UNIVERSITY OF CALIFORNIA, MERCED

Generative Models for Novel Views and Video Synthesis

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering and Computer Science

by

Hsin-Ping Huang

Committee in charge:

Professor Ming-Hsuan Yang, Chair
Professor Shawn Newsam
Professor Meng Tang
Doctor Yu-Chuan Su

2025

The dissertation of Hsin-Ping Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Professor Shawn Newsam)

(Professor Meng Tang)

(Doctor Yu-Chuan Su)

(Professor Ming-Hsuan Yang, Chair)

University of California, Merced

2025

DEDICATION

To my beloved parents and brother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

VITA

2020–2025	Ph. D. in Electrical Engineering and Computer Science, University of California, Merced
2017–2020	M. S. in Computer Science, The University of Texas at Austin
2013–2017	B. S. in Electrical Engineering, National Taiwan University

PUBLICATIONS

Hsin-Ping Huang, Yang Zhou, Jui-Hsien Wang, Difan Liu, Feng Liu, Ming-Hsuan Yang, Zhan Xu, "Move-in-2D: 2D-Conditioned Human Motion Generation," *CVPR*, 2025.

Hsin-Ping Huang, Xinyi Wang, Yonatan Bitton, Hagai Taitelbaum, Gaurav Singh Tomar, Ming-Wei Chang, Xuhui Jia, Kelvin C.K. Chan, Hexiang Hu, Yu-Chuan Su, Ming-Hsuan Yang, "KITTEN: A Knowledge-Intensive Evaluation of Image Generation on Visual Entities," a submission to *NeurIPS*, 2025.

Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, Ming-Hsuan Yang, "Fine-grained Controllable Video Generation via Object Appearance and Context," *WACV*, 2025.

Hsin-Ping Huang, Yu-Chuan Su, Ming-Hsuan Yang, "Generating Long-take Videos via Effective Keyframes and Guidance," *WACV*, 2025.

Hsin-Ping Huang, Charles Herrmann, Junhwa Hur, Erika Lu, Kyle Sargent, Austin Stone, Ming-Hsuan Yang, Deqing Sun, "Self-supervised AutoFlow," *CVPR*, 2023.

Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, Ming-Hsuan Yang, "Adaptive Transformers for Robust Few-shot Cross-domain Face Anti-spoofing," *ECCV*, 2022.

Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, Ming-Hsuan Yang, "Learning to Stylize Novel Views," *ICCV*, 2021.

Hsin-Ping Huang, Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, "Semantic View Synthesis," *ECCV*, 2020.

ABSTRACT OF THE DISSERTATION

Generative Models for Novel Views and Video Synthesis

by

Hsin-Ping Huang
Doctor of Philosophy in Electrical Engineering and Computer Science
University of California Merced, 2025
Professor Ming-Hsuan Yang, Chair

Content creation aims to empower users to visualize complex ideas and generate rich media content with minimal manual effort. Recent advancements in generative models have significantly automated the creative process, producing increasingly realistic and high-quality outputs. These developments lower the barrier for non-professional users to create compelling artistic content without requiring extensive domain expertise.

In particular, techniques such as Neural Radiance Fields (NeRF), 3D Gaussian Splatting, and recent text-to-video generation models have shown great promise in facilitating 3D content creation and synthesizing high-fidelity videos from natural language prompts. Despite these impressive results, several key challenges remain. Existing models often struggle with: 1) maintaining temporal and spatial consistency across multiple views and video frames; 2) offering fine-grained controllability over individual subjects; and 3) generating complex content involving articulated subjects.

This thesis proposes several solutions to address these challenges. First, we introduce a point cloud-based method for consistent and efficient 3D scene stylization. Next, we present a video generation model that enables fine-grained control over object appearance and spatial placement. We further design a human motion generation framework to support video synthesis involving complex and articulated human actions. Finally, we propose a

long-take video generation model capable of producing temporally coherent content over extended durations.

First, we propose a 3D scene stylization framework that generates consistent, stylized novel views. Our method back-projects image features into 3D space and modulates the scene using a learnable linear transformation matrix. The stylized scene is then projected back into 2D space to render novel views with consistent style and geometry.

Second, we present a controllable video generation framework that provides an intuitive interface for manipulating the trajectory and appearance of individual objects. To achieve this, we integrate control signals into text-to-video models through a multimodal conditioning module that incorporates a joint encoder and control-attention layers.

Third, we propose an image-conditioned human motion generation method to improve the quality of human motion in video synthesis tasks. Specifically, we employ a diffusion transformer model trained on our collected video dataset with annotated human motion to generate motion sequences that are semantically and spatially aligned with a given scene image.

Finally, we introduce a framework to generate long-take videos containing multiple coherent events. Specifically, we decouple video generation into two stages. The keyframe generation stage focuses on creating multiple coherent events, while the frame interpolation stage generates smooth intermediate frames to ensure temporal continuity.

Chapter 1

Introduction

1.1 Overview

Recent advances in content creation have led to remarkable progress across a wide range of applications, significantly supporting the creative process and enabling users to visualize their imagination with increasing fidelity. While image generation has achieved impressive quality and flexibility, several challenges remain in the areas of novel view synthesis and video generation.

In the field of novel view synthesis, methods such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting have demonstrated strong capabilities in modeling fine scene details with improved training efficiency. However, difficulties arise when applying these models to visual editing tasks such as object removal, scene extrapolation, and style transfer, where maintaining consistency across multiple synthesized views remains a significant challenge. In the area of video generation, researchers have developed powerful base models in recent years. Despite these advances, limitations persist in enabling controllability and coherence during generation. These limitations include achieving user-defined trajectory control, camera control, customization of appearances across multiple subjects, maintaining scene and object consistency across multiple shots, and generating long and

complex motions for articulated subjects.

To address these limitations, we propose effective and efficient solutions that enhance the capabilities of generative models for novel view synthesis and video generation. Our methods aim to achieve the following goals: 1) consistency across multiple views and video frames, 2) fine-grained controllability at the subject level, and 3) the ability to generate complex and coherent content involving articulated subjects.

1.2 Organization

In Chapter 2, we tackle a 3D scene stylization problem — generating stylized images of a scene from arbitrary novel views given a set of images of the same scene and a reference image of the desired style as inputs. Direct solution of combining novel view synthesis and stylization approaches lead to results that are blurry or not consistent across different views. We propose a point cloud-based method for consistent 3D scene stylization. First, we construct the point cloud by back-projecting the image features to the 3D space. Second, we develop point cloud aggregation modules to gather the style information of the 3D scene, and then modulate the features in the point cloud with a linear transformation matrix. Finally, we project the transformed features to 2D space to obtain the novel views. Experimental results on two diverse datasets of real-world scenes validate that our method generates consistent stylized novel view synthesis results against other alternative approaches.

In Chapter 3, we present FACTOR for fine-grained controllable video generation. While text-to-video generation shows state-of-the-art results, fine-grained output control remains challenging for users relying solely on natural language prompts. FACTOR provides an intuitive interface where users can manipulate the trajectory and appearance of individual objects in conjunction with a text prompt. We propose a unified framework to integrate these control signals into an existing text-to-video model. Our approach involves a multimodal condition module with a joint encoder, control-attention layers, and an ap-

pearance augmentation mechanism. This design enables FACTOR to generate videos that closely align with detailed user specifications. Extensive experiments on standard benchmarks and user-provided inputs demonstrate a notable improvement in controllability by FACTOR over competitive baselines.

In Chapter 4, 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.

In Chapter 5, we tackle the challenge of generating long-take videos encompassing multiple non-repetitive yet coherent events. Existing approaches generate long videos conditioned on single input guidance, often leading to repetitive content. To address this problem, we develop a framework that uses multiple guidance sources to enhance long video generation. The main idea of our approach is to decouple video generation into keyframe generation and frame interpolation. In this process, keyframe generation focuses on creating multiple coherent events, while the frame interpolation stage generates smooth intermediate frames between keyframes using existing video generation models. A novel mask attention module is further introduced to improve coherence and efficiency. Experiments on challenging real-world videos demonstrate that the proposed method outperforms prior methods by up to 9.5% in objective metrics.

In Chapter 6, we summarize the main contributions of this thesis and discuss several potential directions for future research, including 1) real-time style transfer for dynamic scenes and 2) learning to generate human-scene interactions from videos.

Chapter 2

Learning to Stylize Novel Views

We tackle a 3D scene stylization problem — generating stylized images of a scene from arbitrary novel views given a set of images of the same scene and a reference image of the desired style as inputs. Direct solution of combining novel view synthesis and stylization approaches lead to results that are blurry or not consistent across different views. We propose a point cloud-based method for consistent 3D scene stylization. First, we construct the point cloud by back-projecting the image features to the 3D space. Second, we develop point cloud aggregation modules to gather the style information of the 3D scene, and then modulate the features in the point cloud with a linear transformation matrix. Finally, we project the transformed features to 2D space to obtain the novel views. Experimental results on two diverse datasets of real-world scenes validate that our method generates consistent stylized novel view synthesis results against other alternative approaches.

2.1 Introduction

Visual content creation in 3D space has recently attracted increasing attention. Driven by the success of 3D scene representation approaches [141, 166, 249], recent methods have made significant progress on various content creation tasks for 3D scenes, such as

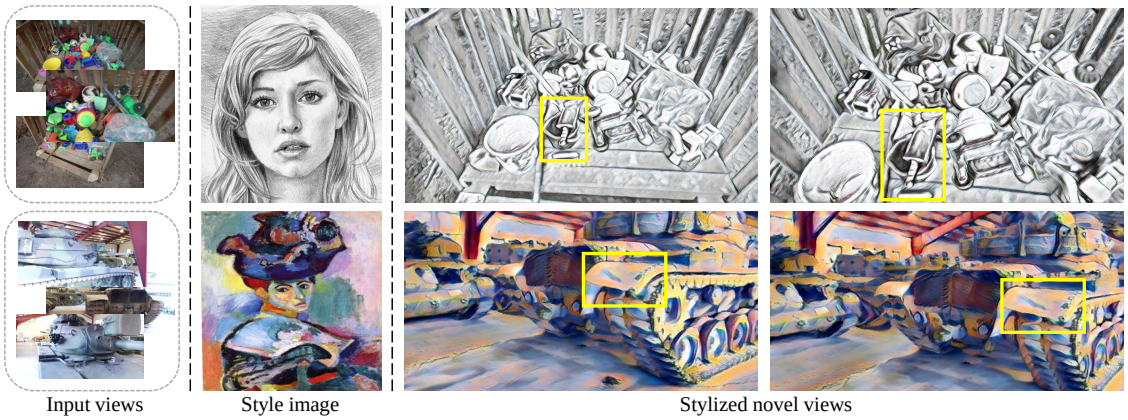

Figure 2.1: **3D scene stylization**. Given a set of images of a 3D scene (*left*) as well as a reference image of the desired style (*middle*), our method is able to modify the style of the 3D scene, and synthesize images of arbitrary novel views (*right*). The novel view synthesis results 1) contain the desired style and 2) are consistent across various novel views, *e.g.,* the texture in the yellow boxes.

semantic view synthesis [71, 87] and scene extrapolation [127]. In this work, we focus on the *3D scene stylization* problem. As shown in Figure 2.1, given a set of images of a target scene and a reference image of the desired style, our goal is to render stylized images of the scene from arbitrary novel views. 3D scene stylization enables a variety of interesting virtual reality (VR) and augmented reality (AR) applications, *e.g.,* augment the street scene at user locations to the *Cafe Terrace at Night* style by van Gogh.

Learning to modify the style of an existing 3D scene is challenging for two reasons. First, the synthesized novel views (*i.e.,* 2D images) of the stylized 3D scene must contain the desired style provided by the reference image. Second, since our goal is to stylize the *holistic* 3D scene, the generated novel views need to be consistent across different viewpoints for the same scene, such as the texture in the yellow boxes shown in Figure 2.1.

To handle these challenges, one plausible solution is to combine existing novel view synthesis [166, 249] and image stylization approaches [116, 184]. However, such straight-

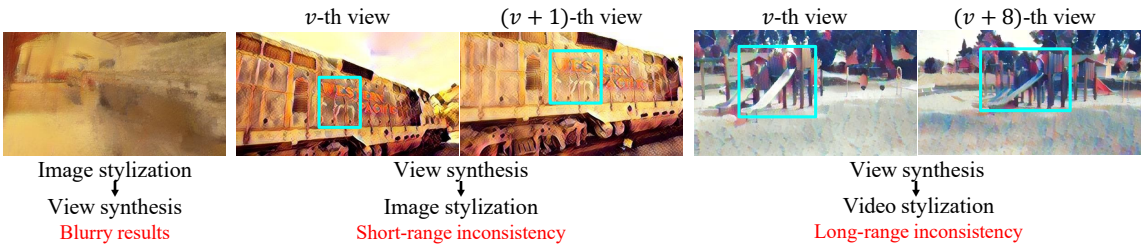

Figure 2.2: **Motivation.** While the existing methods can be used for the 3D scene stylization task, these methods either produce blurry (image stylization → novel view synthesis), short-range inconsistent (novel view synthesis → image stylization), or long-range inconsistent (novel view synthesis → video stylization) results.

forward approaches lead to problematic results since image stylization schemes are not designed to consider the consistency issue across different views for the same scene. We present the examples in Figure 2.2 where the results may be blurry if the input images of the target scene are stylized before conducting novel view synthesis. On the other hand, if we apply image stylization after novel view synthesis, the results are not consistent across different views. Another possible solution is to treat a series of novel view synthesis results as a *video*, and use the video stylization frameworks [43, 55, 202] to obtain temporally consistent results. However, as shown in Figure 2.2, these approaches are not able to enforce long-range consistency (*i.e.,* between two far-away views) as the video stylization schemes only guarantee the short-term consistency.

In this work, we propose a point cloud-based method for consistent 3D scene stylization. To synthesize novel views that 1) match *arbitrary* style images and 2) render images with consistent appearance across different views, the core idea is to operate on the 3D scene representation, *i.e.,* point cloud, of the target scene. Given a set of input images of the target scene, we first construct the point cloud by back-projecting the image features to the 3D space according to the pre-computed 3D proxy geometry. To transfer the style of the holistic 3D scene, we develop a point cloud transformation module. Specifically, we use a series of point cloud aggregation modules to gather the style information of the

3D scene. We then modulate the features in the point cloud with a linear transformation matrix [116] computed according to the style information of the point cloud and reference image. Finally, we project the transformed features from the point cloud to the 2D space to obtain the novel view synthesis results. Since our method synthesizes novel view images from the same stylized point cloud, the rendered results not only demonstrate the desired style, but also are consistent across different viewpoints.

We evaluate the proposed 3D scene stylization method through extensive qualitative and quantitative studies. The experiments are conducted on two diverse datasets of real-world scenes: Tanks and Temples [108] and FVS [165]. We conduct a user preference study to evaluate the stylization quality, *i.e.,* whether the novel view synthesis results match the style of the reference image. In addition, we use the Learned Perceptual Image Patch Similarity (LPIPS) [254] metric to measure the consistency of the results synthesized across different novel views.

We make the following contributions in this work:

- We propose a point cloud-based framework for the 3D scene stylization task.
- We design a point cloud transformation module that learns to transfer the style from an arbitrary 2D reference image to the point cloud of a 3D scene.
- We validate that our method produces high-quality and consistent stylized novel view synthesis results on the Tanks and Temples as well as FVS datasets.

2.2 Related Work

2.2.1 Novel View Synthesis

Given a set of images for a scene, novel view synthesis aims to generate high-quality images at arbitrary viewpoints. It can be categorized by the number of input images that cover the scene. One line of work takes as input a single image or stereo images. These methods use multi-plane images [182, 187, 211, 263], layer depth image [111, 174], or

point cloud [146, 210] representations to synthesize images at novel views near the input views, *e.g.,* 3D photo. To enable the image synthesis at *arbitrary* novel views, several recent frameworks take hundreds of input images of a scene as the input. These frameworks leverage different 3D representations to accomplish the task. Image-based rendering approaches [165, 166] compute 3D proxy geometry of the scene, and generate images by warping the input frames to the desired novel views. Neural radiance field schemes [128, 141, 240, 249] use multi-layer perceptrons to implicitly encode the scene for novel view synthesis. Point cloud-based methods [140, 2] solve different optimization problems to construct the point cloud for a specific 3D scene. Different from these frameworks, our goal is to generate *stylized* novel view images of the 3D scene. As shown in Figure 2.2, while existing algorithms can be used for the 3D scene stylization task, they fail to generate high-quality novel view synthesis results with the desired style.

2.2.2 Image and Video Stylization

Image stylization [56] aims to transfer the style of a reference image to the single input image. Existing methods [24, 98, 118, 173, 189] are designed based on feed-forward networks for transferring a set of *pre-defined* styles. For *arbitrary* image style transfer, Huang and Belongie [91] use first-order statistics to encode the style information, and transform the image style via the AdaIN normalization layers. The WCT [119] approach uses whitening and coloring transformation to match the second-order statistics of the input image to those of the reference image. In addition, the LST [116] scheme leverages the convolutional neural networks to reduce the computational cost of solving the transformation matrix in the WCT method for real-time universal style transfer. Most recently, the TPFR method [184] proposes a regularization layer to facilitate the generalization of image stylization models.

Video stylization aims to transfer the style of a reference image to a sequence of video frames. To address the temporal flickering issue produced by the image stylization ap-

proaches, numerous approaches [23, 31, 54, 69, 86] incorporate optical flow modules to train feed-forward networks for transferring a particular style to the videos. Several recent frameworks [43, 55, 202] enable the video style transfer to *arbitrary* styles. Although significant advances have been made, existing methods are designed specifically for transferring the style of 2D images or video sequences. As shown in Figure 2.2, simply applying these schemes for the 3D scene stylization task leads to problematic results, such as blurry or short/long-range inconsistent images across different novel views.

Several efforts have been made to perform the stylization in 3D space. However, these approaches are only applicable to single objects [102], narrow-baseline stereo images [25, 63], or light field images [73]. In contrast, our method stylizes complex 3D scenes, and produces consistent results at arbitrary viewpoints.

2.2.3 Deep Neural Networks for Point Clouds

Various deep neural network (DNN)-based models [107, 114, 115, 117, 157, 158, 215, 218, 257] that take point clouds as input are widely studied for vision recognition tasks including 3D semantic segmentation [5], 3D shape classification or normal estimation [216], and 3D object part segmentation [236]. Recently, Mallya *et al.* [137] proposes a point cloud colorization approach for the video-to-video synthesis task. In this work, we propose a DNN-based point cloud transformation model for the 3D scene stylization task. We note that the PSNet [19] model aims to transfer the style of the point cloud. Nevertheless, there are two issues for the PSNet method to be applied to the 3D scene stylization task. First, it does not support synthesizing high-quality stylized images at novel views, which makes the PSNet framework limited for real-world (*e.g.,* AR) applications. Second, since the PSNet scheme requires the optimization process for each specific scene, it is time-consuming, and fails to handle large-scale scenes in the real-world with more than 60M points, such as those in the Tanks and Temples dataset [108]. In contrast, we propose a feed-forward point cloud model that is efficient, capable of handling large-scale 3D scenes,

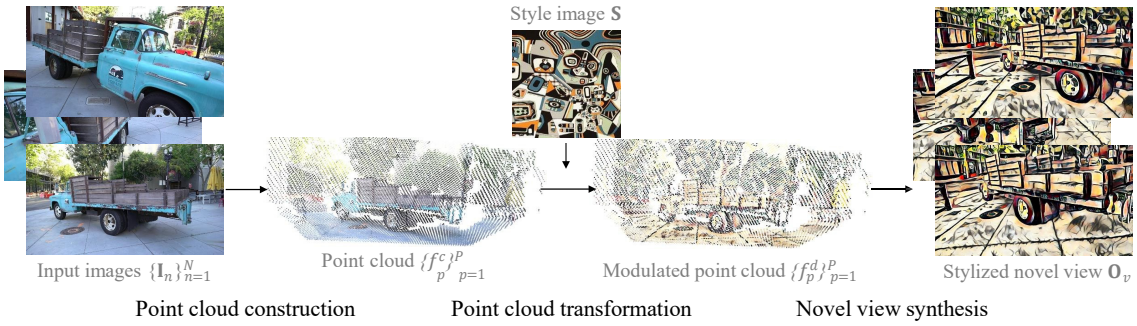

Figure 2.3: **Algorithmic overview.** The proposed method consists of three steps: 1) constructing the 3D point cloud from the set of input images $\{\mathbf{I}_n\}_{n=1}^N$, 2) transforming the point cloud according to the reference image $\mathbf{S}$ with the desired style, and 3) synthesizing the stylized image $\mathbf{O}_v$ at arbitrary novel view v. The coloring of the point clouds is for visualization purposes only. In our approach, the point clouds store the features rather than RGB values.

and generating images with arbitrary styles at various novel views.

2.3 Methodology

We present the overview of the proposed 3D scene stylization framework in Figure 2.3. Given a set of N input images $\{\mathbf{I}_n\}_{n=1}^N$ of a static scene, and a reference image $\mathbf{S}$ with the desired style, our goal is to synthesize the image $\mathbf{O}_v$ at the novel view v with the camera pose $(\mathbf{R}_v, t_v)$ and intrinsic $\mathbf{K}_v$. Specifically, the generated novel view image $\mathbf{O}_v$ needs to 1) match the style of the reference image $\mathbf{S}$ and 2) be consistent for different viewpoints v. To handle such (especially the consistency) requirements, our core idea is to 1) construct a single 3D representation, *i.e.,* point cloud, for the holistic scene, and 2) transform the representation to produce not only stylized but also consistent novel view synthesis results. The proposed approach consists of three steps: point cloud creation,

point cloud transformation, and novel view synthesis, described in the following sections.

2.3.1 Point Cloud Construction

Pre-processing. Our method leverages camera pose and proxy geometry to construct the 3D point cloud. Given the input images $\{\mathbf{I}_n\}_{n=1}^N$, we first use a structure-from-motion algorithm [171] to estimate the camera poses $\{\mathbf{R}_n, t_n\}_{n=1}^N$ and intrinsic parameters $\{\mathbf{K_n}\}_{n=1}^N$. For each image $\mathbf{I}_n$, we use the COLMAP [171, 172] and Delaunay-based reconstruction [95, 110] schemes to obtain the depth map $\mathbf{D}_n$ that can appropriately back-project the points from the image plane to the 3D space.

Feature extraction and back-projection. Since our goal is to transform the point cloud representation for the 3D scene stylization purpose, we need the point cloud representation to encode the style information. Therefore, we use the VGG-19 model [177] pre-trained on the ImageNet [42] dataset to extract the relu3_1 feature maps $\{\mathbf{F}_n^c\}_{n=1}^N$ of the input images $\{\mathbf{I}_n\}_{n=1}^N$. The width and height of each feature map is H and W. According to the depth map $\{\mathbf{D}_n\}_{n=1}^N$, we back-project all the points in each feature map to build the 3D point cloud $\{f_p^c\}_{p=1}^P$, where $P = NHW$ is the total number of points in the constructed point cloud.

2.3.2 Point Cloud Transformation

We model the 3D scene stylization process as a linear transformation [116] between the constructed and stylized point clouds. Intuitively, the goal is to match the covariance statistics of the stylized point clouds and those of the reference image $\mathbf{S}$. To achieve this, we use the pre-trained VGG-19 network to extract the relu3_1 feature map from the reference image $\mathbf{S}$ as the style feature map $\mathbf{F}^s$. Given the constructed point cloud $\{f_p^c\}_{p=1}^P$, we use a predicted linear transformation matrix $\mathbf{T}$ to compute the modulated point cloud $\{f_p^d\}_{p=1}^P$, namely

$$f_p^d = \mathbf{T}(f_p^c - \bar{f}^c) + \bar{f}^s \quad \forall p \in [1, \cdots, P], \tag{2.1}$$

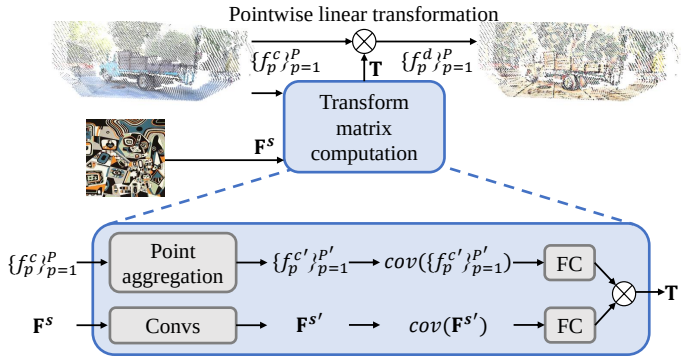

Figure 2.4: **Point cloud transformation.** We model the 3D scene stylization process as the linear transformation between the constructed and stylized point clouds. Specifically, the constructed point cloud is modulated using the predicted linear transformation matrix $\mathbf{T}$, as described in (2.1). We use a series of point cloud aggregation modules to gather the point cloud information, and the convolution layers to process the reference image feature $\mathbf{F}^s$ to compute the matrix $\mathbf{T}$.

where $\bar{f}^c$ is the mean of the features in the point cloud $\{f_p^c\}_{p=1}^P$, and $\bar{f}^s$ is the mean of the style feature map $\mathbf{F}^s$.

Linear transformation matrix T. The transformation matrix $\mathbf{T}$ is computed from the style feature map $\mathbf{F}^s$ and constructed point cloud $\{f_p^c\}_{p=1}^P$. As shown in Figure 2.4, we adopt the strategy similar to the LST [116] method that uses the convolution layers, covariance computation, and fully-connected layers to compute the matrix $\mathbf{T}^s$ from the style feature map $\mathbf{F}^s$. On the other hand, we develop a series of point cloud aggregation modules to process the point cloud $\{f_p^c\}_{p=1}^P$, and use the covariance computation followed by the fully-connected layers to calculate the matrix $\mathbf{T}^c$. Finally, we obtain the transformation matrix $\mathbf{T} = \mathbf{T}^s\mathbf{T}^c$.

Point cloud aggregation. It is challenging to gather the information contained in the constructed point cloud $\{f_p^c\}_{p=1}^P$ due to the sparsity and non-uniformity. We note that the constructed point cloud is non-uniform if the input images cover a particular region of the

3D scene. In this work, we leverage the set abstraction [158] concept to aggregate the point cloud. The input $\{f_p^c\}_{p=1}^P$ to a point cloud aggregation module is a set of P points with feature dimension c, and the output $\{f_p^{c'}\}_{p=1}^{P'}$ is a set of P' points with dimension c'. We first sample a subset of P' points $\{f_p^c\}_{p=1}^{P'}$ using the iterative farthest point sampling algorithm [64, 142]. Viewing the sampled points as the centroids in the 3D space, we use a radius parameter r to find the nearby points to form a point group. By using the MLP layers and the max pooling operator to map each point group to a vector, we obtain the aggregated point cloud $\{f_p^{c'}\}_{p=1}^{P'}$. The output $\{f_p^{c'}\}_{p=1}^{P'}$ is then used as the input for the next module. We use three point cloud aggregation modules sequentially in our pipeline.

2.3.3 Novel View Synthesis and Model Training

We aim to synthesize stylized image $\mathbf{O}_v$ at an arbitrary novel view v. Given the target camera pose $(\mathbf{R}_v, t_v)$ and intrinsic $\mathbf{K}_v$, we use Pytorch3D [163, 210] to render the transformed 2D feature map $\mathbf{F}_v^d$. We then use a decoder network to generate the stylized novel view image $\mathbf{O}_v$ from the 2D feature map $\mathbf{F}_v^d$.

Model training. We keep the pre-trained VGG-19 feature extractor fixed during the whole training phase. We first train the decoder network to perform the non-stylized novel view synthesis. Since the ground-truth (non-stylized) novel view image is available in the training sets, we use the $\ell 1$ reconstruction loss to optimize the decoder network. We then keep the decoder network fixed, and train the proposed point cloud transformation module with the following loss functions:

- **Content loss** $\mathcal{L}_c$ ensures the preservation of the content information by measuring the distance between the pre-trained VGG-19 features of the generated stylized image $\mathbf{O}_v$ and the ground-truth (non-stylized) image $\mathbf{I}_v$.
- **Style loss** $\mathcal{L}_s$ encourages the synthesized image $\mathbf{O}_v$ to match the style of the reference image $\mathbf{S}$. Similar to recent style transfer approaches [98, 116], we extract the features at different layers of the pre-trained VGG-19 model, and compute the gram

matrix differences.

The overall loss function for training the point cloud transformation module is

$$\mathcal{L} = \mathcal{L}_c(\mathbf{O}_v, \mathbf{I}_v) + \lambda \mathcal{L}_s(\mathbf{O}_v, \mathbf{S}), \tag{2.2}$$

where λ controls the importance of each loss term.

2.4 Experimental Results

We conduct extensive experiments on two real-world datasets to validate the efficacy of the proposed 3D scene stylization model.

Datasets. We use the Tanks and Temples [108] dataset for quantitative evaluation. Similar to the setting in FVS [165], we use 17 out of the 21 scenes for the training. The four remaining scenes (Truck, Train, M60 and Playground) are used for testing. We also present qualitative results on the FVS [165] dataset, which consists of 6 scenes: Bike, Flowers, Pirate, Digger, Sandbox and Soccertable. Note that both datasets are collected by *handheld* cameras in un-constraint motions.

Evaluated methods. As the 3D scene stylization task is a relatively new problem, we evaluate our method against alternative approaches built upon the state-of-the-art novel view synthesis NeRF++ [249], SVS [166], and image/video stylization schemes:

- **Image stylization → novel view synthesis:** We first use image stylization schemes LST [116] or TPFR [184] to transfer the style to the input images $\{\mathbf{I}_n\}_{n=1}^N$, then perform novel view synthesis.
- **Novel view synthesis → image stylization:** We apply image stylization to the novel view synthesis results.
- **Novel view synthesis → video stylization:** We use a series of novel view synthesis results to create a *video*, then apply video stylization methods Compound [202], FMVST [55], or MCC [43].

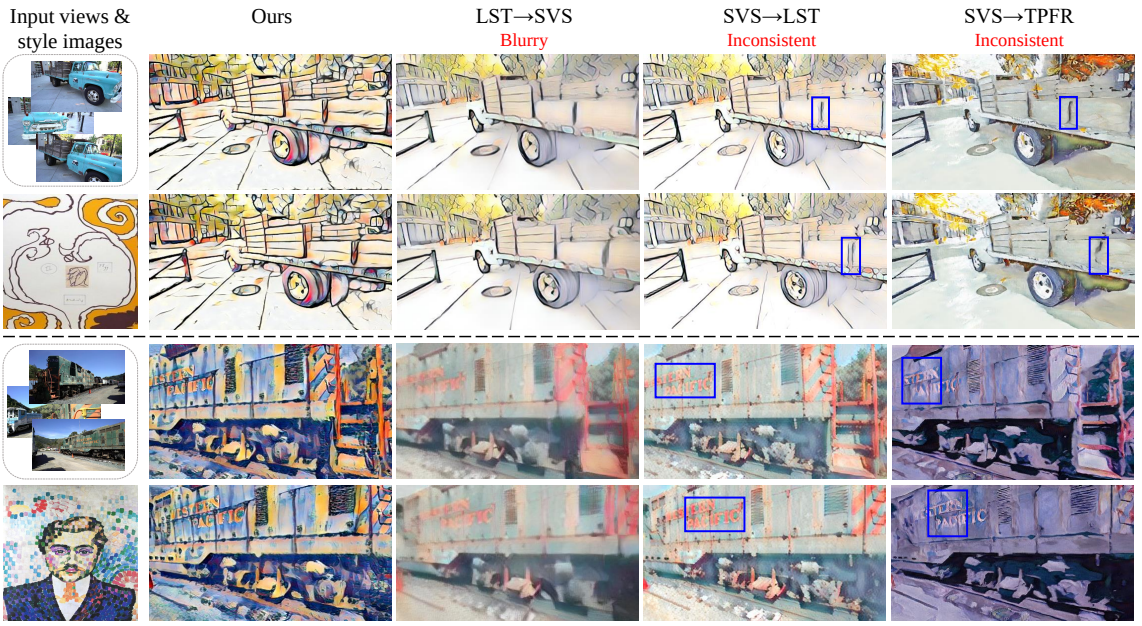

Figure 2.5: **Visual comparisons to image stylization-based approaches.** We compare the stylized novel view images generated by the three image stylization alternative schemes and our model on Tanks and Temples dataset [108].

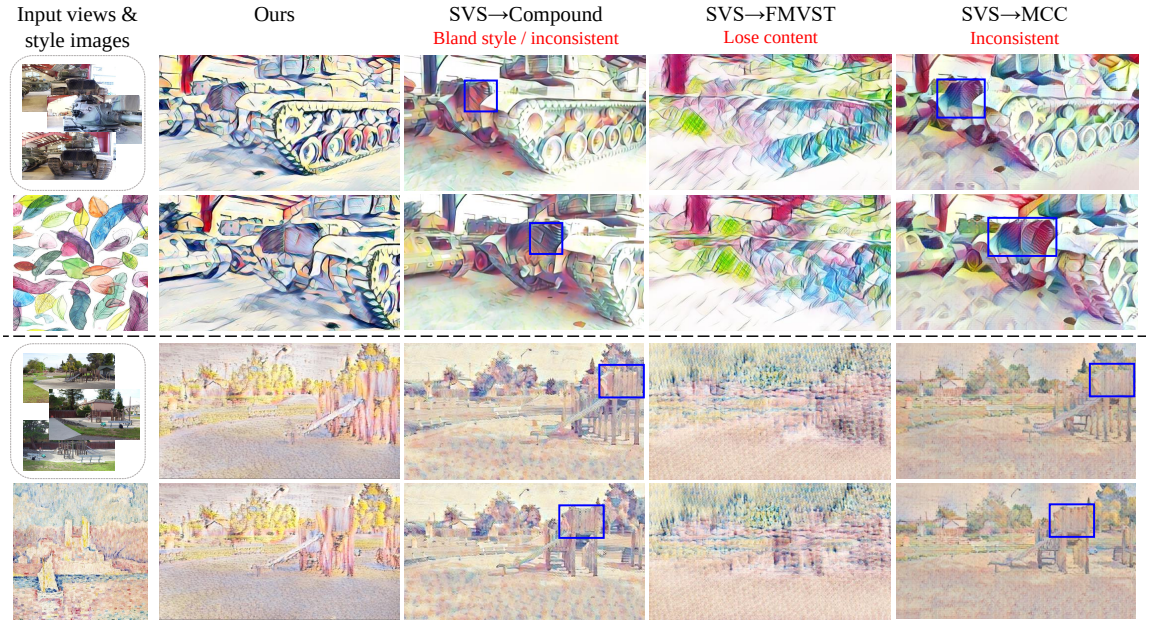

Figure 2.6: **Visual comparisons to video stylization-based approaches.** We compare the stylized novel view images generated by the three video stylization alternative schemes and our model on Tanks and Temples dataset [108].

2.4.1 Qualitative Results

Image stylization. Figure 2.5 presents the qualitative comparison between the stylized novel view images generated by the three image stylization alternative schemes and the proposed method. Since the images are stylized independently without considering the consistency issue across different viewpoints, we observe two issues in the image stylization-based methods. First, LST → SVS generally produces blurry novel view images. Since the stylized input images are not consistent, the novel view synthesis approach tends to *blend* such inconsistency, which leads to blurry results. Second, the novel view synthesis results are not consistent if we operate in the reverse order, *i.e.,* SVS → LST. We highlight the inconsistency using yellow boxes in Figure 2.5. Note that we observe the same problem if we replace SVS with NeRF++.

Video stylization. We qualitatively evaluate the results by the proposed method and three video stylization alternative approaches in Figure 2.6. Specifically, we create the videos using a series of novel view synthesis results. All the alternative approaches generate inconsistent results between two relatively far-away viewpoints since the video stylization methods only guarantee *short-term* consistency in the video. Although SVS→Compound generates less inconsistent results, the style of the novel view images is bland and not aligned with that of the reference image. On the other hand, SVS→FMVST creates images that better match the desired style, but fails to preserve the content of the original scene.

In contrast to the image and video stylization alternative approaches, our method 1) generates sharp novel view images with correct scene contents and the desired style, and 2) guarantees the short/long-range consistency. Furthermore, we demonstrate the generalization of the proposed framework in Figure 2.7 , where we use the model trained on the Tanks and Temples dataset to perform the 3D scene stylization task on the FVS dataset.

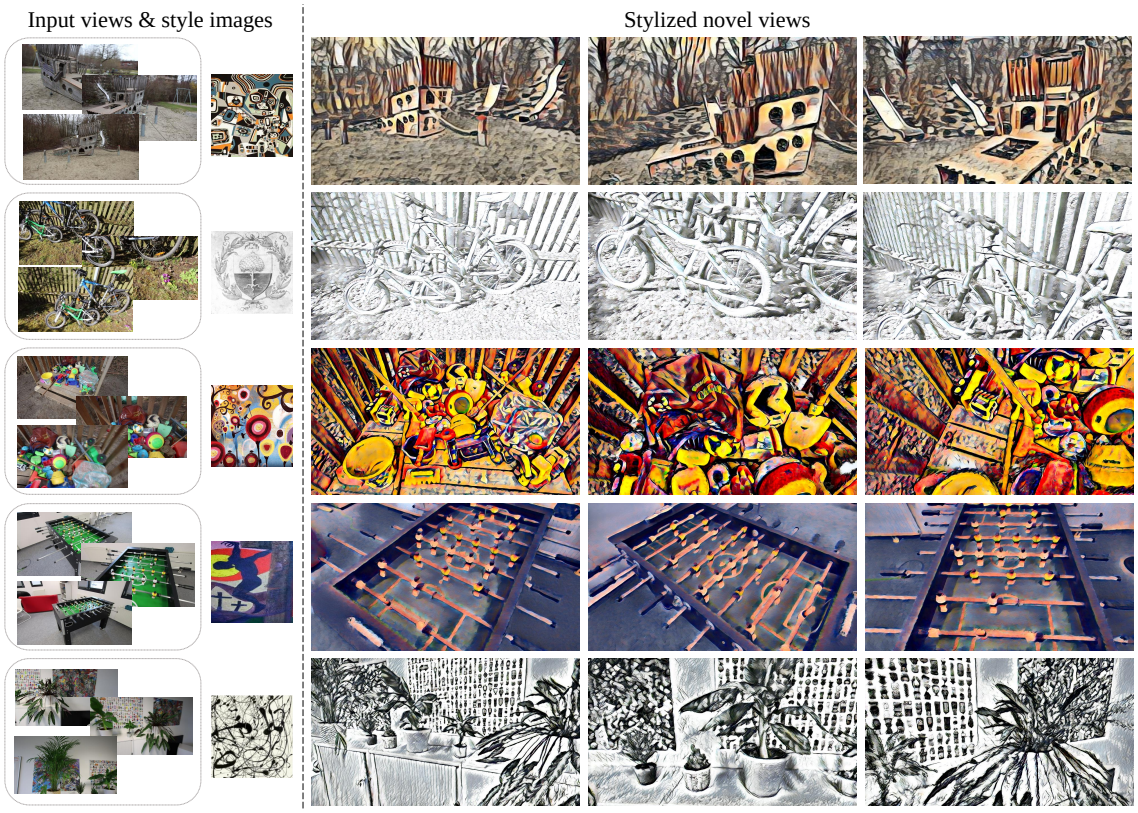

Figure 2.7: **Qualitative results on the FVS dataset.** We demonstrate the generalization of the proposed approach by training on the Tanks and Temples dataset, then testing on the FVS dataset.

2.4.2 Quantitative Results

Stylization quality. We conduct a user study to understand the user preference between the proposed and the alternative approaches. For each testing scene in the Tanks and Temples dataset, we create a video using a series of stylized novel view synthesis results. By presenting two videos generated by different methods for the same scene, we ask the participants to select the one that (1) has more consistent contents across different video frames (e.g., less flickering), and (2) better matches the style of the reference image. As the results shown in Figure 2.8, the synthesized images by the proposed method are

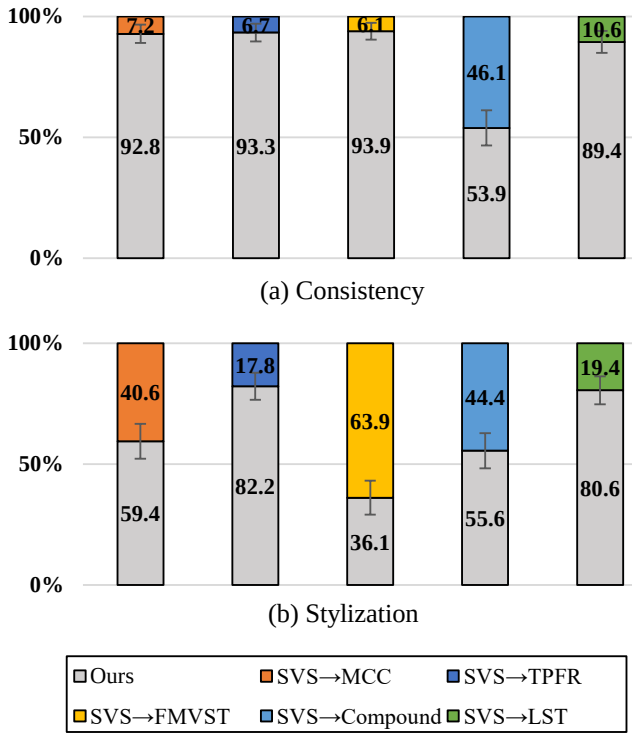

Figure 2.8: **User preference study.** We conduct a user study and ask subjects to select the results that (a) have more consistent contents across different video frames (*e.g.,* less flickering), (b) better match the style of the example image. The number indicates the percentage of preference.

consistent and close to the reference style. We observe that the users slightly prefer the style generated by SVS→FMVST. However, as illustrated in Section 2.4.1 and Figure 2.6, SVS→FMVST fails to preserve the content of the original scene.

Short-range consistency. We use the warped LPIPS metric [254] to measure the consistency of the results across different viewpoints. Given a stylized image at a novel view v, we warp the results generated at another novel view v' to the view v according to the 3D proxy geometry described in Section 2.3.1. We then compute the score by

$$E_{\text{warp}}(\mathbf{O}_v, \mathbf{O}'_v) = \text{LPIPS}(\mathbf{O}_v, W(\mathbf{O}'_v), \mathbf{M}_{v'v}), \tag{2.3}$$

Table 2.1: **Short-range consistency.** We compare the long-range consistency using the warping error (↓) between the viewpoints of $(t-1)$-th and t-th testing video frames in the Tanks and Temples dataset [108]. We report the average errors of 15 diverse styles. The best performance is in **bold** and the second best is underscored.

Method	Truck	Playground	Train	M60	Average
NeRF++→LST	0.215	0.168	0.250	0.274	0.231
SVS→LST	0.192	0.159	0.220	0.241	0.206
NeRF++→TPFR	0.216	0.214	0.299	0.279	0.258
SVS→TPFR	0.235	0.237	0.291	0.276	0.264
NeRF++→Compound	0.188	0.169	0.229	0.208	0.202
SVS→Compound	**0.166**	**0.156**	0.199	**0.160**	0.172
NeRF++→FMVST	0.342	0.300	0.405	0.348	0.354
SVS→FMVST	0.343	0.304	0.412	0.337	0.354
NeRF++→MCC	0.250	0.201	0.269	0.255	0.246
SVS→MCC	0.242	0.198	0.260	0.224	0.232
Ours	0.184	0.158	**0.170**	0.172	**0.170**

where W is the warping function and $\mathbf{M}_{v'v}$ is the mask of valid pixels warped from the views v' to v. Note that we only use the values of valid pixels in the mask for the "spatial average" operation in [254]. For each of the testing scenes in the Tanks and Temples dataset, we use 15 style images [55] to compute the average warping error.

We first present the short-range consistency comparison in Table 2.1. In this experiment, we use the nearby view for a specific novel view to compute the warping error.[1] In general, the image stylization alternative methods produce short-range inconsistent results as they process each novel view independently. In contrast, the proposed method performs

[1] We use the viewpoints of $(t-1)$-th and t-th testing video frames as the views v' and v, respectively.

Table 2.2: **Long-range consistency.** We compare the long-range consistency using the warping error (↓) between the viewpoints of $(t-7)$-th and t-th testing video frames in the Tanks and Temples dataset [108]. We report the average errors of 15 diverse styles. The best performance is in **bold** and the second best is underscored.

Method	Truck	Playground	Train	M60	Average
NeRF++→LST	0.570	0.349	0.520	0.639	0.521
SVS→LST	0.567	**0.327**	0.470	0.603	0.489
NeRF++→TPFR	0.579	0.436	0.503	0.655	0.541
SVS→TPFR	0.605	0.430	0.470	0.581	0.513
NeRF++→Compound	0.586	0.398	0.477	0.557	0.498
SVS→Compound	0.573	0.388	0.422	0.460	0.449
NeRF++→FMVST	0.742	0.525	0.636	0.695	0.644
SVS→FMVST	0.732	0.519	0.620	0.662	0.626
NeRF++→MCC	0.691	0.450	0.535	0.646	0.571
SVS→MCC	0.693	0.447	0.516	0.584	0.548
Ours	**0.559**	0.337	**0.412**	**0.458**	**0.431**

comparably against the video stylization-based approach SVS→Compound that considers the short-term consistency in videos. Nevertheless, SVS→Compound synthesizes bland styles that do not match the desired styles, as the results demonstrated in Figure 2.6 and Figure 2.8.

Long-range consistency. We also consider the long-range consistency issue in our experiments. In this experiment, we compute the warping error between the results of two (relatively) far-away views.[2] As demonstrated in Table 2.2, the proposed method performs favorably against the alternative approaches. Despite the capability of ensuring short-range

[2]We use the viewpoints of $(t-7)$-th and t-th testing video frames as the views v' and v, respectively.

Figure 2.9: **Ablation study on the number of point cloud aggregation modules.** We compare the visual results of using 0/1/**3**/5 modules. We empirically decide to use 3 modules for better visual quality.

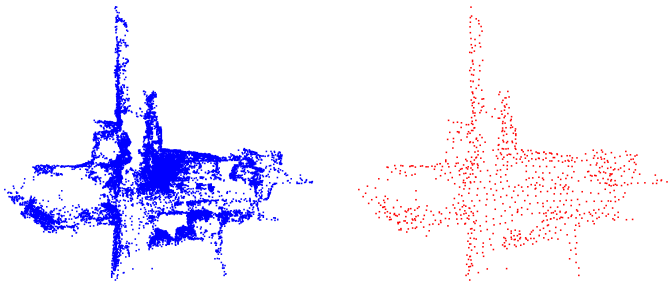

Figure 2.10: **Role of point aggregation.** We visualize the point distribution before and after the point aggregation. Our point aggregation module obtains a more uniform-distributed point set to fairly estimate the transformation matrix **T** that achieves better 3D scene stylization results.

consistency, video stylization-based schemes fail to maintain the long-range consistency.

Number of point cloud aggregation modules. We conduct an ablation study to decide the number of point cloud aggregation modules described in Section 2.3.2. The results are presented in Figure 2.9. We empirically choose to use three modules for better visual quality. Moreover, we visualize the point distributions before and after the aggregation in

Figure 2.10 to understand the role of the aggregation module. The point density before the aggregation is higher around the regions of the 3D scene where more input images cover. As a result, the prediction of the transformation matrix **T** is dominated by such regions, which leads to low-quality stylization results (2nd column in Figure 2.9). By using the point cloud aggregation modules, we obtain a more uniform-distributed point set that fairly estimates the matrix **T** for the 3D scene stylization task.

2.5 Conclusions

In this work, we introduce a 3D scene stylization problem that aims to modify the style of the 3D scene and synthesize images at arbitrary novel views. We construct a single 3D representation, *i.e.,* point cloud, for the holistic scene, and design a point cloud transformation module to transfer the style of the reference image to the 3D representation. Qualitative and quantitative evaluations validate that our method synthesizes images that 1) contain the desired style and 2) are consistent across various novel views.

Chapter 3

Fine-grained Controllable Video Generation via Object Appearance and Context

While text-to-video generation shows state-of-the-art results, fine-grained output control remains challenging for users relying solely on natural language prompts. In this work, we present FACTOR for fine-grained controllable video generation. FACTOR provides an intuitive interface where users can manipulate the trajectory and appearance of individual objects in conjunction with a text prompt. We propose a unified framework to integrate these control signals into an existing text-to-video model. Our approach involves a multimodal condition module with a joint encoder, control-attention layers, and an appearance augmentation mechanism. This design enables FACTOR to generate videos that closely align with detailed user specifications. Extensive experiments on standard benchmarks and user-provided inputs demonstrate a notable improvement in controllability by FACTOR over competitive baselines.

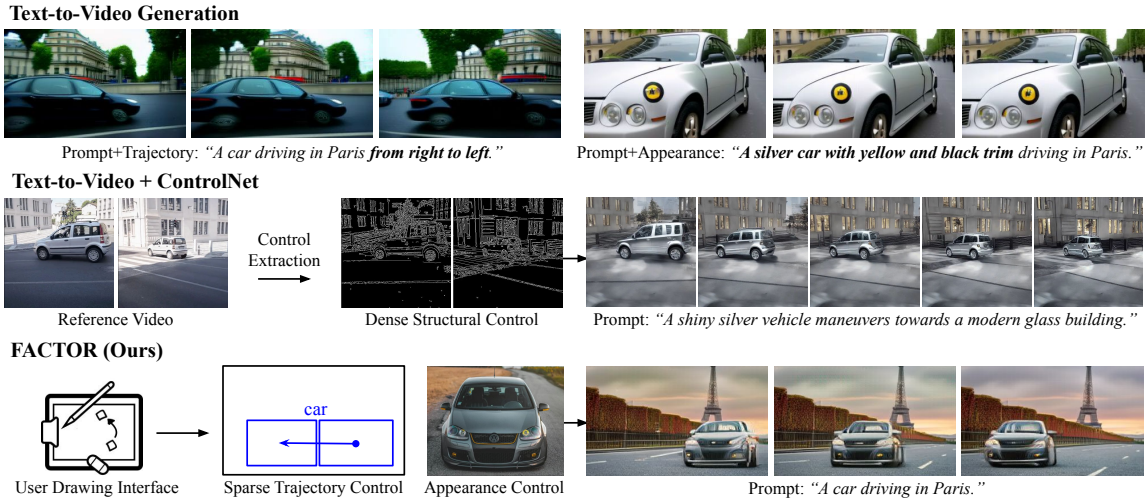

Figure 3.1: **Text-to-video generation** [192] has limited controllable ability through user-provided prompts. **Text-to-video+ControlNet** [255] requires dense control signals extracted from a reference video. To enable controllability through user-friendly inputs, **FACTOR** controls precise subject movements through hand-drawn trajectories and their visual appearance using reference examples.

3.1 Introduction

Recent text-to-video models [83, 213, 212, 80, 178, 14, 104, 198, 58, 192, 262] allow users to translate their creative ideas into video content easily. However, achieving precise control over the composition of these videos remains a challenge. Specifying detailed object movements and appearances through text alone is often difficult and requires iterative revisions. Even when users provide additional descriptions like "from right to left" or "yellow and black trim," models can still struggle to generate the desired output, as shown in Figure 3.1 top. It is highly desirable to have a user-friendly system that enables fine-grained control over the appearance and motion of individual objects.

Recent advances such as ControlNet [250] offer potential solutions, integrating structural controls (e.g., optical flow, depth) into text-to-video generation [214, 104, 255, 28].

Similar techniques have been developed in video editing [201, 129, 159, 59], where natural language guides the manipulation of a video's content and style. While these methods offer promising results, they rely on dense control inputs for every frame – typically extracted from a reference video (Figure 3.1 middle). This limits the generation of novel videos with structures different from the reference and makes the process impractical for manual user input.

On the appearance control front, methods like DreamBooth and its extensions [104, 261, 214, 168, 68, 14] allow for subject customization in video generation. However, these methods are limited to adjusting the appearance of individual subjects and cannot manipulate their appearances and locations simultaneously. VideoComposer [204] takes a step forward by integrating text with spatial and temporal controls. However, it requires dense motion information from a reference video and applies these controls globally to the entire video. A method for generating videos composed of multiple entities, with precise control over appearance and motion, remains an exciting and open problem.

In this work, we introduce FACTOR, a framework for fine-grained controllable video generation. We demonstrate that for precise control, users should be able to easily manipulate individual entities when generating videos with multiple objects. FACTOR prioritizes user-friendliness by accepting sparse and intuitive inputs: a text prompt, user-drawn bounding boxes, and user-provided reference images. To this end, we build FACTOR upon an off-the-shelf text-to-video model [192]. We use a joint encoder to integrate the multimodal input control signals and insert control-attention layers into the model's transformer blocks. Training involves only these newly inserted layers, with the pre-trained model's weights frozen. This design ensures high-quality video generation while introducing object-level control mechanisms. We further enhance the diversity of reference appearance images through color and geometric transformations to improve appearance control. Experiments on the MSR-VTT benchmark [222] demonstrate that FACTOR significantly improves generation quality, trajectory, and appearance control compared to the alternative approaches. A user study further validates the effectiveness of the proposed

method.

The main contributions of this work are:

- We target the new form of *fine-grained controllable video generation* that aims to synthesize videos via multimodal context (text, appearance, and trajectory) of individual objects from easy-to-give user inputs.
- We propose a unified framework to achieve multimodal control within a single training process. This is achieved by injecting control signals with control-attention and appearance augmentation.
- We validate that our method offers fine controllability of multiple objects compared to existing works and shows the additional benefit of creating interactions, which is challenging for existing text-to-video models.

3.2 Related Work

3.2.1 Text-to-video Generation

Text-to-video models have made impressive progress. Token-based methods [212, 213, 83] utilize an auto-regressive model to predict videos in the latent space. Diffusion-based models [178, 82, 80] extend the 2D diffusion model [167] to generate videos [104, 198, 262, 14, 58, 247, 205, 26, 13, 198, 58, 132, 203, 27, 70, 68, 104, 200, 220] by incorporating temporal layers. Although these models produce promising results, they have limited control over the generated videos since text prompts cannot accurately convey precise control signals. In this work, we develop a fine-grained video generation framework that controls the location and appearance of objects. We adopt a generative transformer model [192] as our base model, and our approach can be adapted to diffusion-based models.

3.2.2 Controllable Text-to-image Generation

Various models have been proposed to enhance the controllability of text-to-image models, particularly in terms of structure and appearance. ControlNet [250, 145, 260, 223, 234, 121], structure-guided generation [7, 225, 232, 246, 72], and training-free approaches [11, 242, 217, 106, 47] incorporate various spatial control signals such as edges, depth, segmentation, and human pose into pre-trained diffusion models. However, the dense structure inputs these models require can be challenging for users to provide when generating videos. Fine-tuning on reference images [52, 168, 113] and encoder-based methods [53, 208] are employed to control the appearance of subjects. However, these methods do not control the locations of subjects. In contrast, we propose a fine-tuning-free method to jointly control the appearance and location of subjects for video generation. ControlNet is extended to control global appearance and style [223, 234], while FACTOR focuses on individual subject appearance control.

3.2.3 Controllable Text-to-video Generation

Structural control for video generation [204, 214, 104, 133, 255, 219, 28, 124, 33, 35, 48] focuses on producing temporally consistent videos using a temporal attention layer integrated with ControlNet. Video editing approaches [201, 231, 21, 175, 129, 159, 85, 195, 59, 220, 150, 89, 160] aim to alter the appearance and style of videos based on text prompts. They typically require dense structural inputs extracted from reference videos. In contrast, our work focuses on controllable T2V generation using sparse and user-friendly control signals. Video customization through fine-tuning [214, 104, 78, 261] still relies on dense structural inputs and often lacks control over subject location [68, 14]. VideoComposer [204] employs a reference image for global appearance control and a dense motion sequence for spatial control. In contrast, FACTOR focuses on controlling the generation of individual subjects using sparse inputs.

3.3 Method

Our main goal is to develop a user-friendly system allowing entity-level motion and appearance control for video generation. To this end, we enable users to provide multimodal inputs: a text prompt, user-drawn bounding box trajectories, and user-provided reference images. By breaking down the video into manageable entities, users can create the desired video content by adjusting the properties of individual elements. Our intuitive control interface facilitates users to gradually design their videos by adding entities with reference images in one frame and adjusting their positions across multiple frames. We first briefly review the base T2V model and provide an overview of our approach. Then, we introduce each component in detail.

3.3.1 Preliminaries: Text-to-video Generation

Our base T2V model [192] consists of an encoder-decoder model that encodes the video into discrete tokens and a bidirectional transformer model that predicts the video tokens conditioned on the embedding of text prompts. During training, the tokens are replaced with a special token `[MASK]`, and the transformer model is optimized to predict the tokens at `[MASK]` locations based on the text embedding. We minimize the negative log-likelihood of predicting the masked tokens $v_t, t \in M$, where M denotes the subset of video tokens that are masked, v denotes the video token sequence, and v_M denotes the masked version of v.

$$\mathcal{L} = - \mathop{\mathbb{E}}_{v \in \mathcal{D}} \sum_{t \in M} \log p(v_t | v_M, p). \tag{3.1}$$

The transformer model contains a series of attention layers that condition the video token prediction on the text embeddings. At inference time, all tokens are replaced with `[MASK]`, and the model iteratively predicts the tokens. We use a token-based generative transformer model for its fast inference and inherent flexibility in modeling various control signals directly within the same architecture.

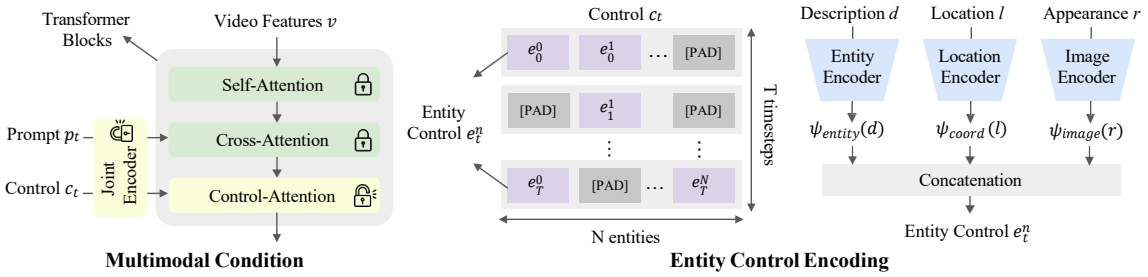

Figure 3.2: **Overview.** a) **Multimodal condition**: a joint encoder is learned to encode the prompt and control to capture their interaction. Control-attention layers are inserted into the transformer blocks of the text-to-video model to incorporate multimodal control signals. Only the inserted layers are optimized to generate videos satisfying the fine-grained control. b) **Entity control encoding**: given T time steps, the embedding of control c_t is formed by the control for N entities, e_t^n, where padding tokens replace the embedding of the non-existing entity. The control for entity n at time t is formed by embeddings of the description, location, and reference appearance.

3.3.2 Overview

We present a novel approach for fine-grained video generation that allows for precise control of individual objects using user-friendly inputs. The problem is defined as follows: given a text prompt p and fine-grained control c as inputs, our model aims to generate a video that satisfies both input conditions. Specifically, users provide multimodal control c by 1) describing the desired entities in the video, 2) drawing their trajectories, and 3) providing a reference appearance image for each entity. This pipeline helps users create videos intuitively. The input control signal is given at T timesteps. Assuming there are N entities in the video, we define our control as $c = \{c_t\}_{t=1}^{T}$. At a single timestep t, the embeddings c_t are formed by a sequence of entity controls, i.e., the embeddings of the N entities $c_t = \{e_t^n\}_{n=1}^{N}$. The embeddings e_t^n encode the desired conditions to generate the entity indexed by n at time t, as shown in Figure 3.2.

3.3.3 Multimodal Condition

We introduce an effective multimodal condition method for generating videos that satisfy fine control of different modalities within a single training pipeline. Our method includes a joint encoder and a control-attention module.

Joint encoder. Existing text-to-image models (*e.g.,* , [250, 121]) use a separate encoder for each input condition and directly input the independent embeddings into the model, which is less effective (see Table 3.1 GLIGEN [121], VideoComposer [204]) as the encoder needs to be trained for new conditions and the interaction between the controls is ignored. In contrast, we use a joint encoder that simultaneously encodes the text prompt p and the fine-grained control c of different modalities within the same encoder. This design facilitates the learning of interactions between multimodal inputs, captured in the contextualized embeddings.

Control-attention. The next task is to incorporate the sequence of control embeddings into the base T2V model. A natural design choice is to include the control embeddings in the original cross-attention module. However, this poses difficulties in generating videos aligned with the new control because the cross-attention layers are optimized for the text prompt, weakening their flexibility for new inputs (see Table 3.1 ELITE [208]). Instead, we insert a new control-attention layer in each transformer block to accommodate additional control. During training, we freeze the weights of the pre-trained model and train the joint encoder and the new control-attention layer. This allows the T2V model to generate videos aligned with both text and the new multimodal condition. By fixing the pre-trained weights, we preserve the capability to generate high-quality videos while updating only 23% of the parameters.

Our control-attention has less computational overhead (19%) compared to gated-attention [121], which struggles to converge with the increasing length of tokens in videos. While we use a generative transformer model, our control-attention layers can be extended to other T2V models, like diffusion models, which have similar attention blocks [167, 14].

This module can also handle other control signals, such as camera poses. We achieve this by transforming the control into tokens and inputting them into our joint encoder and control-attention layers. Our unified training process conditions the model on multimodal control signals in a single pipeline, avoiding the need for multiple condition-specific methods.

3.3.4 Entity Control Encoding

Here, we discuss our entity-level fine-grained control. The entity embeddings e_t^n are constructed by encoding the context of each entity. First, the description of the entity d is given by text, e.g., a cat, and encoded into embeddings $\psi_{\text{entity}}(d)$. Second, the location of the entity l is given by the top-left and bottom-right bounding box coordinates of the entity and encoded as $\psi_{\text{coord}}(l)$. Finally, the reference appearance r of the entity is given by a single example image and encoded by a CLIP image encoder as $\psi_{\text{image}}(r)$. The embeddings e_t^n are the concatenation of the description, location, and appearance embeddings of the entity:

$$e_t^n = \text{Concat}(\psi_{\text{entity}}(d_t^n), \psi_{\text{coord}}(l_t^n), \psi_{\text{image}}(r_t^n)). \tag{3.2}$$

We replace the embeddings of e_t^n with padding embeddings when the entity of index n is missing at timestep t.

User-friendly inputs. To make our method user-friendly, we assume descriptions and appearances are fixed throughout the video, and only the location changes over time although all the conditions can be thoroughly given at T timesteps. Our method takes sparse inputs where the location of the entity is provided by simply drawing a bounding box in the first frame and dragging it to move to the location in the last frame. We obtain the locations in middle frames by linear interpolation. Our sparse input is more user-friendly compared to [204], which is only applicable to dense inputs. In addition, we randomly drop the description, the location, and the appearance embeddings of the entities 20% of the time during training. In this case, the model is trained to generate results when one or more

control signals are absent. At inference time, users can provide partial control signals as inputs (see Figure 3.11).

Data collection. In practice, very few video datasets include annotations for object trajectories and visual examples. To train our model, we employ an off-the-shelf object detector [88] and tracking algorithm [12] to extract N entities within a video clip and their locations across T timesteps. For each entity, we collect a reference visual example r by cropping the region defined by the detected bounding boxes.

3.3.5 Appearance Augmentation

Unlike image synthesis [168, 113], it is difficult to collect multiple images of the same subject as training data for videos. Here, we discuss how we create diverse reference images of the subject. During training, we sample the reference visual example of the subject from a longer video clip outside our training clip within the same video to obtain reference images with more diverse appearances. However, these samples still contain limited backgrounds and poses. To create more diverse references, we apply strong augmentation to the reference images. Specifically, we use color augmentation to create diverse backgrounds and geometric augmentation to introduce different subject poses. These operations force the model to learn the subject's appearance and exclude irrelevant information such as backgrounds and poses, thus preventing the model from overfitting the generated results to the given reference image.

To further enhance the motion of live subjects at inference time, we simulate different poses of the subject by sampling a sequence of transformed variants of the reference image using translation and cropping. Specifically, we randomly crop the reference image to an initial appearance r_0^n. Then, we define a translation vector δ to augment the initial appearance. We create a sequence of reference images with enhanced poses, $r_t^n = T_{\delta \cdot t}(r_0^n)$, where T denotes the translation, as conditions at different timesteps. This strategy greatly improves the motion dynamics of generated live subjects, such as animals (See Table 3.3). In

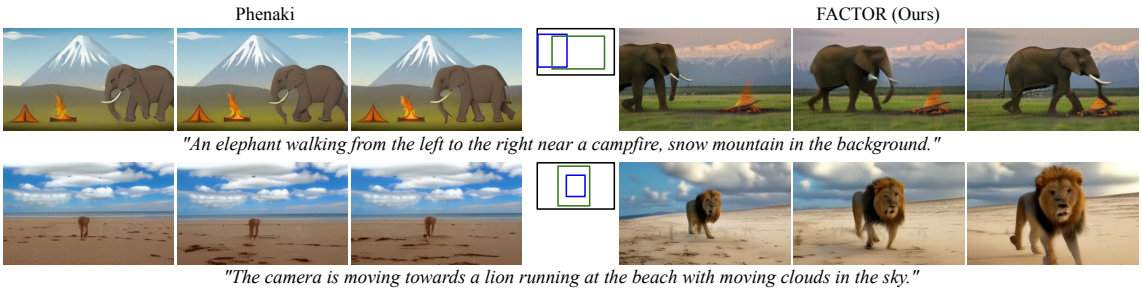

Figure 3.3: **Trajectory control.** We assess FACTOR's ability to control trajectories compared to the base model Phenaki [192]. The text prompt is augmented with trajectory descriptions as inputs to Phenaki. While Phenaki fails to accurately generate object movements, FACTOR successfully controls object movements. The blue and green boxes denote the positions in the initial and final frames.

addition, the test-time augmentation can be prompt-dependent, such as using a translation vector spanning downward to customize the target motion of getting up (See Figure 3.10).

3.4 Experiments

Dataset. Our model is trained on a dataset consisting of 10M videos from the WebVid dataset [9] and a randomly sampled subset of 500M images from the WebLI dataset [29]. Each training batch comprises 20% images and 80% videos.

Evaluation metrics. 1) `FVD` assesses video quality. 2) `CLIP-T` measures the alignment between prompts and generated videos using CLIP embeddings. 3) `AP` evaluates alignment to trajectories by comparing detected bounding boxes in generated videos with ground truth using an object detector [88]. 4) `CLIP-V` quantifies alignment to reference images by comparing image similarity in generated videos with ground truth using CLIP embeddings [168].

Implementation details. We implement the base T2V model following the architecture

Figure 3.4: **Enhanced subject interaction through trajectory control.** By utilizing trajectories of the two main entities as inputs, FACTOR improves the generation of subject interactions, a challenging task for T2V models.

of Phenaki [192]. The base model and FACTOR are trained for 1M and 500K steps with batch sizes of 256 and 128, respectively. Videos are generated at a length of 11 and a resolution of 192×320. The dimension of c_t is 220×512. The model is trained with a maximum of four entities. Further implementation details and model architectures are in the supplementary materials.

3.4.1 Trajectory Control

First, we evaluate FACTOR's ability to control the trajectories of generated entities. Since FACTOR is constructed based on Phenaki [192], we first validate that FACTOR enhances fine controllability compared to its base model without compromising quality. We design several prompts using descriptions of bounding box trajectories as inputs to Phenaki [192]. In Figure 3.3, Phenaki fails to generate object movements aligned with the designed prompts, whereas FACTOR accurately generates the correct movement direction for the entity by conditioning the generation on hand-drawn trajectories. In Figure 3.4, we use trajectories of two main entities as inputs. FACTOR generates videos aligned

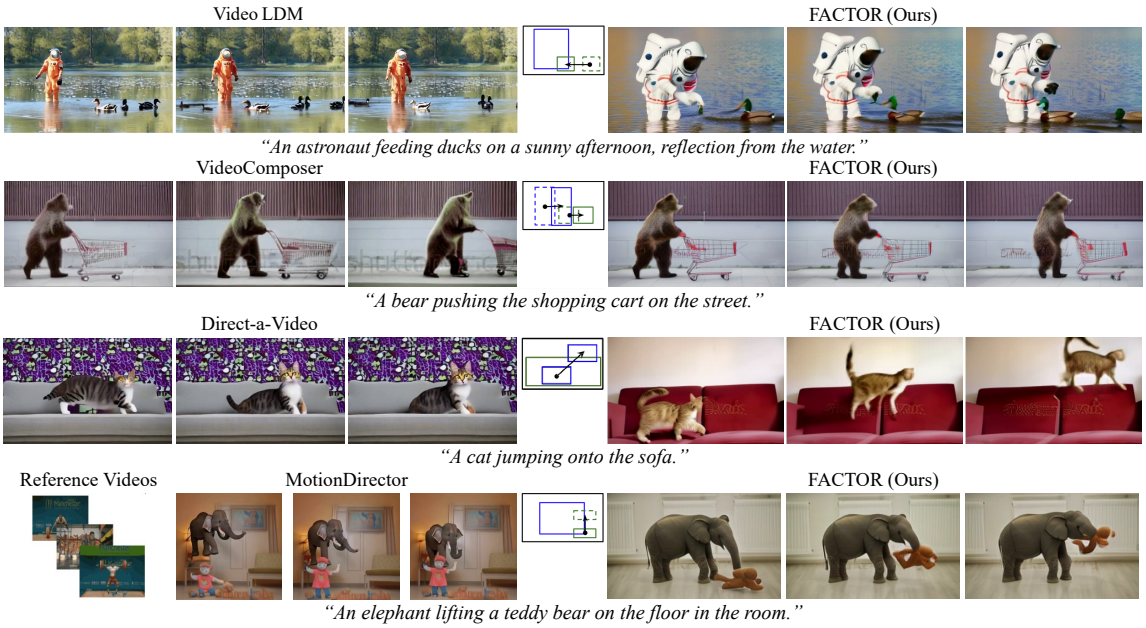

Figure 3.5: **Comparison to state-of-the-art.** By conditioning video generation on trajectories, FACTOR synthesizes videos with larger motions compared to T2V models. When using FACTOR's trajectory inputs, VideoComposer generates fade-out effects, Direct-a-Video struggles with overlapping boxes, and MotionDirector fails to generate videos with structures differing from the reference video.

with these trajectories and additionally benefits by generating interactions between entities, such as "high-fiving", even though our method is not specifically trained on annotated interactions.

In Figure 3.5, we demonstrate FACTOR's advantages by comparing it with state-of-the-art T2V and controllable T2V models. By conditioning the generation on input trajectories, FACTOR synthesizes videos with enhanced semantic meaning and larger motion, such as an astronaut stretching their hand to a duck. In contrast, the T2V model VideoLDM [14], which relies solely on text prompts, generates videos with less dynamic motion.

We compare FACTOR with controllable T2V models [204, 230, 259] designed to con-

Figure 3.6: **Comparison of subject customization.** FACTOR generates subjects with higher fidelity and larger movements compared to DreamBooth and VideoBooth, which lack trajectory control. MotionDirector struggles to generate accurate environments (e.g., New York).

trol subject movements. VideoComposer requires a reference image and dense motion sequence, but with FACTOR's sparse trajectory inputs, it often generates static or fade-out results due to its reliance on dense inputs. Direct-a-Video, which controls object locations by amplifying attention maps, struggles with overlapping boxes. FACTOR better aligns objects to bounding boxes by training on a large-scale video dataset. MotionDirector, which requires reference *videos* for motion control, only approximately controls object locations and cannot generate structures different from the reference (e.g., an elephant lifting an object).

3.4.2 Appearance Control

We demonstrate FACTOR's capability of appearance control. For this evaluation, we use images of subjects from [168, 113]. Each subject's appearance is conditioned on a single reference image for FACTOR. In this section, we compare FACTOR to state-of-the-art subject customization methods. We further evaluate FACTOR's control injection by comparing it to the control modules developed for T2I models. We then show FACTOR's ability to achieve controllable image animation, a special type of appearance control.

Figure 3.7: **Comparison of control modules.** FACTOR generates videos that align accurately with input trajectories and appearances. Other control methods, ELITE and GLIGEN, often generate inaccurate subject appearances and struggle to control their locations.

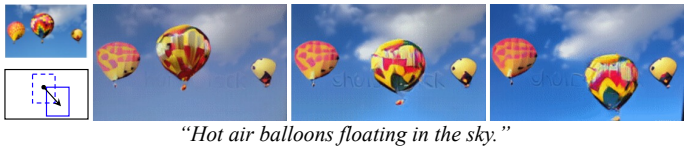

Figure 3.8: **Controllable image animation.** FACTOR achieves image animation conditioned on the initial frame and the trajectory.

First, we compare FACTOR with subject customization images and video synthesis methods including DreamBooth [14], VideoBooth [97], and MotionDirector [259]. Notably, these models cannot control entity trajectories or handle multiple entities. Figure 3.6 shows FACTOR generates high-fidelity customized subjects and exhibits larger movements by controlling trajectories, outperforming DreamBooth and VideoBooth. MotionDirector [259] uses separate temporal and spatial LoRA for motion and appearance controls, resulting in inadequate motion and incorrect environments (e.g., New York).

Next, we evaluate FACTOR's control module by comparing it to the control modules of ELITE and GLIGEN [208, 121] implemented on our video backbone. Since no existing models achieve the same control on videos as FACTOR, we compare methods originally designed for T2I tasks. In Figure 3.7, we show that FACTOR generates subjects that closely align with input trajectories and appearances. In contrast, ELITE and GLIGEN

often generate inaccuracies in subject appearance and struggle to effectively control their locations. GLIGEN's use of separate encoders for different controls introduces higher computational overhead and presents convergence challenges for video models, limiting its ability to inject precise control signals compared to FACTOR's unified encoder approach.

Though beyond the scope of this work, we show FACTOR's capability for appearance control in image animation, aiming to generate videos conditioned on the first frame. In Figure 3.8, FACTOR animates the hot air balloon in the initial frame based on the input trajectory. This capability leverages generative transformers adept at various infilling tasks. FACTOR's training involves predicting video tokens within random masks, enabling it to predict subsequent frames from the initial frame and control inputs, thus achieving controllable image animation.

3.4.3 Quantitative Results

We evaluate our models on the MSR-VTT test set [222], comprising 2,990 examples. We generate one video per example by randomly selecting one prompt from 20 prompts [14]. We use an unseen set of 6,513 videos as the real videos for FVD evaluation and compare two variants: 1) `FACTOR-T`: our model with trajectory control. 2) `FACTOR`: our model with trajectory and appearance control.

In Table 3.1, we first compare FACTOR and FACTOR-T with the base T2V model Phenaki, all using the same video backbone but accepting different levels of control inputs. To test our control module's capability, we use inputs automatically extracted from ground truth videos for FACTOR and FACTOR-T. This allows us to quantitatively evaluate the alignment between generated and ground truth videos, noting the extracted inputs are noisier than user-provided ones. FACTOR improves FVD scores with control inputs that enforce closer alignment to real video distributions in trajectories and appearances. However, it achieves slightly lower CLIP-T scores than Phenaki, likely due to misalignment between extracted controls and text prompts in test data. Longer training enhances control

Table 3.1: **Quantitative results.** FACTOR achieves higher AP and CLIP-V scores, demonstrating its capability to generate videos aligned with trajectory and appearance control inputs.

Methods	FVD ↓	CLIP-T ↑	AP ↑	CLIP-V ↑
T2V models				
MagicVideo	1290	—	—	—
VideoLDM	—	0.2929	—	—
Make-A-Video	—	**0.3049**	—	—
ModelScope	550	0.2930	—	—
Controllable T2V models				
VideoComposer	342	0.2906	0.126	0.742
Direct-a-Video	360	0.2849	0.144	0.658
GLIGEN	217	0.2817	0.173	0.703
ELITE	130	0.2712	0.137	0.733
Same backbones				
Phenaki*	411	0.2870	0.099	0.663
FACTOR-T (Ours)	339	0.2787	0.290	0.683
FACTOR (Ours)	**116**	0.2721	**0.356**	**0.763**

*We implement and train the Phenaki model from scratch using our datasets.

alignment but could reduce prompt alignment, as noted in concurrent works [138]. We also assess AP scores. While Phenaki is not designed for generating videos with specific object trajectories, its AP score reflects chance alignment when prompts sufficiently match input trajectories. FACTOR's higher AP score indicates successful alignment with provided trajectory controls. Additionally, FACTOR outperforms Phenaki in CLIP-V scores, showing superior alignment with appearance controls. We also present results from other state-of-the-art T2V models trained on different data and backbones for reference.

Table 3.2: **User study.** FACTOR consistently achieves performance gains using inputs provided solely by users. Average scores on a 0-2 scale are reported based on 16 prompts and 5 subjects.

	Quality↑	Text↑	Trajectory↑	Appearance↑
Phenaki *v.s.* FACTOR-T	1.13 / **1.63**	0.72 / **1.97**	0.16 / **1.96**	-
Phenaki *v.s.* FACTOR	1.25 / **1.63**	1.70 / **1.75**	0.43 / **1.88**	0.38 / **1.75**

Table 3.3: **Ablation study.** FACTOR consistently outperforms ablated models across various metrics.

Methods	FVD↓	CLIP-T↑	AP↑	CLIP-V↑
FACTOR	**116**	0.2721	**0.356**	**0.763**
w/o joint encoder	133	0.2719	0.307	0.726
w/o control-attention	127	0.2720	0.322	0.750
w/o augmentation	124	**0.2723**	0.341	0.758

Next, we compare FACTOR's control injection module with controllable T2V models. Compared to VideoComposer [204] and Direct-a-Video [230], FACTOR achieves better FVD, AP, and CLIP-V scores. These results highlight FACTOR's enhanced controllability with user-friendly trajectory inputs. Compared to ELITE and GLIGEN [208, 121], FACTOR also outperforms in terms of FVD, AP, and CLIP-V scores, highlighting the effectiveness of FACTOR's control-attention module.

Finally, we conduct a user study using input controls provided by users. Raters assess two videos generated by different methods using the same inputs on a 0-2 scale for the following criteria: quality, text alignment, trajectory alignment, and appearance alignment. Average scores from 16 videos and 5 subjects are reported in Table 3.2. FACTOR consistently achieves higher ratings across all criteria.

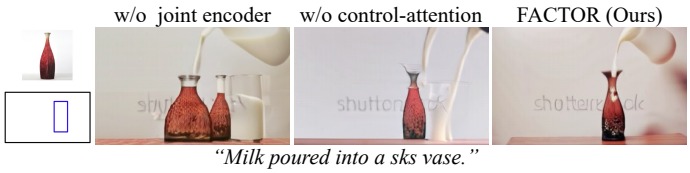

Figure 3.9: **Ablation study.** FACTOR achieves better alignment with control inputs compared with ablated models.

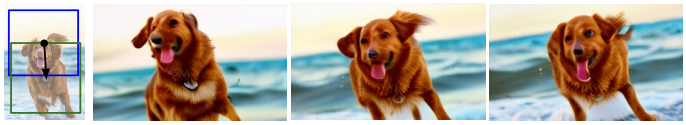

Figure 3.10: **Appearance augmentation.** We demonstrate fine-tuned augmentation at test time as a form of control. For instance, adjusting the translation of the reference image downwards to achieve a desired jumping motion.

3.4.4 Ablation Study

We conduct an ablation study to validate the effectiveness of each proposed module. In Table 3.3 and Figure 3.9, we compare the following models: 1) `w/o joint encoder`: uses separate encoders for prompt and fine control. 2) `w/o control-attention`: concatenates prompt and fine control inputs to the original cross-attention layer. 3) `w/o augmentation`: trained without appearance augmentation. FACTOR consistently achieves higher metrics compared to these alternative models, confirming the importance of the proposed components. Here, our appearance augmentation technique proves effective across different prompts, with augmentations randomly sampled.

In Figure 3.10, we further demonstrate that appearance augmentation can be dynamically adjusted at test time to better meet user preferences. By applying a translation vector, such as moving downwards for a jumping motion, we use variations of the reference image as appearance conditions at different time steps, which enhances the realism of live subjects' motions. In Figure 3.11, we show that FACTOR handles partial inputs where one or

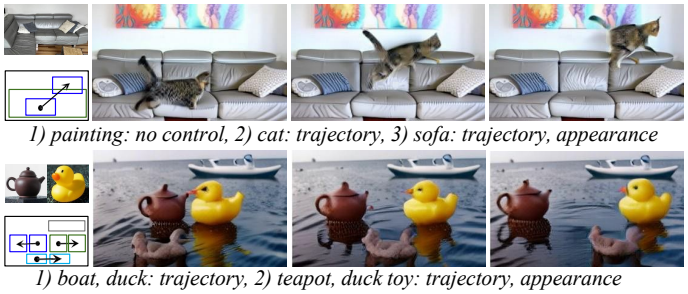

Figure 3.11: **Partial and multiple control.** FACTOR demonstrates capability with partial inputs and supports up to four objects.

more controls are absent for the entities. Also, FACTOR supports up to four objects.

3.5 User Interface

Our goal is to provide a user-friendly, controllable video generation framework that addresses fine-grained control. We present an illustration of the user interface of FACTOR. In our framework, users first specify the objects in the video and then provide fine-grained control to generate each object, including their appearance and context. The context to generate each entity includes 1) a description of the entity, 2) a user-drawn trajectory, and 3) a user-provided reference image to achieve both structural and appearance control. The control inputs are provided by the users following these steps:

- Step 1: The user types the text prompt.
- Step 2: The user types the description of the entity.
- Step 3: The user draws the trajectory for the entity by first drawing a bounding box for its starting location and then dragging the bounding box to its end location.
- Step 4: The user uploads a reference appearance image for the entity.
- Step 5: The user repeats Steps 2-4 to input the control for all entities.

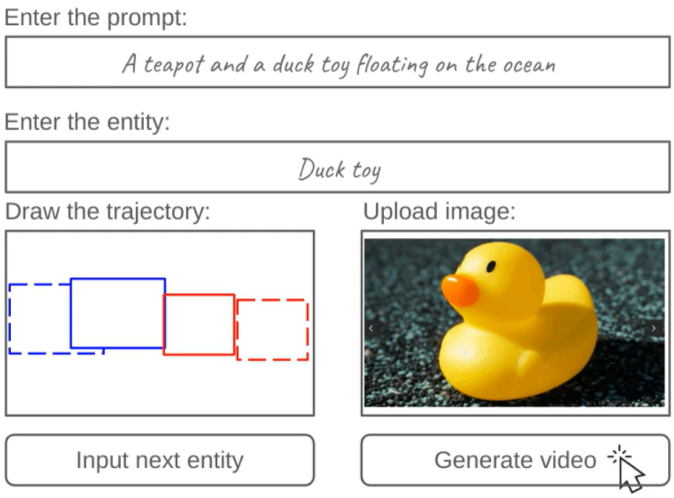

Figure 3.12: We present an illustration of the user interface of FACTOR.

3.6 Limitations

First, FACTOR's performance relies on the capabilities of the base T2V model. Integrating the proposed control module into more advanced models could enhance overall quality. Second, FACTOR implicitly manages camera poses (Figure 3.6). Explicit control over camera poses could be achieved by integrating them into FACTOR's control sequences. Third, FACTOR faces challenges when text prompts do not align well with fine-grained control inputs. Lastly, generating subject poses significantly different from a single reference image is challenging for FACTOR. Training FACTOR on datasets with multiple references and upgrading the base model could improve its performance.

3.7 Conclusions

We introduce a novel approach for fine-grained, controllable video generation. Our framework allows users precise control over video creation by specifying entity names, drawing trajectories, and providing visual appearance examples. Key components include

a joint encoder for learning interaction among controls, control-attention modules for precise control injection, and appearance augmentation for diverse subject poses. This approach enables nuanced control over the appearance and trajectory of multiple objects in generated videos.

Chapter 4

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

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.

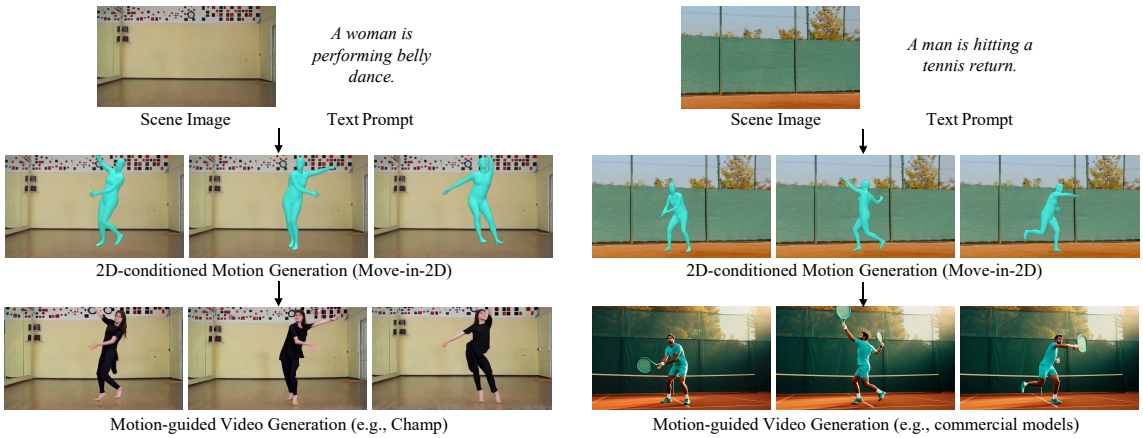

Figure 4.1: **2D-conditioned human motion generation.** Given an image representing the target scene and a text prompt describing the desired motion, we generate a motion sequence that aligns with the text description and projects naturally onto the scene image. This generated motion then serves as the control signal for the subsequent video generation tasks.

4.1 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 [265, 84, 224, 199, 100] 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 Figure 4.1. 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 [185, 30, 1], which, while straightforward, may not produce motion that seamlessly integrates into specific target environments, requiring further adjustments for scene compatibility. Other methods [206, 90] 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 perspective, 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 [62]. 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) [3, 207, 152], 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.

4.2 Related Work

4.2.1 Human Video Generation

Human-centric video generation typically leverages ControlNet to condition the generation process on motion guidance signals, such as OpenPose [199, 84], DensePose [224, 100] keypoints, or SMPL mesh sequences [265]. 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.

4.2.2 Human Motion Datasets

Several datasets have been proposed to facilitate research on human motion understanding and generation. Datasets such as CMU Mocap [37], Human3.6M [92], and MoVi [60] capture human motion but lack textual descriptions of the actions. The KIT Motion Language Dataset [156] provides both motion sequences and textual prompts, containing approximately 3.9K motion sequences. HumanML3D [65] expands this number to 14.6K by sourcing motion data from HumanAct12 [67] and AMASS [136]. The Motion-X dataset [126] scales up further 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 [164, 226] include SMPL models with global positions in 3D scenes. The PROX dataset [76] utilizes optimization techniques with RGB-D data to reconstruct human motions, while others [4, 196] compile synthesized data of human-scene interactions. Additionally, datasets like [206, 258] incorporate both scene context and language descrip-

tions 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.

4.2.3 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 [66] and TEACH [6], enabling improved control and the generation of longer, more coherent sequences. Building on these advancements, models such as MotionGPT [96] leverage large language model (LLM) pretraining for text-driven motion synthesis.

More recently, many works have applied diffusion models in motion generation tasks. MotionDiffuse [251] and MDM [185] generate motion sequences aligned with text prompts from an initial random noise. ReMoDiffuse [252] introduces a retrieval-augmented model, where knowledge from retrieved samples enhances motion synthesis. MLD [30] performs motion diffusion in a latent space using a variational autoencoder. EMDM [264] 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.

4.2.4 Scene-aware Human Motion Generation

Given a 3D indoor scene, prior works [256, 256, 77, 75, 196, 197] have demonstrated the generation of physically plausible human poses or motion sequences. HUMANISE [206] aligns captured human motion sequences with various 3D indoor scenes, using

text prompts as conditioning inputs. LaserHuman [40] 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.

4.3 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 [112, 16, 151], 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 Section 4.2 and Table 4.1 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.

4.3.1 Data Collection

While action recognition datasets [103, 144] 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 [61] and OpenPose for keypoint prediction [20], following [16]. 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

Table 4.1: **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 [156]	3.9k	6.2k	No	No	Indoor
HumanML3D [65]	14.6k	44.9k	No	No	Indoor
HUMANISE [206]	19.6k	19.6k	643	RGBD	Indoor
PROX [76]	28k	No	12	RGBD	Indoor
LaserHuman [40]	3.5k	12.3k	11	RGBD	Indoor/Outdoor
Motion-X [126]	81.1k	81.1k	81.1k	Video	Indoor/Outdoor
HiC-Motion	300k	300k	300k	Video	Indoor/Outdoor

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.

4.3.2 Data Preprocessing

To obtain human motion annotations from the selected videos, we use the off-the-shelf method 4D-Humans [62] 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 [93] 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 images [101] as well as random cutout augmentations.

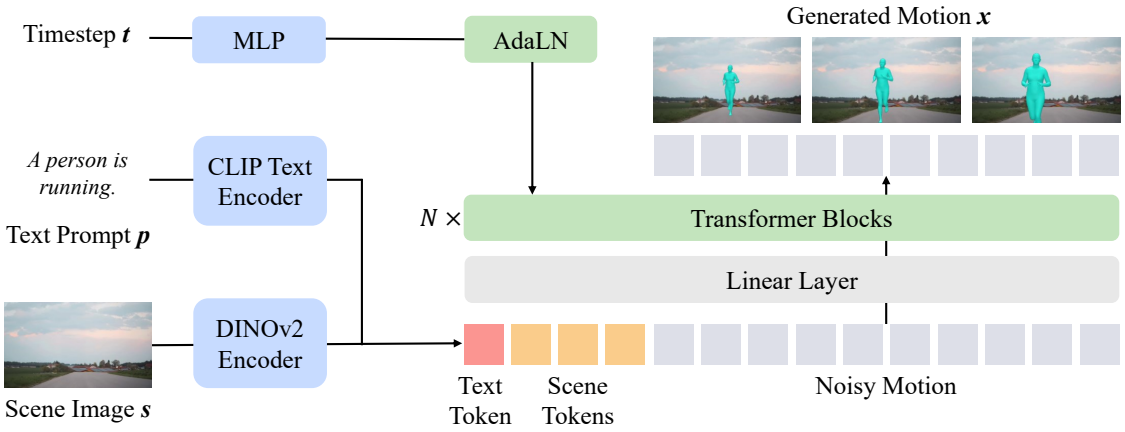

Figure 4.2: **Overview.** The text prompt and background scene image are encoded by the CLIP and DINO encoders, and incorporated into the model via in-context conditioning. The AdaLN layer receives the diffusion timestep as input. Our multi-conditional transformer model then generates a human motion sequence through a diffusion denoising process, aligning the generated motion with both input conditions.

4.4 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 models in Section 4.4.1. In Section 4.4.2, we propose a conditional motion diffusion model. We then introduce a multi-conditional transformer in Section 4.4.3. Finally, in Section 4.4.4, we present our training strategy. Figure 4.2 provides an overview of our approach.

4.4.1 Preliminaries on Diffusion Models

Diffusion models like DDPM [81, 185] 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) [79] is applied to enhance alignment with conditions.

4.4.2 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 [131], along with a global orientation parameter $\theta_g \in \mathbb{R}^6$ that defines the overall human body orientation. Unlike previous motion generation approaches [185, 30], 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.

$$\mathcal{M}_{\text{cfg}} = \mathcal{M}(\mathbf{x}_t|t) + g\left(\mathcal{M}(\mathbf{x}_t|t, p, s) - \mathcal{M}(\mathbf{x}_t|t)\right). \tag{4.1}$$

4.4.3 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 [152, 185]. 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 [162], and the background image condition s by a DINO encoder [149] 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 [152] 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) [153].** 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 Section 4.5.2).

4.4.4 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.

4.5 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* [185] and *MLD* [30] 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 [228] 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* [90], which uses 3D point clouds as input, and *HUMANISE* [206], 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 [67, 154] 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 [154, 241] 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

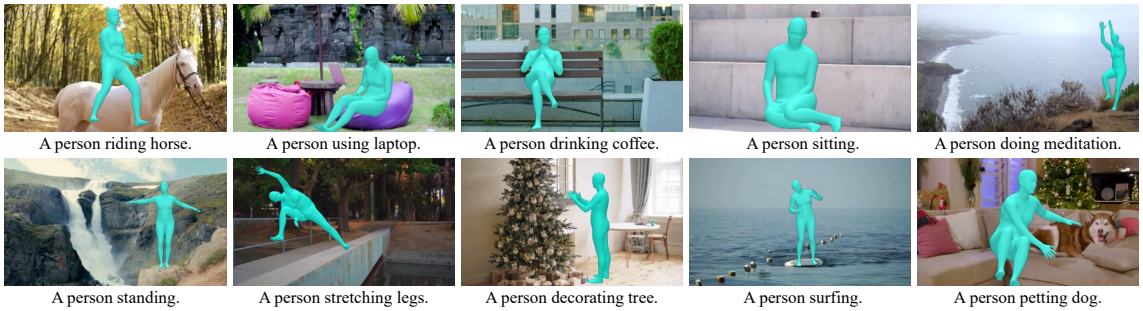

Figure 4.3: **Affordance-aware human generation.** Our model generates human poses consistent with both text prompts and scene context, such as standing on a cliff. It also supports complex human-scene interactions, including activities like petting a dog.

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 $D = 147$, 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 [149], and text prompts into a single token using CLIP-B [162], resulting in a sequence of 497 tokens.

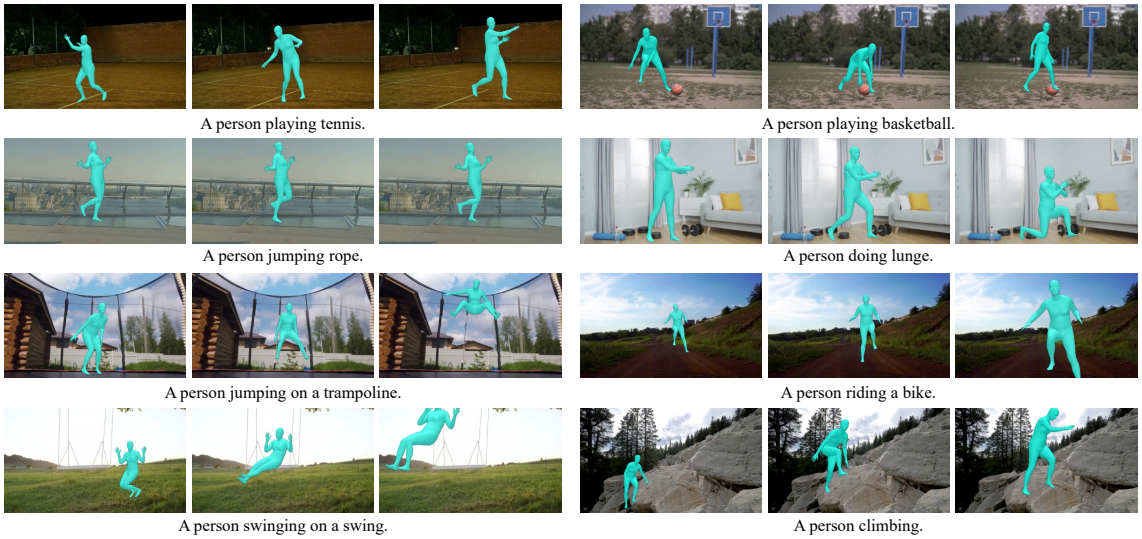

Figure 4.4: **Motion generation with large dynamics.** Our results show motion sequences that are accurately placed and move within scenes, such as playing tennis, enabling the generation of complex human activities that are challenging for video generation models.

4.5.1 Qualitative Results

Affordance-aware human generation. Figure 4.3 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 Figure 4.4, 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.

Figure 4.5: **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.

Comparison to state-of-the-art. As shown in Figure 4.5, 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 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.

4.5.2 Quantitative Results

The evaluation results are shown in Table 4.2. 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-

Table 4.2: **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 [185]	164.595	0.325	24.758	18.924
MLD [30]	85.913	0.322	25.119	19.464
SceneDiff [90]	543.769	0.203	4.217	3.861
HUMANISE [206]	159.935	0.225	23.287	19.956
MDM+ [185]	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

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*

Table 4.3: **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 [185]	2.25	1.35	1.50	5.10
MLD [30]	2.85	1.95	1.90	6.70
SceneDiff [90]	2.05	1.20	1.20	4.45
HUMANISE [206]	2.20	1.45	1.30	4.95
MDM+ [185]	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**

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 [148] 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 Table 4.3. Our method consistently outperforms 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 Section 4.4, we train our 2D-conditioned motion diffu-

Table 4.4: **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

sion 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 Table 4.4. 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.

4.5.3 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 [265] 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 [48] 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

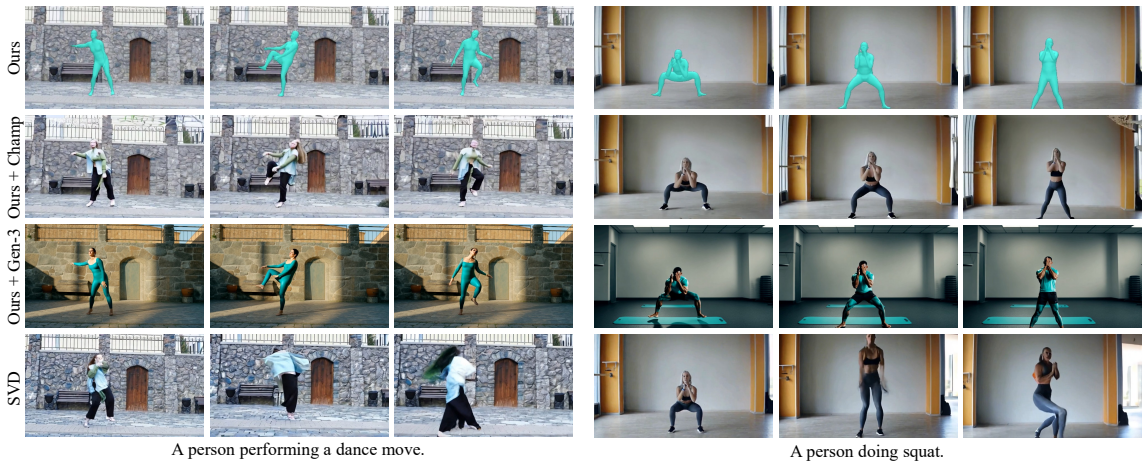

Figure 4.6: **Motion-guided human video generation.** Our approach generates scene-compatible motion sequences from a scene image and text prompt, which are then used to animate a reference human using Champ [265] or the commercial model [48]. The generated motion ensures accurate human shapes and smooth motion in the resulting videos, outperforming SVD [13] in preserving human geometry and motion consistency.

image provides the desired background's layout and semantic information. As shown in Figure 4.6, 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 Figure 4.1). We also include results from Stable Video Diffusion (SVD) [13] 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.

4.6 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.

Chapter 5

Generating Long-take Videos via Effective Keyframes and Guidance

We tackle the challenge of generating long-take videos encompassing multiple non-repetitive yet coherent events. Existing approaches generate long videos conditioned on single input guidance, often leading to repetitive content. To address this problem, we develop a framework that uses multiple guidance sources to enhance long video generation. The main idea of our approach is to decouple video generation into keyframe generation and frame interpolation. In this process, keyframe generation focuses on creating multiple coherent events, while the frame interpolation stage generates smooth intermediate frames between keyframes using existing video generation models. A novel mask attention module is further introduced to improve coherence and efficiency. Experiments on challenging real-world videos demonstrate that the proposed method outperforms prior methods by up to 9.5% in objective metrics.

5.1 Introduction

Recent advancements in text-to-video generation have shown significant progress in enhancing visual fidelity, including improvements in resolution, frame rate, and video length [247, 205, 26, 13, 198, 58, 132, 203, 27, 70, 68, 104, 200, 220]. These models excel in generating impressive visuals within a scene but face challenges when it comes to constructing complex sequences involving multiple events, due to several limitations. First, they produce video clips of a predefined fixed length, which is significantly shorter than real-world videos due to substantial computational and memory overhead. Second, the input conditions, such as a class label [183, 50, 190, 147, 139, 194, 169, 188, 170, 36] or a text prompt describing the entire video [83, 213, 212, 80, 178, 14, 104, 58, 192, 262, 78], often lack complexity and richness. However, real-world videos encompass diverse content that surpasses the capability of encoding within such simplistic conditions. Thus, optimizing the video generation model alone cannot resolve these limitations and close the gap between generated and real-world videos. For example, generating a real-world video as shown in Figure 5.1 requires creating a meaningful storyline with different events, such as cutting food, moving a plate, and putting it into the fridge. To achieve this goal, it is necessary to develop a video generation system that creates the high-level contents of multiple events and employs existing models to create detailed frames for each event.

Several methods address the challenge of generating long videos. Autoregressive approaches [57, 238] create video clips sequentially, with each frame building upon previously generated ones. Latent-based methods [179, 253] utilize a shared latent content vector across all frames. While successfully generating lengthy videos, these techniques often produce homogeneous content featuring natural scenes and recurring human actions. Video generation conditioned on multiple guidance sources [10, 245, 57, 74, 237] generate diverse events but are typically limited to synthetic environments and cartoons. Text-to-video models successfully handle transitions within videos [161, 32], but efficiently prompting the model to generate long videos across multiple clips remains an open prob-

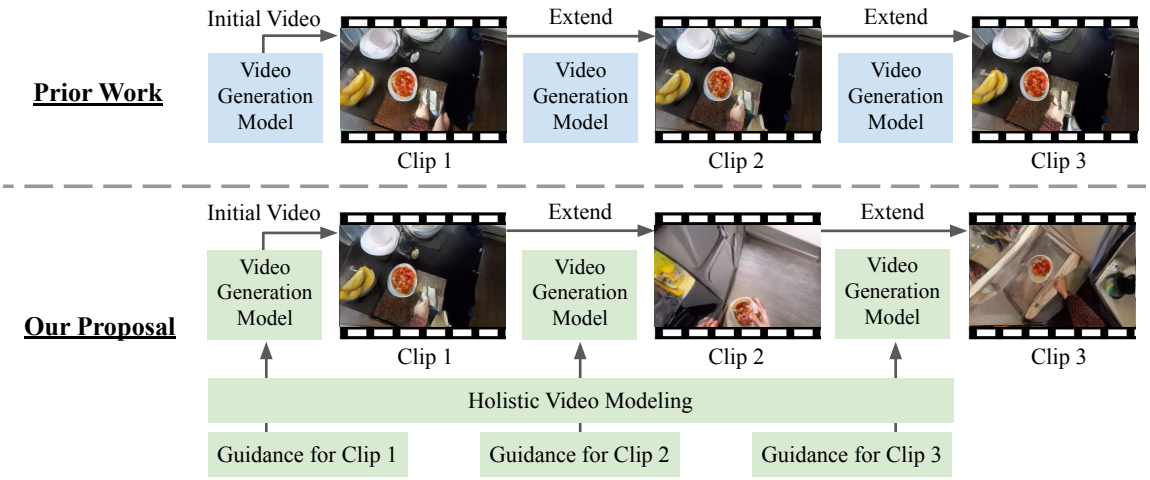

Figure 5.1: Existing schemes [57] generate long videos by iteratively extending the generated frames, often leading to videos with repetitive patterns. In contrast, we generate *long-take videos encompassing multiple non-repetitive yet coherent events* using multiple guidances.

lem, often requiring extensive trial-and-error and manual post-processing. Recently, multi-shot video generation [125, 266, 122] has been introduced to synthesize videos consisting of multiple distinct snippets lacking direct temporal continuity between shots. However, generating a *long-take single-shot* video featuring *multiple semantically different events*, as illustrated in Figure 5.1, remains a challenging problem that requires further exploration.

We tackle the challenges of generating long-take videos encompassing multiple real-world events. To achieve this, guidance signals provided at multiple timesteps are essential for effectively controlling the generation process. We utilize various levels of guidance, including object labels as high-level guidance that describe objects in the generated video and image layouts as low-level guidance. The core concept of our approach is to divide the video generation process into keyframe generation and frame interpolation. Keyframe generation focuses on creating multiple coherent frames with varied video content, each representing key events specified by the guidance. The frame interpolation stage then

generates smooth intermediate frames to connect these keyframes using existing video models. Our method is orthogonal to existing efforts to optimize the generation quality of video models and is compatible with different video generation models by incorporating them into our system. To ensure temporal consistency in the generated videos, we introduce a novel masked attention module to enhance efficiency and object coherence during long-take video generation. We conduct extensive experiments on real-world datasets, demonstrating that our approach outperforms state-of-the-art video generation models by 9.5% on the FVD metric. Our main contributions are as follows:

- We address the problem of generating long-take videos with multiple non-repetitive yet coherent events.
- We propose a multi-stage approach involving the generation and the connection of keyframes.
- We introduce a masked attention module to enhance continuity and efficiency for long-take videos.

5.2 Related Work

5.2.1 Video Generation

GAN-based methods [194, 169, 188, 36, 170, 179, 186, 17] have demonstrated early success in generating short video clips, while the generation quality decreases significantly when applied to long videos. Recently, diffusion models [193, 229, 82] have shown promising visual quality in video generation. Auto-regressive models [99, 209, 8] have benefited from developments in vector quantization [49, 191] and transformer [44] models, allowing them to predict discrete tokens in the latent space [227, 143, 57, 22, 243, 109] with impressive visual quality. Despite the recent success in video generation, these models mostly focus on generating videos with a predefined length, often shorter than real videos, or are limited to synthesizing long videos with repetitive contents [192, 57], such

as human actions [181, 176, 15] and sky timelapse [221] under the setting of unconditional video generation [17, 179, 253, 238]. Few works explore video generation conditioned on input guidance at multiple timesteps [10, 245, 57, 74, 237], but they are limited to synthetic environments. In contrast, our work aims to generate real-world videos. Most recently, text-to-video models [212, 213, 83, 178, 80, 14, 78, 205, 68, 247, 205, 26, 13, 198, 58, 132, 203, 27, 70, 68, 104, 200, 220] have been developed to generate videos given natural language inputs. However, the generated videos are limited in length and complexity due to computation constraints and limitations in text description. Instead of trying to increase the capacity of existing models, we focus on an alternative approach to improving video generation, namely generating videos with multiple events. Our approach is generic, and off-the-shelf video generation models can be adopted as a component in our pipeline. Concurrent video generation works address different topics such as video-to-video translation [195, 59, 239, 38], generating style or static-dynamic transitions [161], and creating artificial transitions between distinct scenes [32]. In contrast, we focus on long-take video generation from high-level guidance, connecting keyframes by natural transitions to model semantic changes (e.g., moving to different rooms, picking up objects) in real videos.

5.2.2 Story Visualization

Story visualization [120, 180, 135, 134] aims to synthesize a sequence of images that depict a story composed of multiple sentences. GeNeVA [46, 51, 130, 39] is a conditional text-to-image generation task developed on the CoDraw [105] dataset, which addresses the problem of constructing a scene iteratively based on a sequence of descriptions. However, this line of work primarily focuses on experiments conducted with synthetic data and targets the generation of images. In contrast, our focus lies in generating videos using real-world data.

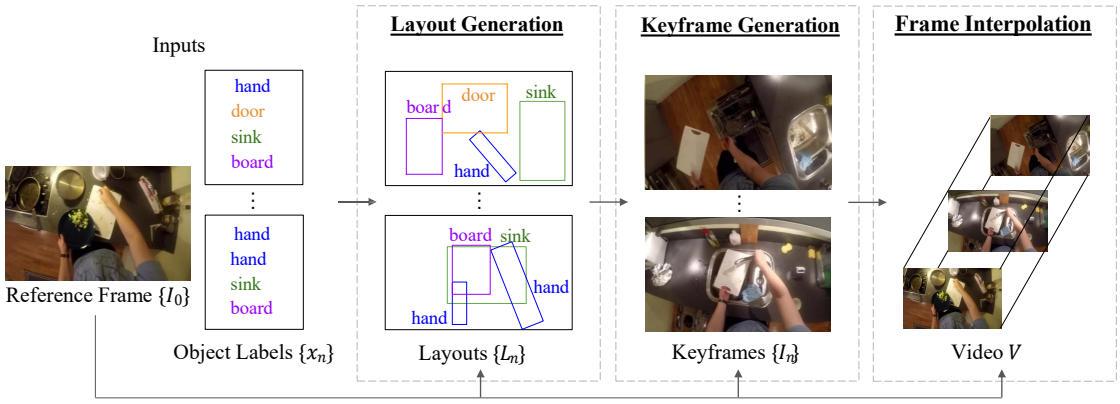

Figure 5.2: **Approach overview.** The proposed method consists of three stages: 1) *Layout generation* predicts a series of layouts from the object label sets. 2) *Keyframe generation* predicts a sequence of keyframes for the video using the layout sequences and the reference frame as inputs. 3) *Frame interpolation* synthesizes intermediate frames based on the keyframe sequence to obtain the complete video.

5.3 Method

5.3.1 Overview

We first provide an overview of our long-take video generation system and then describe the three stages in detail. Given a series of N guidance signals and a reference frame I_0 as input, our objective is to synthesize a video V that encompasses the content specified by the guidance, starting from the initial frame I_0. The input guidance consists of object labels $x_1, x_2, \cdots, x_N$. Each set x_n contains k^n object labels, which may vary across timesteps n, indicating the objects appearing in the video. To address the challenge of generating videos covering multiple events, our approach decouples the video generation process into layout generation, keyframe generation, and frame interpolation, as shown in Figure 5.2. Layout generation explicitly predicts the structures of keyframes from object labels, addressing the deterioration of structure prediction over time during long video

generation. This layout can also be provided and edited by users. Keyframe generation focuses on jointly modeling the entire video to create multiple coherent events, with each keyframe representing core events specified by the guidance. Using existing video models, the frame interpolation stage then synthesizes smooth intermediate frames between keyframes. Specifically, our system produces a series of N keyframes in the video based on the object label sets. Each keyframe serves as the initial and final frames of short video clips. An existing video generation model predicts intermediate frames between keyframes to complete the video.

Our approach offers several advantages: First, it simplifies generating long videos by predicting keyframes that encapsulate high-level event concepts. Second, it leverages advanced video generation models to generate intermediate frames, enhancing flexibility in video synthesis. We utilize generative transformer models capable of handling various modalities by converting them into a unified vocabulary. This capability allows for incorporating different levels of input guidance (e.g., labels and layouts) in our task. Transformers are adept at tasks such as interpolation and frame prediction, making them versatile across our multi-stage approach. They also offer significant speed advantages and achieve competitive output quality compared to diffusion models [244].

5.3.2 Layout Generation

We first generate a series of layouts from the object label sets to explicitly constrain the structures of keyframes. Given a series of N object label sets $\{x_n\}$ and a reference frame I_0, we synthesize a series of layouts $\{L_n\}$ representing the layouts of the N keyframes in the video. Since the reference first frame I_0 is given, we assume L_0 is known and can be used as a reference layout. We define the layout L as a set of bounding boxes with a variable length k^n, i.e., $\{b_1, b_2, \cdots, b_{k^n}\}$. The bounding box attributes include its object label, x-coordinate, y-coordinate, width, and height, i.e., $b = \{c, x, y, w, h\}$. We apply tokenization to encode the layouts $\{L_n\}$ and the object label set $\{x_n\}$ into discrete token sequences, where

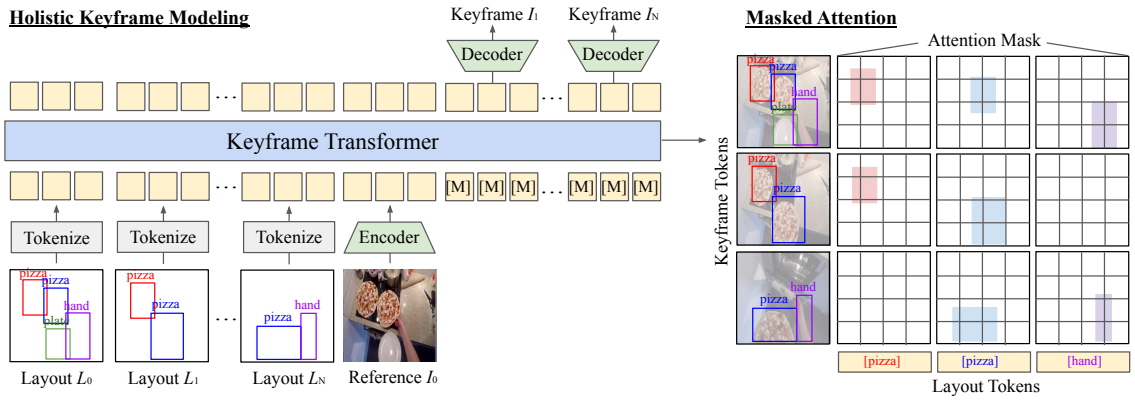

Figure 5.3: **Keyframe generation.** Our key ideas are: 1) *Holistic keyframe modeling*: the model predicts multiple keyframes jointly to maintain coherence, using the layouts and the reference frame as inputs. 2) *Masked attention*: the attention mask between irrelevant keyframe and layout tokens is set to zero, encouraging the model to focus on relevant tokens, thus improving coherence and efficiency. Tokens of a specific color in a bounding box are associated with layout tokens of the same color, indicating they represent the same object.

the values of the bounding box attributes are discretized by uniform quantization.

Given the tokens of the reference layout L_0 and the object labels c as inputs, we replace the other attributes of the bounding boxes $\{x, y, w, h\}$ with a `[MASK]` token and train the model to predict the masked bounding box attributes in the layout series. Our layout generation module predicts multiple layouts jointly to preserve the temporal consistency of the generated layout sequence. Although few datasets contain annotations of labels and layouts, we utilize an off-the-shelf segmentation model to obtain annotations through an automatic pipeline, which is readily scalable to general-domain real-world datasets.

5.3.3 Keyframe Generation

Next, we generate a sequence of keyframes from the predicted layout sequences. Given a reference first frame I_0 and a series of layouts $\{\hat{L}_n\}$ either provided by users or synthesized in the previous stage, our goal is to generate a sequence of keyframes $\{\hat{I}_n\}$. Specifically, the keyframe sequence $\{I_n\}$ is encoded into visual tokens, $\{e_n\}$, by a VQ-VAE [49] model. During training, the tokens of layouts and keyframes are concatenated as $s = \{L_0, \hat{L}_1, \cdots, \hat{L}_N, e_0, e_1, \cdots, e_N\}$ in dataset $\mathcal{D}$. We randomly replace tokens in the sequence with the `[MASK]` token, obtaining the masked sequence s_M. The generative transformer model is trained to predict the masked tokens of the target keyframes $s_t, t \in M$, by minimizing the negative log-likelihood:

$$\mathcal{L} = - \mathop{\mathbb{E}}_{s \in \mathcal{D}} \sum_{t \in M} \log p(s_t \mid s_M) \tag{5.1}$$

At test time, given the tokens of the layout sequence and the first frame $\{\hat{L}_n\}$ and e_0, the model predicts the tokens of the subsequent keyframes $\{\hat{e}_n\}$, which are then reconstructed into raw images $\{\hat{I}_n\}$ by the decoder of VQVAE, as depicted in Figure 5.3. Our keyframe generator predicts all frames holistically throughout the entire video, enhancing consistency across keyframes and the final long video.

Masked attention. We further propose a novel masked attention module to improve the efficiency and coherence of keyframe generation. Since our input to the keyframe generator is a concatenation of layout and keyframe tokens, each token attends to every other token without considering the relationship between the layout and the keyframe. However, this approach may be redundant because the visual tokens within a specific bounding box in the keyframe ($\hat{e}_n$) are highly related to the layout tokens of the corresponding bounding box in $\hat{L}_n$. Moreover, since we generate a series of keyframes jointly, the tokens of each bounding box in a single layout $\hat{L}_n$ are correlated with every relevant visual token in all the keyframes $\{e_n\}$, as long as the bounding box represents the same instance in the sequence. We encode this prior knowledge into our model through the attention mask,

setting irrelevant attention to zero to guide the model to focus on relevant tokens. This prior knowledge helps the model respect the relationship between layout and keyframe to improve their alignment and temporal coherence for the same instance. Additionally, masked attention reduces computational costs and improves efficiency in handling long keyframe sequences, as shown in Table 5.4.

5.3.4 Frame Interpolation

Finally, given the reference frame I_0 and a sequence of generated keyframes $\{\hat{I}_n\}$, we input the keyframes into an existing video generation model [57, 243, 14] trained for infilling tasks to generate intermediate frames between the keyframes. Specifically, we use a model that predicts intermediate frames based on the initial and final keyframes as input. We first convert the video into discrete tokens using a 3D-VQVAE model. Then, a transformer model is employed to predict masked video tokens for the intermediate frames. Finally, the interpolated video tokens are decoded back into raw videos by the 3D-VQVAE decoder.

During inference, given two consecutive keyframes $\hat{I}_{n-1}$ and $\hat{I}_n$, the model predicts video token sequences that connect these keyframes, denoted as $\hat{z}_n$. N clips of video tokens $\{\hat{z}_n\}$ are predicted and concatenated into a sequence to generate the long video using the 3D decoder. Since our keyframe generation module has already generated consistent objects over a longer duration, it facilitates the frame interpolation stage to maintain object consistency and temporal coherence when predicting intermediate frames. Our frame interpolation module can also be adapted to models trained on open-domain data, such as [80, 68, 14], to enhance the quality and diversity of generated videos.

5.4 Experiments

Dataset. We experiment with two real-world video datasets: 1) `EPIC Kitchen` [41] comprises egocentric videos of kitchen activities. Unlike commonly used datasets such as [181, 103, 45], EPIC Kitchen videos feature complex and dynamic scenes. Foreground objects move in and out of the camera's field of view, while background scenes and camera viewpoints frequently change, including transitions between different rooms. To synthesize such videos, the video generation model needs to generate multiple non-repetitive events within a short time window, such as various actions performed by the subject. 2) `nuScenes` [18] consists of driving videos that capture both object motion (e.g., moving cars) and dynamics caused by egocentric motion, where background scenes change over time.

Evaluation metrics. We evaluate the generated videos using the following metrics: 1) `FVD` and `FID` assess the quality of generated videos by measuring the feature distribution compared to real videos. 2) `LPIPS` assesses the similarity between the generated and ground truth video frames. 3) `AP` measures the semantic correctness by comparing the detected objects between the generated and ground truth frames. 4) `CLIP` computes the CLIP similarity of the generated frame and a text prompt constructed from the input labels. 5) `Cons.` assesses frame consistency using the CLIP image similarity between two frames [204, 48].

Implementation details. We utilize a segmentation model [34], not specifically trained on our datasets, to automatically extract sparse labels without human effort. The model covers common objects and is readily extendable to various real-world video datasets in general domains. These labels describe both foreground objects and dynamic background scenes (e.g., walls, tables, floors). We set the maximum number of objects to 14 and pad sequences if fewer than 14 objects are detected. We sample one keyframe every 16 frames, resulting in 16×N frames per video. Our quantitative evaluation uses 64-frame videos at 128×128 resolution, with longer videos of up to 512 frames included in visual results. Inference costs are 0.01 seconds for layout generation, 1.5 seconds for keyframe

Table 5.1: **Quantitative results of video generation.** Metrics are averaged across frames. Our method consistently outperforms state-of-the-art video generation models.

(a) EPIC Kitchen

Methods	FVD ↓	LPIPS ↓	AP ↑	CLIP ↑	Cons. ↑
TATS	737.6	0.615	0.072	0.2452	0.9971
StyleGAN-V	581.8	0.550	0.074	0.2257	0.9960
MAGVIT (frame pred.)	291.4	0.475	0.133	0.2785	0.9965
MAGVIT (class cond.)	285.6	0.469	0.140	0.2788	0.9968
Ours	**258.4**	**0.421**	**0.192**	0.2855	**0.9973**
Ours (GT layouts)	214.7	0.346	0.390	**0.2866**	0.9970
Ours (GT keyframes)	174.1	0.242	0.451	0.2865	0.9972

(b) nuScenes

Methods	FVD ↓	LPIPS ↓	AP ↑
MAGVIT (frame pred.)	171.1	0.478	0.044
Ours	**158.9**	**0.448**	**0.059**
Ours (GT layouts)	106.0	0.384	0.137
Ours (GT keyframes)	84.5	0.306	0.219

generation, and 37 fps for frame interpolation, achieving an inference speed of 25 fps, which is 10 times faster [243] than autoregressive and diffusion models.

5.4.1 Video Generation

We evaluate video generation results to verify the effectiveness of additional guidance and the capability of generating multiple coherent events in the videos.

Baselines. We compare with the following state-of-the-art video generation methods: 1) `MAGVIT (frame pred.)` [243]: given the reference frame as input, we apply MAGVIT to generate a 16-frame clip. We then use the last predicted frames iteratively to generate

the entire video. This follows the sliding window approach for long-take video generation. 2) `MAGVIT (class cond.)`: we condition MAGVIT on the reference frame and the object label. This extends the sliding window approach to incorporate additional guidance, similar to our method. 3) `StyleGAN-V` [179]: we train their unconditional model to generate a 64-frame video and apply the GAN projection method to condition on the reference frame. 4) `TATS` [57]: similar to MAGVIT, we generate a 16-frame clip and use a sliding window approach. 5) `Ours (GT layouts)` and `Ours (GT keyframes)`: to understand how the quality of layout and keyframe generation affects video results, we compare two variants of our method that use the ground truth layouts and keyframes as inputs.

Quantitative results. Table 5.1 shows our method achieves substantial improvements of 9.5% and 7.6% over baselines in FVD. The videos closely resemble ground truth, confirmed by LPIPS. AP and CLIP scores validate our model's success in generating videos with specified object inputs. Consistency scores verify maintained coherency, demonstrating our approach's efficacy. `MAGVIT (class cond.)` outperforms `MAGVIT (frame pred.)`, highlighting the benefits of additional guidance. Further, our approach surpasses `MAGVIT (class cond.)`, showing the combined effectiveness of additional guidance and our system, which generates videos by predicting keyframes and infilling the intermediate frames. `Ours (GT layouts)` and `Ours (GT keyframes)` significantly enhance video quality, suggesting room for improvement in long-take video generation despite using the same video generation model. This underscores the importance of developing video generation systems alongside enhancing base model quality. Our multi-stage approach, generating videos via keyframes and guidance, complements existing efforts and improves addressing real-world video generation challenges. Our method also enables user refinement of intermediate representations, like detailed layouts, to enhance generated videos interactively.

The per-frame LPIPS scores are shown in Figure 5.4. The X-axis represents the frame number, ranging from 0 to 64, while the Y-axis shows the LPIPS score averaged across all test videos for a specific frame. Image quality degrades rapidly in MAGVIT, particu-

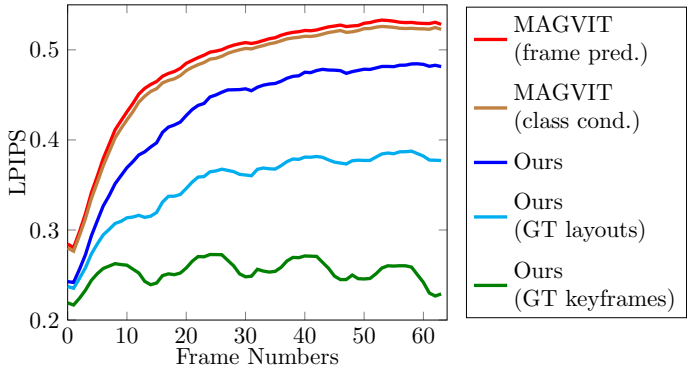

Figure 5.4: **Per-frame results.** The results of MAGVIT degrade over time, while our method shows slower quality degradation.

Table 5.2: **User study.** We report the percentage of raters who prefer our method based on video quality and fidelity to the ground truth.

Methods	Quality	Reproduction
Ours *v.s.* MAGVIT (frame pred.)	76.3%	82.1%
Ours *v.s.* MAGVIT (class cond.)	68.4%	66.7%

larly when generating videos beyond the length of the training clip (i.e., 16 frames). In contrast, our approach exhibits slower quality degradation, highlighting its effectiveness in generating videos covering diverse content.

User study. We conduct a user study to complement the quantitative evaluation. In this study, participants are presented with two videos generated by different methods alongside the ground truth video. They are then asked to evaluate the videos based on two criteria: 1) `Quality`: which video has better visual quality, and 2) `Reproduction`: which video better reproduces the content of the ground truth video. The study includes 40 videos and 11 participants using the EPIC Kitchen dataset. The results, presented in Table 5.2, align with the quantitative findings and further validate the significant performance gain of our method compared to the baselines.

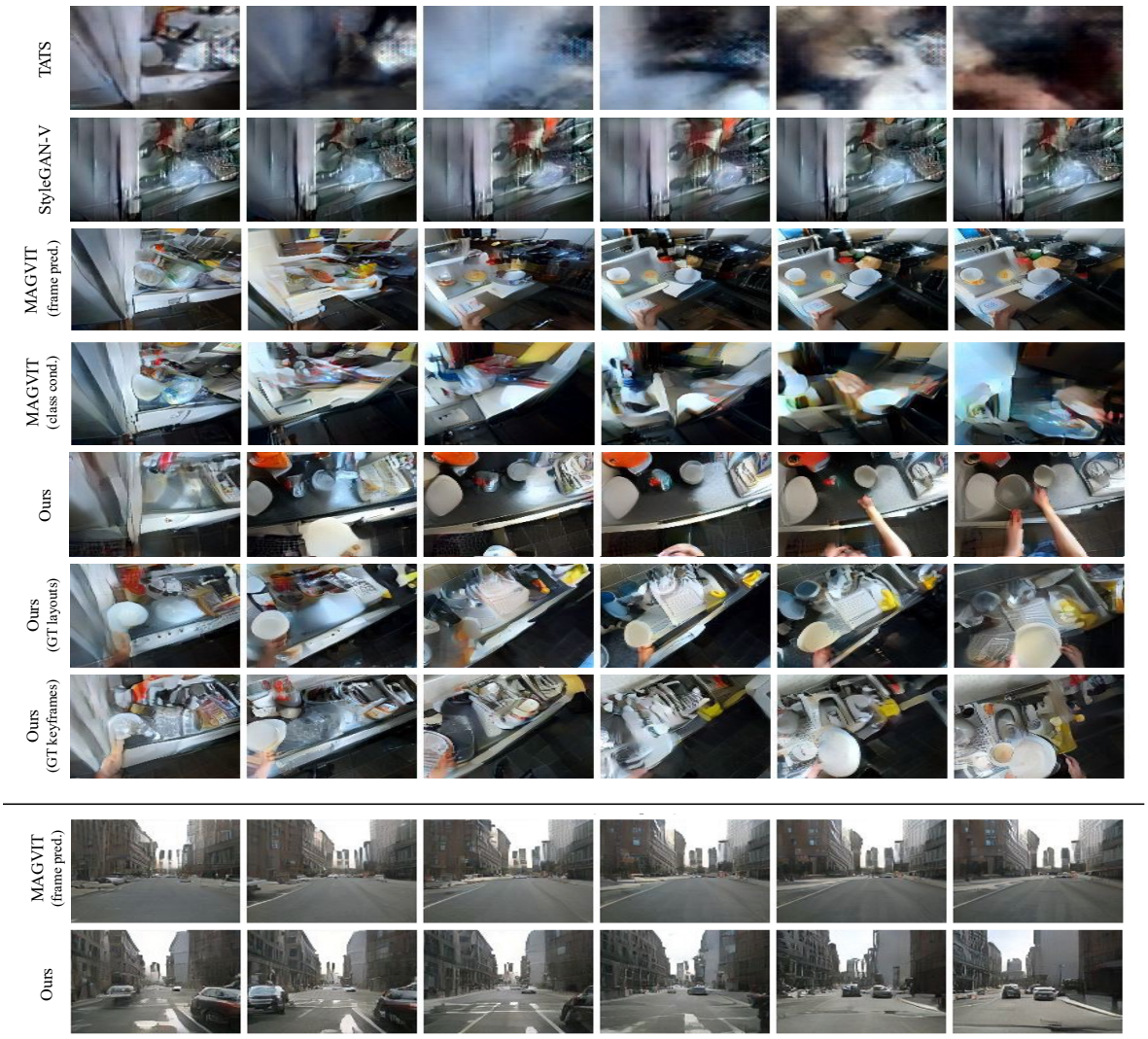

Figure 5.5: **Video generation on EPIC Kitchen (top) and nuScenes (bottom).** TATS, StyleGAN-V, and MAGVIT predict relatively static videos with repetitive patterns and quality degradation when inferring videos beyond the model's output length. In contrast, our method generates videos with non-recurrent events, including scene changes, object insertion, and deletion, maintaining meaningful content.

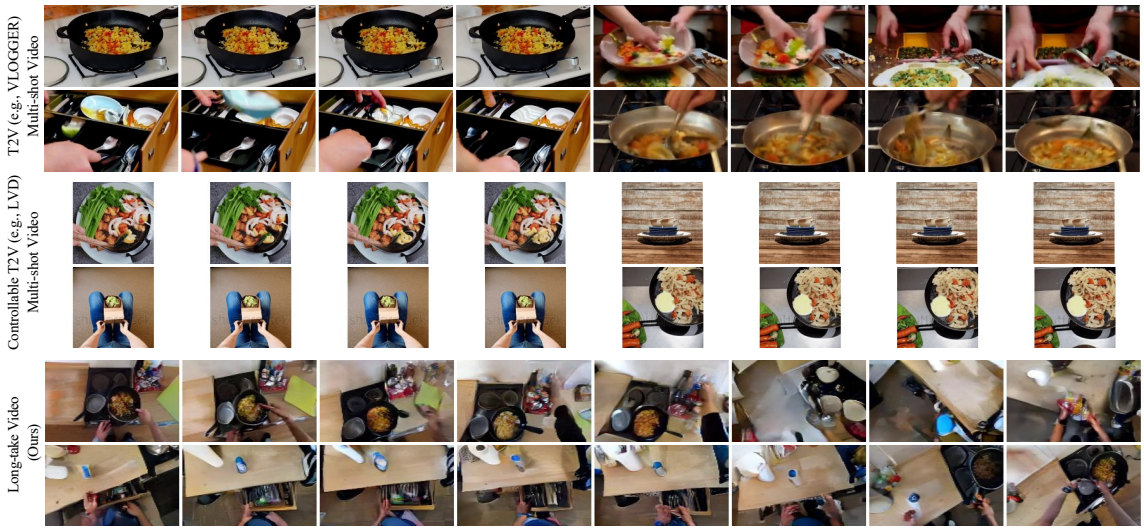

Figure 5.6: **Comparison to concurrent works.** Our method generates a single-shot long-take video with multiple non-recurrent events, such as cooking, obtaining condiments, and opening drawers. Video frames are presented sequentially from left to right and top to bottom. In contrast, multi-scene video generation VLOGGER and LVD [122, 125, 266], creates multiple snippets without direct temporal continuity.

Qualitative results. In Figure 5.5, `TATS`, `StyleGAN-V`, and `MAGVIT` tend to produce videos with homogeneous content, with quality deteriorating for long sequences. In contrast, our method generates videos featuring scene changes (e.g., from wall to table), object deletions (e.g., plate at the bottom), and object insertions (e.g., left/right hand), demonstrating the model's ability to generate videos with multiple events. Additionally, `Ours` generates videos that match the label sets but with different layouts than `Ours (GT layouts)`, validating our model's capability to generate videos satisfying different levels of input guidance, i.e., object label sets or (more constrained) layouts. Overall, we achieve satisfying results for the challenging video generation problem, where objects move in and out of camera views with significant motion and changing scenes.

Comparison to concurrent works. T2V and controllable T2V models create impressive

Table 5.3: **Quantitative results of keyframe generation.** Metrics are averaged across the keyframes. Our method consistently outperforms alternative approaches.

Methods	FID ↓	LPIPS ↓
MaskGIT	46.9	0.633
HCSS	50.2	0.653
Ours	**29.9**	**0.548**
Ours (GT)	24.2	0.416
Ours single (GT)	27.5	0.480

visuals within a scene but typically produce short clips (e.g., 32 frames) and struggle with complex sequences. VLOGGER [266] and VideoDirectorGPT [125] extend T2V models to generate longer videos, but focus on generating multiple snippets without direct temporal continuity in the adjacent shots. We address the challenge of generating single-shot, long-take videos up to 512 frames. By applying LVD [122] and VLOGGER [266], we generate long videos using a series of text prompts. In Figure 5.6, LVD and VLOGGER create multiple distinct shots, while our model produces a long-take video with non-recurrent events (e.g., cooking, fetching condiments, opening drawers). Our method complements multi-scene T2V models and can be integrated into VLOGGER's pipeline to create long takes as part of multiple snippets.

Limitations. The maximum video length our system generates is constrained by the number of keyframes, and the output length of the interpolation model. To generate longer videos, our approach can be integrated with multi-step inference processes to iteratively generate keyframes conditioned on preceding keyframes. As shown in Figure 5.6, up to 512 frames are generated using 32 keyframes. In addition, the video tokenizer from MAGVIT limits the proposed method, and using a better tokenizer such as MAGVITv2 could reduce flickering artifacts. We plan to train our models on caption-video data and swap the backbone to T2V models to generate videos in broader domains.

Figure 5.7: **Qualitative results of keyframe generation.** MaskGIT predicts similar content, while HCSS fails to generate consistent frames. Our superior results validate the importance of providing additional input guidance and modeling the keyframes holistically.

5.4.2 Keyframes Generation

We evaluate our keyframe generation model to assess the importance of jointly generating all keyframes and the effectiveness of masked attention and layout generation.

Baselines. We compare the following methods: 1) `MaskGIT` [22]: the model takes the reference as input and iteratively predicts the next keyframe, representing keyframe generation without input guidance. 2) `HCSS` [94]: the model takes a single layout as input and generates a single keyframe, representing independent keyframe generation without full video modeling. 3) `Ours`: our keyframe generation given the predicted layouts as inputs. 4) `Ours (GT)`: our keyframe generation using the ground truth layouts (upper bound). 5) `Ours single (GT)`: an iterative approach that predicts each keyframe conditioned on the previous keyframe and the ground truth layout, representing generation without full video

Table 5.4: **Ablation study.** We validate the effectiveness of the proposed masked attention module and layout generation stage in synthesizing high-quality and accurate keyframes.

(a) Masked attention

	5 objects			14 objects		
Methods	FID ↓	LPIPS ↓	AP ↑	FID ↓	LPIPS ↓	AP ↑
Ours w/o mask	33.6	0.584	0.183	34.8	0.575	0.216
Ours	**31.3**	**0.565**	**0.192**	**29.9**	**0.548**	**0.223**

(b) Layout generation

Methods	FID ↓	LPIPS ↓	AP ↑
Ours w/o layout	60.3	0.672	0.095
Ours	**29.9**	**0.548**	**0.223**

modeling.

Quantitative results. In Table 5.3, our method consistently outperforms `HCSS`, confirming the importance of joint prediction for all keyframes. `MaskGIT` tends to predict repetitive keyframes with minimal changes across frames, suggesting that the resulting video may appear static and unsuitable for video generation, highlighting the importance of guidance. We also compare different variants of our method. The superior performance of `Ours (GT)` over `Ours single (GT)` further verifies that a model considering the entire video jointly leads to better keyframe generation.

Qualitative results. Figure 5.7 shows that `MaskGIT` tends to generate keyframes with similar content, validating the importance of providing input guidance at multiple timesteps. In contrast, `HCSS` fails to generate consistent results across keyframes, as evidenced by variations in sink styles. Similarly, the iterative approach `Ours single (GT)` exhibits inconsistencies compared to `Ours (GT)`. These examples emphasize the significance of joint modeling for the entire video.

Ablation study. We conduct ablation studies on the masked attention module and the

layout generation process. The proposed masked attention module enhances coherence and efficiency in generating long videos. We analyze the module's impact using different numbers of input objects. As shown in Table 5.4 (a), removing the masked attention module and allowing each layout token to attend to every keyframe token leads to a decrease in performance. With masked attention, our model scales effectively to videos with up to 14 objects, generating keyframes that better align with the reference frame and contain more details, resulting in improved FID and LPIPS scores. The improvements in AP score further confirm its effectiveness in concentrating generation within specific bounding box areas.

The layout generation process helps preserve the 2D frame structure, which can easily deteriorate over time during long video generation. In Table 5.4 (b), we conduct an ablation study using object labels instead of the generated layouts as inputs to the keyframe generator. The FID and LPIPS scores increase without the layouts, validating the importance of the layout generation process.

5.5 Conclusions

This work addresses the challenge of generating long-take videos that encompass multiple events. We introduce a video generation system that utilizes multiple guidance inputs to control the generation process. A multi-stage approach has been developed to generate core events through keyframes and connect these events using existing video generation models. Empirical results validate the effectiveness of our approach, emphasizing the importance of guidance in improving video generation quality.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we explore how to improve the effectiveness of generative models in achieving consistency, controllability, and complexity, with a focus on applications in novel view and video synthesis. Specifically, we make the following key contributions.

Multiview consistency for novel view synthesis. In Chapter 2, we present a point cloud-based approach to enforce multiview consistency in 3D scene stylization. We first construct a point cloud by back-projecting image features into 3D space. We design point cloud aggregation modules that extract and integrate style information. The features are then modulated via a linear transformation matrix and projected back into 2D to synthesize stylized novel views. By performing stylization on constructed 3D representations, our method naturally achieves consistent stylization across views without requiring per-scene finetuning.

Subject-level controllability for video generation. In Chapter 3, we present a video generation model that enables controllability over both the trajectory and appearance of

individual subjects. Our approach incorporates a multimodal conditioning module composed of a joint encoder, control-attention layers, and an appearance augmentation mechanism. These design choices enable the model to generate videos that accurately reflect user inputs across different modalities in a single training pass, while maintaining the quality of the base model.

Modeling complex human motion for video generation. In Chapter 4, we present a human motion generation method conditioned on a scene image and a text prompt. The resulting motion sequences facilitate the synthesis of complex human actions for video generation. Our approach is built upon a diffusion transformer model with in-context conditioning. The model is trained on a large-scale video dataset that we collected, in which each video is annotated with its corresponding human motion. The proposed method effectively generates human motions that align well with the scene background, enabling the synthesis of long and complex human-centric video content.

Long-take video generation with consistent content. In Chapter 5, we propose a framework that uses multiple guidance signals to generate long-take videos with diverse yet coherent events. The generation process is divided into two stages: keyframe generation, which focuses on producing a sequence of semantically consistent events, and frame interpolation, which synthesizes smooth transitions between keyframes. We introduce a novel mask attention module to further enhance coherence and computational efficiency. Experimental results show that our approach effectively produces long-take videos with consistent content.

6.2 Future Work

We have explored four frameworks that improve different aspects of generative models for novel view and video synthesis in this thesis. Next, we discuss potential directions for

future exploration.

6.2.1 Real-time Style Transfer for Dynamic Scenes

In Chapter 2, we explore the task of transferring artistic styles from a reference style image to multiple novel view images of a static scene. Our results show that the stylized novel views maintain view consistency in 3D space. However, the model's performance heavily relies on the quality of 3D point cloud estimation provided by COLMAP and is limited to static 3D scenes. In the future, I plan to explore real-time style transfer for dynamic scenes or space-time videos. Given a set of images captured from different viewpoints and at different timesteps, along with a reference style image, the goal is to generate a stylized space-time video with consistent style transfer effects, all in a single feed-forward pass.

Recently, feed-forward models for real-time static 3D scene reconstruction [233] have gained increasing attention. These models take multi-view images of a static scene as input and predict 3D Gaussian parameters. At inference time, they can produce real-time 3D Gaussian reconstructions in a single forward pass, which can then be used for novel view synthesis. Building on this rapid progress, researchers have begun developing feed-forward frameworks for 4D reconstruction of dynamic scenes [248, 123], which aim to estimate the geometry of dynamic, deformable scenes across multiple timesteps. By training on large-scale datasets of static and dynamic scene videos captured from varying poses and times, these models reconstruct 3D Gaussian splatting or point cloud representations for each timestep in real time. These methods not only advance scene reconstruction capabilities but also open up exciting directions for scene editing, manipulation, and visual effects such as style transfer. I plan to extend 4D reconstruction models to support style transfer in dynamic scenes.

6.2.2 Learning to Generate Human-Scene Interactions from Videos

In Chapter 4, we develop a model that learns to generate human motion conditioned on a text prompt and a background image. Trained on our collected video datasets, the model produces diverse and complex human motions that are highly compatible with the given scene. However, the model relies on a fixed 2D image to represent the scene geometry, limiting its ability to capture dynamic human-scene interactions and the broader 3D environment present in videos. In future work, I plan to extend our human motion generation framework to model explicit human-scene interactions, conditioned on 3D representations of the environment extracted from monocular videos. This would enable more realistic and physically grounded motion generation in complex, dynamic environments.

Prior work on scene-aware human motion generation [256, 77, 75, 196, 197] has shown promising results in generating physically plausible human poses. However, these approaches are often trained on limited datasets, due to the challenges of acquiring paired 3D human and scene data at scale. More recently, researchers have started leveraging large-scale 2D in-the-wild video data to learn human motion [155], but learning human-scene interactions in 3D directly from video remains underexplored. Given the abundance of 2D video sources available online, I aim to collect human-scene interaction data in 3D by applying recent advances in 4D scene reconstruction [248] and human mesh recovery [235]. This data will serve as a foundation for training a motion generation model conditioned on 3D scene context, allowing us to better understand and synthesize physically plausible human behavior in realistic 3D environments.

Bibliography

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 47, 50

[2] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European Conference on Computer Vision (ECCV)*, 2020. 8

[3] 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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 48

[4] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 49

[5] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 9

[6] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *International Conference on 3D Vision (3DV)*, 2022. 50

[7] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 27

[8] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2020. 69

[9] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 33

[10] Amir Bar, Roei Herzig, Xiaolong Wang, Gal Chechik, Trevor Darrell, and Amir Globerson. Compositional video synthesis with action graphs. In *International Conference on Machine Learning (ICML)*, 2021. 67, 70

[11] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 27

[12] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*, 2016. 32

[13] 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. xi, 26, 64, 67, 70

[14] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 24, 25, 26, 27, 30, 35, 37, 38, 67, 70, 75

[15] Navaneeth Bodla, Gaurav Shrivastava, Rama Chellappa, and Abhinav Shrivastava. Hierarchical video prediction using relational layouts for human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 70

[16] Tim Brooks and Alexei A Efros. Hallucinating pose-compatible scenes. In *European Conference on Computer Vision (ECCV)*, 2022. 51

[17] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos

of dynamic scenes. In *Neural Information Processing Systems (NeurIPS)*, 2022. 69, 70

[18] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 76

[19] Xu Cao, Weimin Wang, Katashi Nagao, and Ryosuke Nakamura. Psnet: A style transfer network for point cloud stylization on geometry and color. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 9

[20] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019. 51

[21] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023. 27

[22] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 69, 83

[23] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 9

[24] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[25] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stereoscopic neural style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 9

[26] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024. 26, 67, 70

[27] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv preprint arXiv:2312.04557*, 2023. 26, 67, 70

[28] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 24, 27

[29] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR)*, 2023. 33

[30] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 47, 50, 54, 57, 61, 62

[31] Xinghao Chen, Yiman Zhang, Yunhe Wang, Han Shu, Chunjing Xu, and Chang Xu. Optical flow distillation: Towards efficient and stable video style transfer. In *European Conference on Computer Vision (ECCV)*, 2020. 9

[32] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023. 67, 70

[33] Yutao Chen, Xingning Dong, Tian Gan, Chunluan Zhou, Ming Yang, and Qingpei Guo. Eve: Efficient zero-shot text-based video editing with depth map guidance and temporal consistency constraints. *arXiv preprint arXiv:2308.10648*, 2023. 27

[34] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Neural Information Processing Systems (NeurIPS)*, 2021. 76

[35] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. *arXiv preprint arXiv:2308.10079*, 2023. 27

[36] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 67, 69

[37] CMU. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. 49

[38] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. In *International Conference on Machine Learning (ICML)*, 2024. 70

[39] Gaoxiang Cong, Liang Li, Zhenhuan Liu, Yunbin Tu, Weijun Qin, Shenyuan Zhang, Chengang Yan, Wenyu Wang, and Bin Jiang. Ls-gan: Iterative language-based image manipulation via long and short term consistency reasoning. In *ACM International Conference on Multimedia (ACM MM)*, 2022. 70

[40] 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. 51, 52

[41] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal on Computer Vision (IJCV)*, 130:33–55, 2022. 76

[42] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 11

[43] Yingying Deng, Fan Tang, Weiming Dong, haibin Huang, Ma chongyang, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021. 6, 9, 14

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 69

[45] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *Conference on Robot Learning (CoRL)*, 2017. 76

[46] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W. Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 70

[47] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Neural Information Processing Systems (NeurIPS)*, 2023. 27

[48] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. xi, 27, 63, 64, 76

[49] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 69, 74

[50] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Neural Information Processing Systems (NeurIPS)*, 2016. 67

[51] Tsu-Jui Fu, Xin Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. SSCR: Iterative language-based image editing via self-supervised counterfactual reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 70

[52] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 27

[53] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 27

[54] Chang Gao, Derun Gu, Fangjun Zhang, and Y. Yu. Reconet: Real-time coherent video style transfer network. In *Asian Conference on Computer Vision (ACCV)*, 2018. 9

[55] Wei Gao, Yijun Li, Yihang Yin, and Ming-Hsuan Yang. Fast video multi-style transfer. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 6, 9, 14, 19

[56] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[57] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision (ECCV)*, 2022. xi, 67, 68, 69, 70, 75, 78

[58] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 24, 26, 67, 70

[59] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations (ICLR)*, 2024. 25, 27, 70

[60] 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. 49

[61] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 51

[62] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 48, 52

[63] Xinyu Gong, Haozhi Huang, Lin Ma, Fumin Shen, Wei Liu, and Tong Zhang. Neural stereoscopic image style transfer. In *European Conference on Computer Vision (ECCV)*, 2018. 9

[64] Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985. 13

[65] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 49, 52

[66] 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 *European Conference on Computer Vision (ECCV)*, 2022. 50

[67] 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 International Conference on Multimedia (ACM MM)*, 2020. 49, 57

[68] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 25, 26, 27, 67, 70, 75

[69] Agrim Gupta, Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Characterizing and improving stability in neural style transfer. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 9

[70] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 26, 67, 70

[71] Tewodros Habtegebrial, Varun Jampani, Orazio Gallo, and Didier Stricker. Generative view synthesis: From single-view semantics to novel-view images. In *Neural Information Processing Systems (NeurIPS)*, 2020. 5

[72] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*, 2023. 27

[73] David Hart, Bryan Morse, and Jessica Greenland. Style transfer for light field photography. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 9

[74] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *Neural Information Processing Systems (NeurIPS)*, 2022. 67, 70

[75] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 50, 89

[76] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 49, 52

[77] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 50, 89

[78] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 27, 67, 70

[79] Jonathan Ho. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 54

[80] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 24, 26, 67, 70, 75

[81] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Neural Information Processing Systems (NeurIPS)*, 2020. 53

[82] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *International Conference on Learning Representations Workshop (ICLRW)* , 2022. 26, 69

[83] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations (ICLR)*, 2023. 24, 26, 67, 70

[84] 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. 46, 49

[85] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 27

[86] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 9

[87] Hsin-Ping Huang, Hung-Yu Tseng, Hsin-Ying Lee, and Jia-Bin Huang. Semantic view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[88] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 32, 33

[89] Nisha Huang, Yuxin Zhang, and Weiming Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. *arXiv preprint arXiv:2305.05464*, 2023. 27

[90] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 47, 57, 61, 62

[91] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[92] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2014. 49

[93] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015. 52

[94] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 83

[95] Michal Jancosek and Tomas Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 11

[96] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Neural Information Processing Systems (NeurIPS)*, 2024. 50

[97] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 37

[98] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. 8, 13

[99] Nal Kalchbrenner, Aäron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning (ICML)*, 2017. 69

[100] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 46, 49

[101] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Neural Information Processing Systems (NeurIPS)*, 2020. 52

[102] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 9

[103] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 51, 76

[104] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 24, 25, 26, 27, 67, 70

[105] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 70

[106] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 27

[107] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 9

[108] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4), 2017. viii, xiii, 7, 9, 14, 15, 19, 20

[109] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: Bidirectional layout transformer for controllable layout generation. In *European Conference on Computer Vision (ECCV)*, 2022. 69

[110] Johannes Kopf, Michael Cohen, and Richard Szeliski. First-person hyper-lapse videos. *ACM Transactions on Graphics (TOG)*, 33:1–10, 07 2014. 11

[111] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM Transactions on Graphics (TOG)*, 39(4):76–1, 2020. 7

[112] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 51

[113] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 27, 32, 36

[114] Truc Le and Ye Duan. Pointgrid: A deep network for 3d shape understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 9

[115] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 9

[116] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 7, 8, 11, 12, 13, 14

[117] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Neural Information Processing Systems (NeurIPS)*, 2018. 9

[118] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[119] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Neural Information Processing Systems (NeurIPS)*, 2017. 8

[120] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 70

[121] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 27, 30, 37, 40

[122] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. xii, 68, 81, 82

[123] Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, and Jiahui Huang. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. *arXiv preprint arXiv:2412.03526*, 2024. 88

[124] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 27

[125] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. xii, 68, 81, 82

[126] 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 *Neural Information Processing Systems (NeurIPS)*, 2023. 49, 52

[127] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. *arXiv preprint arXiv:2012.09855*, 2020. 5

[128] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Neural Information Processing Systems (NeurIPS)*, 2020. 8

[129] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 25, 27

[130] Zhenhuan Liu, Jincan Deng, Liang Li, Shaofei Cai, Qianqian Xu, Shuhui Wang, and Qingming Huang. Ir-gan: Image manipulation with linguistic instruction by increment reasoning. In *ACM International Conference on Multimedia (ACM MM)*, 2020. 70

[131] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 2015. 54

[132] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 26, 67, 70

[133] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 27

[134] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 70

[135] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. 70

[136] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 49

[137] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 9

[138] Joanna Materzyńska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2023. 39

[139] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2015. 67

[140] Moustafa Meshry, Dan B. Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[141] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 4, 8

[142] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003. 13

[143] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. CCVS: Context-aware controllable video synthesis. In *Neural Information Processing Systems (NeurIPS)*, 2021. 69

[144] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1–8, 2019. 51

[145] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 27

[146] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (TOG)*, 38(6):1–15, 2019. 8

[147] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Neural Information Processing Systems (NeurIPS)*, 2015. 67

[148] OpenAI. Chatgpt, 2024. 62

[149] 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. 55, 58

[150] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 27

[151] 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 *Neural Information Processing Systems (NeurIPS)*, 2024. 51

[152] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 48, 55

[153] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018. 55

[154] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 57

[155] Huaijin Pi, Ruoxi Guo, Zehong Shen, Qing Shuai, Zechen Hu, Zhumei Wang, Yajiao Dong, Ruizhen Hu, Taku Komura, Sida Peng, and Xiaowei Zhou. Motion-2-to-3: Leveraging 2d motion data to boost 3d motion generation. *arXiv preprint arXiv:2412.13211*, 2024. 89

[156] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, dec 2016. 49, 52

[157] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 9

[158] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Neural Information Processing Systems (NeurIPS)*, 2017. 9, 13

[159] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 25, 27

[160] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv preprint arXiv:2305.12328*, 2023. 27

[161] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 67, 70

[162] 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 *International Conference on Machine Learning (ICML)*, 2021. 55, 58

[163] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 13

[164] 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 Transactions on Visualization and Computer Graphics (TVCG)*, 2023. 49

[165] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 7, 8, 14

[166] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4, 5, 8, 14

[167] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 26, 30

[168] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 25, 27, 32, 33, 36

[169] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 67, 69

[170] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal on Computer Vision (IJCV)*, 2020. 67, 69

[171] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 11

[172] Johannes Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 11

[173] Falong Shen, Shuicheng Yan, and Gang Zeng. Neural style transfer via meta networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[174] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[175] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 27

[176] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Neural Information Processing Systems (NeurIPS)*, 2019. 70

[177] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 11

[178] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023. 24, 26, 67, 70

[179] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 67, 69, 70, 78

[180] Yun-Zhu Song, Zhi-Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. Character-preserving coherent story visualization. In *European Conference on Computer Vision (ECCV)*, 2020. 70

[181] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 70, 76

[182] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[183] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning (ICML)*, 2015. 67

[184] Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 8, 14

[185] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 47, 50, 53, 54, 55, 57, 61, 62

[186] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 69

[187] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[188] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 67, 69

[189] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning (ICML)*, 2016. 8

[190] Joost van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-based models of video sequences. *arXiv preprint arXiv:1701.08435*, 2017. 67

[191] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Neural Information Processing Systems (NeurIPS)*, 2017. 69

[192] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *International Conference on Learning Representations (ICLR)*, 2023. ix, 24, 25, 26, 28, 33, 34, 67, 69

[193] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *Neural Information Processing Systems (NeurIPS)*, 2022. 69

[194] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Neural Information Processing Systems (NeurIPS)*, 2016. 67, 69

[195] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 27, 70

[196] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 49, 50, 89

[197] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 50, 89

[198] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 24, 26, 67, 70

[199] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 46, 49

[200] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 26, 67, 70

[201] Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 25, 27

[202] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. Consistent video style transfer via compound regularization. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020. 6, 9, 14

[203] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2024. 26, 67, 70

[204] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Neural Information Processing Systems (NeurIPS)*, 2023. 25, 27, 30, 31, 35, 40, 76

[205] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 26, 67, 70

[206] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Neural Information Processing Systems (NeurIPS)*, 2022. 47, 49, 50, 52, 57, 61, 62

[207] 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. *Transactions on Machine Learning Research (TMLR)*, 2022. 48

[208] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 27, 30, 37, 40

[209] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations (ICLR)*, 2020. 69

[210] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8, 13

[211] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[212] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 24, 26, 67, 70

[213] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision (ECCV)*, 2022. 24, 26, 67, 70

[214] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 24, 25, 27

[215] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 9

[216] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 9

[217] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 27

[218] Saining Xie, Sainan Liu, Zeyu Chen, and Zhuowen Tu. Attentional shapecontextnet for point cloud recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 9

[219] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 27

[220] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 26, 27, 67, 70

[221] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 70

[222] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 25, 38

[223] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 27

[224] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 46, 49

[225] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 27

[226] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 49

[227] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 69

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

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 57

[229] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 69

[230] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM Transactions on Graphics (TOG)*, 2024. 35, 40

[231] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2023. 27

[232] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 27

[233] Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 88

[234] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 27

[235] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 89

[236] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 9

[237] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over diffusion for eXtremely long video generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 67, 70

[238] Jaehoon Yoo, Semin Kim, Doyup Lee, Chiheon Kim, and Seunghoon Hong. Towards end-to-end generative modeling of long videos with memory-efficient bidirectional transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 67, 70

[239] Sunjae Yoon, Gwanhyeong Koo, Geonwoo Kim, and Chang D. Yoo. Frag: Frequency adapting group for diffusion video editing. In *International Conference on Machine Learning (ICML)*, 2024. 70

[240] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[241] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018. 57

[242] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 27

[243] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. *arXiv preprint arXiv:2212.05199*, 2022. 69, 75, 77

[244] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 72

[245] Wei Yu, Wenxin Chen, Songhenh Yin, Steve Easterbrook, and Animesh Garg. Modular action concept grounding in semantic video prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 67, 70

[246] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and M. Patel, Vishal. Scenecomposer: Any-level semantic image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 27

[247] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 26, 67, 70

[248] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 88, 89

[249] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 4, 5, 8, 14

[250] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 24, 27, 30

[251] 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. 50

[252] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 50

[253] Qihang Zhang, Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Towards smooth video composition. In *International Conference on Learning Representations (ICLR)*, 2023. 67, 70

[254] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7, 18, 19

[255] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. ix, 24, 27

[256] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 50, 89

[257] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. PointWeb: Enhancing local neighborhood features for point cloud processing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 9

[258] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*, 2022. 49

[259] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 35, 37

[260] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Neural Information Processing Systems (NeurIPS)*, 2023. 27

[261] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023. 25, 27

[262] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 24, 26, 67

[263] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 7

[264] 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 *European Conference on Computer Vision (ECCV)*, 2024. 50

[265] 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 *European Conference on Computer Vision (ECCV)*, 2024. xi, 46, 49, 63, 64

[266] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024. xii, 68, 81, 82