the middle frame of each generated sequence for evaluation. Average scores over 20 test set videos are reported in Tab. 3. Our method consistently outperforms

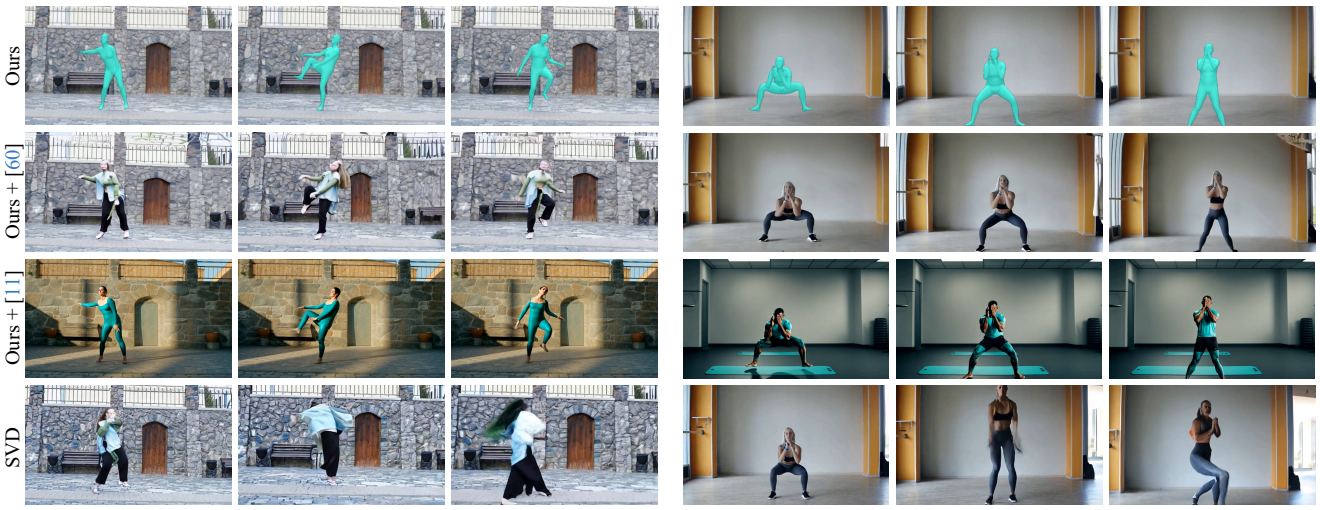

Figure 6. **Motion-guided human video generation.** Our approach generates scene-compatible motion sequences from a scene image and text prompt, which are then used to animate a reference human using Champ [60] or the commercial model [11]. The generated motion ensures accurate human shapes and smooth motion in the resulting videos, outperforming SVD [5] in preserving human geometry and motion consistency.

Table 4. **Ablation study.** We study different transformer block designs, and choose *AdaLN* for timestep conditioning and *In-Context* for text and scene conditions as our main configuration.

Timestep	Text	Scene	FID (↓)	Accuracy (↑)
AdaLN	In-Context	In-Context	**44.639**	**0.661**
AdaLN	In-Context	Cross-Attn	47.656	0.567
In-Context	In-Context	In-Context	62.927	0.554
In-Context	In-Context	Cross-Attn	66.827	0.519

the compared approaches across all criteria, achieving the highest score of 3.55 for alignment with the scene, demonstrating that our in-context framework effectively enforces affordance-aware motion generation on 2D scenes.

Ablation study. As discussed in Sec. 4, we train our 2D-conditioned motion diffusion models using various transformer block designs, including AdaLN, in-context, and cross-attention layers, to condition on the timestep, text, and scene inputs. We evaluate four models, each with a different combination of these conditioning methods in Tab. 4. Our results demonstrate that the model incorporating *AdaLN* for timestep conditioning and *In-Context* conditioning for text and scene inputs achieves the best FID and accuracy. Thus, we adopt this setup in our main framework.

5.3. Motion-guided Human Video Generation

One important downstream application supported by our approach is human video generation guided by motion sequence. We employ a two-stage approach: first, given a scene image and text prompt, our model generates a scene-compatible motion sequence. Next, using this generated motion sequence and a reference human in the scene, we apply Champ [60] to animate the reference human guided by the generated motion, enabling the creation of affordance-aware human videos that align with the target background. Additionally, we use the commercial model [11] to generate a motion-guided video. Although the commercial model does not preserve the original background, our generated motion still serves as an effective guidance signal for the human subject, while the scene image provides the desired background's layout and semantic information. As shown in Fig. 6, the accurate and smooth motion sequences generated by our model allow both Champ and the commercial model to produce videos with detailed human shapes and clean motion. Our method generates 256-frame sequences of complex activities such as dancing and playing tennis (see Fig. 1). We also include results from Stable Video Diffusion (SVD) [5] using the same reference frame. Without pose guidance, SVD generates incomplete human geometry and inconsistent, blurry results, underscoring the advantages of using our method to generate intermediate pose sequences for video generation.

6. Conclusions

We introduced a novel task of generating human motion conditioned on a scene image. Our approach employs a conditional diffusion model enhanced by in-context learning techniques. To support this, we collected a large-scale dataset of diverse human activities and environments for model training. Our method effectively predicts 2D-aligned human motion and improves motion quality in video generation. Despite these advancements, our framework does not control camera movement in generated motions, and the two-pass video generation pipeline has not been jointly optimized. We leave these aspects for future work.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, 2018. 1, 3

[2] Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. How do in-context examples affect compositional generalization? In *ACL*, 2023. 2

[3] 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. 3

[4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, 2022. 3

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 8

[6] Tim Brooks and Alexei A Efros. Hallucinating pose-compatible scenes. In *ECCV*, 2022. 3

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 3

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 1, 3, 4, 5, 7

[9] CMU. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. 2

[10] 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

[11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 8

[12] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. Movi: A large multi-purpose human motion and video dataset. *Plos one*, 2021. 2

[13] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 3

[14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 3

[15] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, 2020. 2, 5

[16] 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. 2, 3

[17] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, 2022. 3

[18] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019. 3

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

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, 2021. 3

[21] Jonathan Ho. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4

[23] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 1, 2

[24] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 1, 5, 7

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2014. 2

[26] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015. 3

[27] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2024. 3

[28] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, 2023. 1, 2

[29] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 4

[30] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[31] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *CVPR*, 2023. 3

[32] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *NeurIPS*, 2023. 2, 3

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 4

[34] 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. 2

[35] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE TPAMI*, pages 1–8, 2019. 3

[36] OpenAI. Chatgpt, 2024. 7

[37] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 6

[38] Boxiao Pan, Zhan Xu, Chun-Hao Paul Huang, Krishna Kumar Singh, Yang Zhou, Leonidas J. Guibas, and Jimei Yang. Actanywhere: Subject-aware video background generation. In *NeurIPS*, 2024. 3

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 4

[40] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 5

[41] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021. 5

[42] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. 2, 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 6

[44] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE TVCG*, 2023. 3

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1, 3, 4, 5, 7

[46] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, 2021. 3

[47] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021. 3

[48] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *CVPR*, 2024. 1, 2

[49] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *NeurIPS*, 2022. 1, 3, 5, 7

[50] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *TMLR*, 2022. 2

[51] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024. 1, 2

[52] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *CVPR*, 2023. 3

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 5

[54] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *AAAI*, 2018. 5

[55] 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

[56] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023. 3

[57] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *CVPR*, 2020. 3

[58] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, 2022. 3

[59] 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. 3

[60] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 1, 2, 8