DropletVideo: A Dataset and Method Exploring Integral Spatio-Temporal Consistent Video Generation

shuidi*
System Co.,Ltd.
Project Website: https://github.com/

Abstract

aaa
aaa
aaa

1 Introduction

Figure 1: Illustration of the problem of spatio-temporal consistency. (a) Only object motion consistency. (b) Only camera motion consistency. (c) More complex fusion of object motion and camera motion.

Video generation, as opposed to static image synthesis, accords greater importance to the dynamic variations within the frames and the associated challenges of maintaining temporal and spatial consistency. This involves two critical aspects. First, with respect to the **temporal consistency** of plot progression, it is necessary to ensure that the variations within the frames are coherent and adhere to physical laws, allowing the video to evolve plausibly according to the given storyline, as illustrated in Fig. 1(a). Second, concerning the **spatial consistency** across viewpoint changes, it is required that the appearance of objects remains consistent from different perspectives, including aspects such as shape, size, texture, and color, as depicted in Fig. 1(b).

Recently, numerous studies in video generation have begun to focus on the issue of consistency in visual transitions. A and B achieved significant progress in enhancing image quality and the plausibility of event transitions, thereby ensuring the coherence and accuracy of the generated video narratives. C and D considered camera movement and conducted in-depth explorations into the unification of objects across multiple viewpoints. However, most existing open-source works primarily focus on either temporal consistency or spatial consistency, with few addressing the integration of both, as illustrated in Fig. 1(c). Although research like E and F has demonstrated

*Corresponding author.

Preprint. Under review.

excellent performance in terms of spatiotemporal consistency, these models are closed-source, which limits the development of openly available alternatives and, consequently, hampers the potential for algorithmic innovation within the public community. This paper aims to address the spatiotemporal consistency issues in video generation from both the dataset and methodological perspectives, and introduces an open-source foundational video generation model named DropletVideo.

To support the pre-training of the proposed generation model, we construct the largest spatiotemporal consistency video dataset in the community, termed as DropletVideo-10M. The unique feature of this dataset lies in the simultaneous inclusion of both object motion and camera movement within the videos. Currently, most existing video datasets include only object motion, with some incorporating camera motion, but samples that simultaneously cover both are extremely rare. DropletVideo-10M fills this gap by including a large number of videos that feature both types of motion. Furthermore, the training of spatiotemporal consistency models necessitates captions that accurately describe both object motion and camera movement. Traditional video captions often only describe the scenery and plot, lacking detailed descriptions of motion, particularly the content variations caused by camera movement. DropletVideo-10M offers precise captions that encapsulate these details, effectively covering descriptions of camera movements and their resultant effects. The average caption length is 240 words, which significantly exceeds that of comparable datasets.

In addition, we propose an open-source foundational model for video generation, DropletVideo, designed to maintain spatiotemporal consistency during the video generation, enabling the generation of camera movement and plot progression simultaneously. Within DropletVideo, we also develope a variable frame rate sampling strategy, which allows for controllable video generation speed, thereby constraining the pace of visual transitions. Extensive experimental results validate the effectiveness of DropletVideo in preserving the consistency of generated content across both temporal and spatial dimensions. Moreover, visualization results indicate that, compared to other open-source models, DropletVideo exhibits greater potential in terms of three-dimensional consistency.

The contributions of this work are as follows:

- **First,** we explore the issue of integral spatiotemporal consistency in video generation for the first time and propose an open-source solution.
- **Second,** we have constructed the largest spatiotemporal video dataset to date, DropletVideo-10M, which is 50 times larger than the MVImageNet[47] multi-view image dataset.
- **Third,** we have designed and trained an effective foundational video generation model, DropletVideo, which meets both temporal and spatial consistency requirements. Experimental results demonstrate its significant potential in 3D consistency.
- **Last,** we have open-sourced the dataset, along with the code and weights for DropletVideo. Given that closed-source models like Sora[**?**], keling[19], and Pika[**?**] have already outpaced open-source models in generation capabilities, this is detrimental to algorithmic innovation in the public domain[**?**]. By releasing the dataset and model weights, we aim to stimulate innovation among researchers.

2 Related Work

2.1 Video-language Datasets

In order to facilitate the video generation task, particularly the text-conditioned video generation tasks, lots of video-language datasets [26, 3, 43, 39, 38, 5, 17] are proposed in recent years. For example, Panda-70M [5] proposed a large-scale dataset containing 70M video clips with automatic caption annotations. This dataset covers 166.8Khr with average 13.2 words. MiraData [17] proposed a high quality dataset with 788K videos containing detailed captions with an average 318.0 words. These datasets mainly deal with the temporal consistency during video generation, and seldom of them deal with the spatial consistency problem, where the length of caption need to be enlarged to afford the task, since new semantic information emerged after camera motion.

In order to deal with the spatial consistency problem during video generation, lots of multi-view images [30, 47] or video datasets [22, 48] are proposed. CO3Dv2 [30] and MVImageNet [47] mainly conclude object-level multi-view images, whereas DL3DV-10K [22] and Real-estate-10K [48] mainly conclude scene-level videos. However, most of the targets in the datasets are static and the data

quantity is far less than datasets that dealing with the temporal consistency problem. In addition, most of the videos in these dataset have limited frames and are usually used in multi-view image generation tasks rather than video generation tasks.

More recently, many more researchers starts to deal with the spatio-temporal consistent video generation problem and few datasets are proposed [15]. MV-Video [15] comprises about 115K animations that are available under a public license, consisting of about 53K animated 3D objects at all, which are rendered into over 1.8M multi-view videos. Comparing with them, we curate the world's largest Video-language dataset DropletVideo-10M, as shown in Table 1, to deal with the spatio-temporal consistency problem integrally. All videos in DropletVideo-10M involve camera motions and the quantity is 45× larger than the multi-view images dataset MVImageNet [47], and comparable with the large-scale video generation dataset Panda-70M [5]. Additionly, the average caption of our dataset is 240 words, which is nearly 20× in comparison with Panda-70M [5].

Table 1: Various video-language datasets.

	Words	Year	Clips	Avg dur.	Total dur.	Resolution	Category
HowTo100M [26]	4.0 words	2019	100M	3.6s	135Khr	240p	Temporal
WebVid-10M [3]	12.0 words	2021	10M	18.0s	52Khr	360p	Temporal
HD-VILA-100M [43]	17.6 words	2022	100M	11.7s	760.3Khr	720p	Temporal
InternVid [39]	32.5 words	2023	7M	13.4s	371.5Khr	720p	Temporal
HD-VG-130M [38]	~9.6 words	2024	130M	~5.1s	~184Khr	720p	Temporal
Panda-70M [5]	13.2 words	2024	70M	8.5s	167Khr	720p	Temporal
MiraData [17]	318.0 words	2024	788K	72.1s	16Khr	720p	Temporal
CO3Dv2 [30]	-	2021	36k	-	-	-	Spatial
DL3DV-10K [22]	-	2023	10K	-	-	-	Spatial
RealEstate-10K [48]	-	2023	10K	-	-	-	Spatial
MVImageNet [47]	-	2023	229K	-	-	-	Spatial
MV-Video [15]	-	2024	1.8M	2s	1Khr	512*512	Spatio-Temporal
DropletVideo-10M(Ours)	240.0 words	2025	10M	7.3s	20.4Khr	720p	Spatio-Temporal

2.2 Spatial-temporal Consistent Video Generation

Due to the high continuity and dynamic variability of video data, it is extremely challenging to directly generate videos with dynamic consistency in both temporal and spatial dimensions, which tends to result in the generated videos of the model being difficult to satisfy people's practical requirements.

Meanwhile, many video generation studies focus on temporal consistency. A. Blattmann et al. [4] introduces a high-resolution video framework using pre-trained Latent Diffusion Models (LDM). It adds a temporal dimension to the latent space and integrates learnable temporal layers, ensuring inter-frame alignment. Videofusion [25] proposes a decomposed diffusion model, separating spatial and temporal optimization to enhance cross-frame consistency. It employs time-aware latent representations and a hierarchical strategy, reducing temporal jitter effectively. Naturally, jointly considering spatiotemporal consistency has become a crucial challenge in generation tasks.

For spatial consistency, researchers initially propose a series of models based on the U-Net [23] architecture. He et al. [13] introduce CameraCtrl, enabling accurate camera pose control for text-to-video(T2V) models. It realizes the spatial control of the camera viewpoint in text-to-video generation. To enhance control, ObjCtrl-2.5D [41] is designed as an object control method without training. It employs 3D trajectories extended from 2D trajectories with depth information as control signals. ViewCrafter [46] merges point cloud rendering into U-Net as a conditional signal to achieve spatial consistency in video generation. Diffusion Transformer (DiT) [12] combines the advantages of Visual Transformer (ViT) and Diffusion Diffusion Model (DDPM), and gradually replaces U-Net as the dominant architecture in image generation tasks. Aiming at the problem that DiT suffers from a serious degradation of camera motion accuracy, Cheong et al. [7] explore and introduce camera motion guidance (CMG) and a sparse camera control pipeline. Based on ViT model, the VD3D [2] utilizes a spatio-temporal camera based on Plucker coordinates to achieve spatial controllability of video generation. DiT-based video generation methods have made progress in generating high-quality long videos [1, 40].

For 4D generation task, spatiotemporal consistency remains an essential and unavoidable research focus [16, 28, 21, 44, **?**]. Recent advances in video generation studies also have focused on spatial-temporal consistency. Singer et al. [35] lverages pretrained text-to-image diffusion models and

introduces pseudo-3D convolutional layers to extend temporal coherence without text-video paired data. ModelScope [37] proposes a hybrid structure combining spatial-temporal blocks and cross-frame attention to maintain multi-scale consistency. Qing et al. [29] explicitly disentangles spatial and temporal modeling through a two-stage framework, first generating keyframes then interpolating motions. Agrim et al. [11] enhances temporal alignment via a cascaded diffusion pipeline with optical flow-guided latent propagation. Chen et al. [6] addresses inconsistency through a training-free approach that unifies spatial and temporal attention controls in diffusion sampling. Although these studies address spatial-temporal consistency in video generation, they do not rigorously evaluate whether the generated videos truly exhibit consistent spatiotemporal characteristics, and they lack a well-defined benchmark.

2.3 Open-source Situation of Video Generation Models

Although large amounts of video generation models [18, 19, 24, 34, 32] are proposed, most of them are commercial closed-source models. Tencent's Hunyuan [18] generate diverse short-form videos and assist in film and television production by creating realistic virtual scenes and character animations. Kuaishou Keling [19] focuses on personalized 10 seconds video creation, leveraging user data to offer tailored templates and effects, and enabling easy creation through gesture and voice commands. Luma Dream Machine [24] combines deep learning and reinforcement learning to generate videos that reflect user emotions and intentions. Meta's Make-A-Video [34] explores the potential of text-to-video synthesis with a focus on scalability and accessibility. RunwayML [32] allows users to generate, edit, and transform videos using simple text-based instructions, bridging the gap between technical algorithms and creative workflows.

Except from these commercial models, some of the video generation models in community are open-sourced. However, few of them totally open-source their models, training data, codes and checkpoints all together, as shown in Table 2. Whereas, we open-source all the information about DropletVideo, hoping to raise the development of video generation technology and open up new possibilities for research and application in the field.

Table 2: Open situation of video generation models

	Model	Data	Code	Ckpts
CogVideoX	✓	×	✓	✓
DynamiCrafter	✓	✓	✓	✓
Animate-Anything	✓	×	✓	✓
I2VGen-XL	✓	×	✓	✓
SVD-XT-1.1	✓	×	✓	✓
OpenAI Sora	×	×	×	×
HunyuanVideo	partial	×	✓	✓
Keling	×	×	×	×
Luma Dream Machine	×	×	×	×
Meta Make-a-Video	×	×	×	×
RunwayML	×	×	×	×
Nvidia Cosmos	✓	×	✓	✓
DropletVideoGen(Ours)	✓	✓	✓	✓

3 Dataset

To facilitate the research on spatio-temporal consistent video generation, we decide to explore this problem from the dimension of dataset. We expect videos that have not only temporal changes but also spatial changes. In other words, videos should have camera motion and targets in the video should also changes overtime. We first select these videos from existing datasets for video generation, such as OpenVid-1M [27], Open-Sora-plan [20] and MiraData [17]. However, a large portion of videos from these datasets are static with no camera motion. What's more, the captions involved are not spatio-temporal aware. Regarding this, we decide to curate this kind of videos by our own.

Compared with stationary cameras, moving cameras typically offer a greater variety of angles and perspectives of the subjects or environments being captured, which illustrates more spatial change information. That means videos with camera movement contribute to the model's ability to learn more comprehensive information about different objects, while also mitigating the issue of monotonous

content generation. To this end, we decide to curate a video dataset that encompasses a diverse range of camera motion shots, which will illustrate more spatial information as well as the original temporal change information. We propose a novel dataset curation pipeline, particularly for collecting camera motion videos. Our pipeline contains four main parts, which are collecting raw videos from YouTube, video segmentation to obtain clips, clips filtering to obtain high-quality camera motion clips and generating spatio-temporal awared captions for clips. Pipeline of the curation of DropletVideo-10M is illustrated in Fig. 2.

Finally, we curated a spatio-temporal dataset with 10M high-quality videos, covering 2.21 Billion frames and have a 20.4K hour total video length. We made DropletVideo-10M open source in order to facilitate the research on spatio-temporal consistent video generation. Note that since the original videos in DropletVideo-10M are collected from the internet, they are only available for academic and non-commercial usage under the license of CC BY-NC-SA 4.0.

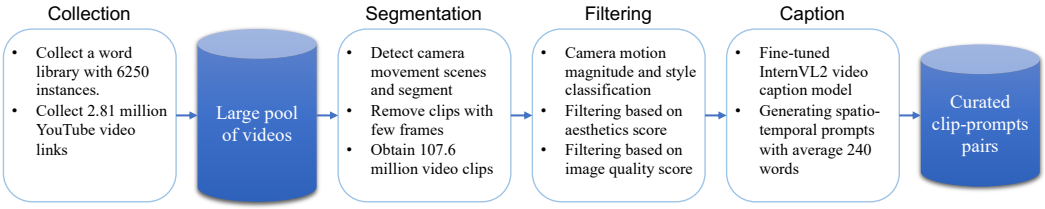

Figure 2: Pipeline of the curation of DropletVideo-10M: raw video collection, video segmentation, video clip filtering, video clip caption.

3.1 Raw Video Collection

We select the YouTube video website as the video source, since it is one of the world's largest video sharing websites, contains hundreds of billion videos covering virtually every topic and genre. The videos sourced from YouTube include a vast array of clips with camera movement, encompassing self-recorded videos, aerial shots, animations, gaming, and various other scenes. In order to obtain appropriate links to download, we decided to search to obtain the links. Firstly, we collected a word library that contains nearly 6250 instances existing in the world, resulting in 2.81 million links with average 450 links per searching word. Then, we download these links and obtain a raw video dataset with 2.81 million raw videos.

3.2 Video Segmentation

Videos sourced from the internet are often excessively long and do not consistently feature camera movement throughout. To address this, we implemented a series of processing steps to trim and select usable segments. We developed a detection program for camera movement scenes based on the prediction of optical flow between adjacent video frames, which was then used to crop and extract compliant segments from the videos. We considered the camera to be in motion between two frames only when the optical flow displacement exceeded a predefined threshold. Continuous sequences of moving frames were then excised to isolate segments of camera movement. This functionality was achieved by modifying the codebase of scenedetect [8]. Furthermore, we set an upper limit for the Euclidean distance of optical flow between adjacent frames to avoid scene transitions and cuts, thereby ensuring the continuity of the samples. Ultimately, we obtain 107.6 million video clips from 2.81 million raw videos, with 38.29 clips per raw video.

3.3 Video Clip Filtering

Clips segmented varies in content and camera motion, and they have different aesthetics quality and image quality. Therefore, we conduct three filters to filter high-quality video clips.

Firstly, the detected video clips vary in camera motions, and we claim that only clips with camera motions benefit for spatio-temporal consistent video generation. In addition, we detail the kind of camera motions according the magnitude and style of the motion, by observing large amounts of

video clips. Four detailed camera motion categories are concluded which are camera surrounding or target self-rotating, camera locally tilting horizontally or vertically, camera move following targets, camera linearly moving. Video clips with static or nearly static camera motions should be filtered out from the initial video clips dataset. In addition, since the videos in YouTube have a large proportion which are created or edited from video software, these kind of video clips should also be filtered out. We manually labeled a dataset with 20 thousand video clips for training. Video swim transformer is selected as the video classification model for its accurate classification performance. After training, we obtained an classification accuracy of 0.89. We selected clips belonging to the first four categories to constitute the spatio-temporal awared video dataset.

Secondly, we also utilize the public LAION aesthetics image model [33] to compute the aesthetics score and DOVER-Technical score [42] to compute the image quality in order to further improve the data quality. By setting appropriate thresholds, we obtain 10 million high-quality video clips, named DropletVideo-10M, at last. We conclude the aesthetics distribution and the image quality distribution of DropletVideo-10M, as shown in Fig. 3. We can see that nearly 90% clips in DropletVideo-10M are above 3.5, and nearly 78% clips are above 4.0, which demonstrates DropletVideo-10M's high quality.

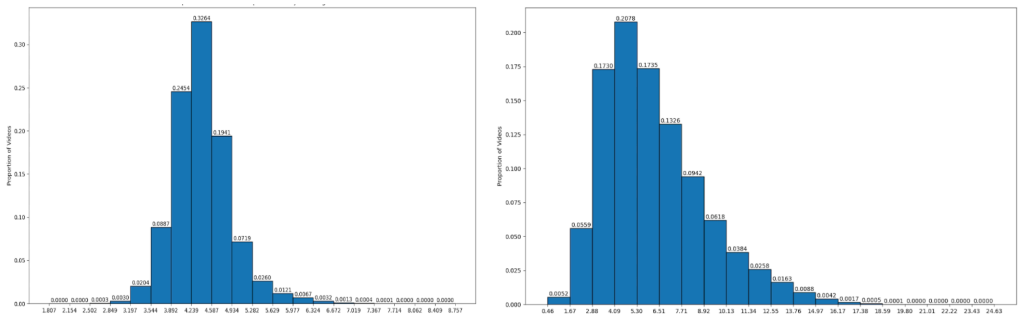

Figure 3: The aesthetics distribution and the image quality distribution of DropletVideo-10M.

3.4 Video Captioning

Existing video caption models usually generate captions with few words, which could not support the validation of the ability of spatio-temporal consistency in video generation. Therefor, we decide to construct our own video caption models for clips in DropletVideo-10M. Through carefully investigation and experiments, we selected InternVL2 [36] as the base model, and fine-tuned a video summarization model based on InternVL2 with LoRA, with the objective of generating detailed descriptions for video samples that include various transitions, camera movements, and content shifts due to lens changes, thereby providing more precise information for model training. Figure 4 presents two complete samples, clearly showing that the captions encompass abundant details of camera operations and captures the visual transitions induced by camera movement. Furthermore, we require that the descriptions contain adequate details regarding the lighting, style, and atmosphere of objects or backgrounds, thereby ensuring that the model receives sufficient guidance during the training process. We supply captions of an average length of 240 words for each video segment.

4 The DropletVideo Architecture

4.1 Preliminary

Diffusion Formulation. The core of the diffusion model(DM)[9] is that the generation process is viewed as a gradual addition of noise. Its overall structure is based on a Markov chain that is divided into a forward diffusion process and a reverse noise reduction process.

In the diffusion phase, the model adds noise sequentially according to the intensity until the original data is completely changed into Gaussian noise data. Given a real data distribution $x_0 \sim q(x)$, and it is sampled T times to add Gaussian noise. The variation schedule of the noise is defined as a_t, and the

This video showcases an abandoned ship moored in a tranquil sea, surrounded by lush green vegetation and a rocky coastline. The footage is captured from a high-altitude vantage point, revealing the detailed structure of the ship and its surroundings.
As the video begins, **the camera zooms in on an old**, rusted ship with a brown hull, tilting towards the shore. The ship is equipped with two long masts, devoid of sails. The ship is encircled by clear seawater, which reflects the sunlight in a sparkling array, displaying varying shades of blue and green. With the movement of the camera, the rocky coastline to the left of the ship comes into view, lined with green vegetation and scattered with small stones. The shoreline extends into the distance, meeting the sea. Throughout the video, the ship remains stationary as **the camera gradually pulls back to reveal the broader environment**. To the right of the ship lies an open expanse of sea, calm and serene, with the faint outlines of other ships visible in the distance.
The entire video conveys a sense of tranquility with a touch of desolation ,contrasting the ship's dilapidation with the vitality of the natural surroundings.

This video depicts a fantastical forest scene, where a small figure dressed in white is seen walking through a lush green forest.
At the beginning of the video, **the camera focuses on the depths of the forest**, revealing a small figure in white moving from the right to the left side of the screen. Surrounding him are dense green plants, including tall trees and low shrubs. In the background, sunlight filters through the leaves onto the ground ,creating a serene and mysterious atmosphere. As the video progresses, **the camera slowly pans to the left**, with the small figure continuing to walk forward, revealing more details of the trees and vegetation in the background. There are also some large mushrooms with vibrant orange and red colors, adding a splash of brightness to the scene. In the latter half of the video, **the camera continues to move left**, with the figure gradually exiting the frame, while the forest landscape in the background becomes even clearer. It is evident that there are rocks and moss-covered boulders, which add to the natural beauty of the forest.
The entire video, through its slow camera movement, evokes a sense of exploring an unknown world, offering a tranquil and mysterious journey through the forest.

Figure 4: Samples for model pre-training. The average length of theses captions is significantly longer than that of other existing datasets, containing more details about objects and scenarios. Additionally, the text includes extensive descriptions of camera angle changes and the resulting visual shifts in the scene.

data thus sampled is denoted as x_t, where $t \in [1, T]$. The process obeys a Markov chain, and after a reparameterization trick, the model can directly obtain any intermediate state, and the sampling formula for x_t is $q(x_t) = N(x_t; \sqrt{\bar{a}_t}x_0, (1 - \sqrt{\bar{a}_t})I)$, where $\bar{a}_t = \prod_{i=1}^{t} a_i$.

In the reverse noise reduction phase, the model learns the real data distribution from the standard Gaussian noise $p(x_T)$, where $p(x_T) = N(x_T; 0, I)$. The noise reduction function $\epsilon_\theta(x_t, t)$ usually obtained by designing a U-Net network stacked by residual networks, and then the optimization objective is defined as $L_{DM} = E_{x_0,t,\epsilon_\theta \sim N(0,I)}[\|\epsilon_t - \epsilon_\theta(x_t, t)\|^2]$, where ϵ_t is the sampled noise at time t, which is used as the ground truth.

Latent Diffusion Model. To address the problem of high computational and resource consumption of traditional diffusion models in high-dimensional data generation, a series of latent diffusion models(LDM)[31] have been proposed. Variational autoencoder(VAE)[10] and generative adversarial network(GAN)[45] are the common autoencoders. In the LDM, a pre-trained perception compression model is introduced, which consists of an encoder ε and a decoder D. In this way, the diffusion process can be transferred from the high-dimensional pixel space to the low-dimensional potential space, thus allowing the model to learn in the potential representation space. The objective function of the LDM is $L_{LDM} = E_{\varepsilon(x_0),t,\epsilon_\theta \sim N(0,I)}[\|\epsilon_t - \epsilon_\theta(z_t, t)\|^2]$, where z_t is the output of the encoder.

Our proposed DropletVideo encodes textual prompts and video frames into the latent space according to the idea of LDM. Furthermore, we introduce the diffusion transformer (DiT) model, a generative model that combines the transformer architecture with the diffusion model. DiT replaces the denoising network (e.g., U-Net) in traditional diffusion models with the transformer framework, which allows the DropletVideo to better capture the global dependencies in the data and generate high-quality videos.

4.2 Design Space

In this section, we design a motion-controllable video generation model, termed DropletVideo, as shown in Figure 5. During the training process, its input consists of a textual prompt, a video, and a motion-control parameter M. Text and image inputs are embeded into the latent feature space through the corresponding encoders, respectively. The text encoder is text-to-text transfer transformer(T5)[**?**] and the 3D causal VAE is applied for visual information. Subsequently, the potential features of these two modalities synergize the time parameter T and the motion parameter M , and then they are embedded into the motion adaptive generation module of vision and text, respectively. Finally, a new video is generated progressively from the noisy data by a 3D causal VAE decoder, which will satisfy the desired motion speed.

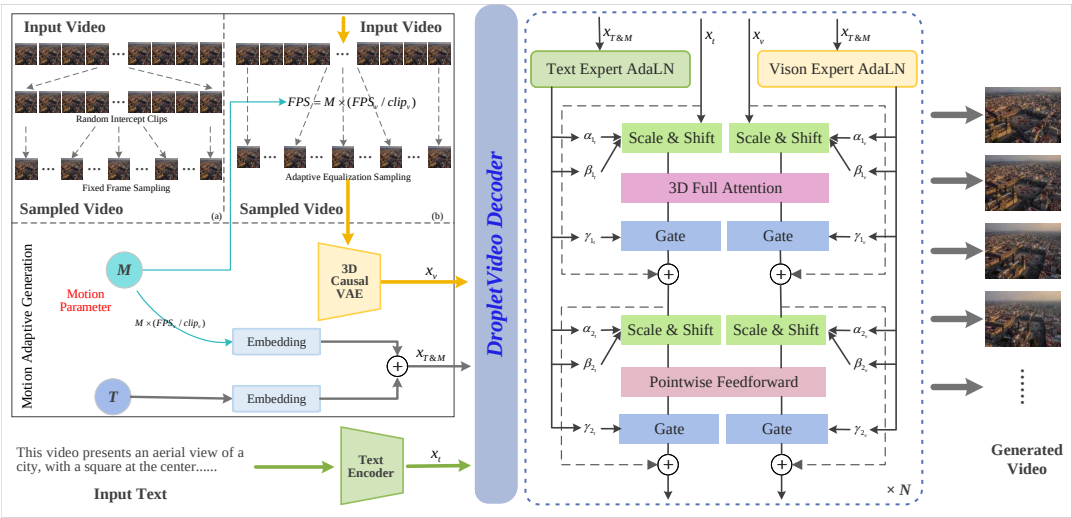

Figure 5: The overall framework of DropletVideo.

3D Causal VAE. Different from other auto-encoders, the outputs of VAE's encoder and decoder are subject to parameter-constrained probability density distributions. Yang et al.[**?**] applied 3D structures to video reconstruction. It is demonstrated that the 3D structure can reduce the jitter problem in the rebuilt video and improve the recovery quality. Therefore, in our DropletVideo architecture, we apply 3D causal VAE extended with VAE and 3D structure to the encoding and decoding of vision data.

Specifically, 3D causal VAE has two main roles in the DropletVideo model, spatio-temporal compression and disentangled representation learning, ensuring the efficiency and continuity of the generated video. Firstly, the video has both spatial and temporal dimensions, which results in a computationally intensive video generation task. 3D causal VAE performs multiple encodings on the original video, including spatio-temporal and spatial dimensions, and then DropletVideo obtains the potential space distribution with a 4×8×8 compression. The original video is expanded from the input image and random noise. The decoder is designed symmetrically with the encoder, which is used to implement the generative process of diffusion model, i.e., generating the video data by stepwise denoising the Gaussian noise from the latent space. Moreover, in real-world scenarios, videos with consecutive frames are causally correlated in the latent space, and based on this assumption, DropletVideo requires a new framework that supports causal disentanglement. 3D causal VAE can separate different generative factors in the data into separate dimensions of the latent representation. The learned latent factors can be forced to be as independent as possible through the regularization term of the kullback-leibler (KL) dispersion between the latent factor posterior and the standard multivariate Gaussian prior.

3D Transformer. The input for DropletVideo consists of two modalities, textual prompt and video. To ensure smooth embedding of each modality, 3D positional embedding is applied in the transformer architecture, and multimodal attention is employed to handle text and vision data simultaneously.

3D rotary position embedding (3D-RoPE) extends the rotation mechanism on a 3D sphere based on RoPE[**?**], which constructs positional coding by partitioning long sequences into blocks and setting the rotation angles within and between blocks. DropletVideo applies 3D-RPE individually to each dimension of the video tensor in the latent space and concatenates the resulting hidden states to generate the final coding. This design brings two benefits to video generation, controlling long-term decay and mitigating positional resolution degradation. The former prevents DropletVideo from losing the memory of earlier frames in the later generation stages, thus ensuring the overall consistency of the video. The latter enables accurate modeling of position information in the time dimension, thus generating smooth video.

3D full attention is a technique that has evolved with the widespread application of transformer in computer vision, e.g., in object detection[**?**] and point clouds[**?**]. We apply it to video generation and adapt it to the specific task of DropletVideo. Specifically, 3D full attention packs text and vision information with different sizes in one batch by frame pack technique, and performs attention computation on the fused new sequence along both temporal and spatial dimensions. Compared with the previous separation approach, it can better capture dynamic variations in the video and enhance the semantic consistency and diversity of the generated content. In addition, 3D full attention deploys a sparse attention mechanism that focuses only on localized domains or key frames, which enables parallel acceleration and efficient computation.

Motion Adaptive Generation. To generate videos with arbitrary motion speeds, we creatively propose the motion adaptive generation(MAG) strategy in our DropletVideo.

The generated videos from previous models usually have a fixed motion speed, mainly because they adopt a fixed frame rate to sample the raw data and deploy diffusion generation accordingly. It fails to meet the customers' requirements for more details on the video, and the user experience is poor. To this end, MAG is designed in DropletVideo to dynamically capture video frames through an adaptive algorithm and ensure that the generated video is motion-controlled. In addition, this strategy allows the sampled data to span the entire video stream, thus capturing global dependencies and obtaining more complete semantic information.

In the DropletVideo framework, the MAG strategy jointly modulates the input coding with time T. The T is a control parameter for the time step and noise level in the diffusion process, and is also used to guide the inverse generation. Since the feature states of the two input modalities, text and image, are quite diverse, we apply the expert adaptive layernorm strategies independently in the text and image latent spaces, which are vision expert adaptive layernorm (Vison Expert AdaLN) and text expert adaptive layernorm (Text Expert AdaLN). Through the auxiliary calculation of these two modules, the control parameter M is smoothly delivered to the diffusion transformer, and DropletVideo can precisely control the dynamics of the generated video data.

5 Experiments

5.1 Experimental Settings

5.1.1 Dense Prompt Rewrite

To effectively address the variability in language style and length of user-provided prompts, and to offer detailed guidance for video generation, we implement a dense prompt generation preprocessing step. This step serves as a bridge between the DropletVideo system and user input. Specifically, considering the superior performance of large language models in tasks such as text reasoning and image summarization, we have fine-tuned the InternVL2 [36] model with instruction tuning. This fine-tuning is done using the LoRA [**?**], utilizing caption pairs from a high-quality training set. Experimental results indicate that approximately 600 such samples are sufficient to achieve the desired level of fine-tuning.

The module is designed to rephrase user prompts while keeping their original semantics intact. It transforms them into a standardized information architecture, akin to the trained captions. The module parses plot and camera movement details from the user input. It expands the content based on the input image, ensuring that the user's intent is preserved and detailed information is added. Furthermore, the module offers support for multiple languages.

Finally, we rewrite 1118 standard prompts provided by vbench++ [14], and obtained 1118 long prompts which contains temporal and spatial changes.

5.1.2 Progressive Training

Videos from the Internet are varied, and to make full use of the data and save computational cost, we adopt a progressive approach to train DropletVideo. During the training process, we classify the dynamics range of the video viewpoints and select the class with the larger visual variations as much as possible. At the same time, we also evaluated the videos for aesthetics, quality, and transitions, aiming for diversity and reliability of the training data. Specifically, the training process of DropletVideo is divided into three parts. In the first phase, we download 10.73 million samples from the Internet with the aesthetic score of 3.75 and the quality score of 2.0. In the second phase, another 6.46 million better samples are selected from the previous phase's data based on the video scores to further refine the model. In the third phase for the precision of the generated videos, 3.43 million new high-resolution video data are acquired from the Internet. We select 0.51 million higher quality videos (aesthetics score>4.45, quality score>5.65) from them according to the scoring principle, and these videos together with 97,000 high-quality videos (aesthetics score>5.1, quality score>8.5) from the initial data in the first stage form the new training set.

5.1.3 Implementation Details Runze

We adopt xx for weight initialization and employ xx as the text encoder. We use xx as our optimizer, and the learning rates is set to xx. We sample video clips xx. All experiments are conducted on NVIDIA A100 80G GPUs. The training hour is xx.

5.2 Comparison with State-of-the-Art I2V Models

5.2.1 Qualitative Comparison

Dynamic Scene Generation with Object Consistency Maintenance. DropletVideo focuses on the temporal and spatial controllability during video generation. On one hand, DropletVideo prioritizes the camera movement capabilities in the generated content, aiming to avoid pseudo-videos with minor shakes or static frames. On the other hand, DropletVideo addresses the spatial distortion issues caused by camera movement, ensuring smooth plot progression during camera movement and the spatial consistency of objects within the scene. Figure **??** exemplifies the coexistence of camera movement and plot progression. It is evident that Droplet has the potential to simultaneously attend to the temporal and spatial controllability of videos.

Regulation of Video Generation Speed. DropletVideo regulates the speed of plot progression or camera angle changes by setting a motion control parameter. In the example provided, increasing this parameter allows a video of the same length to encompass more plot elements. Figure **??** presents the video generation outcomes under different motion control parameters with the same text-image input, demonstrating that DropletVideo can effectively control the playback speed of content while maintaining semantic accuracy.

3D Consistency. Videos generated by DropletVideo exhibit robust 3D consistency. As the camera pans and tilts, the video reveals projections of the same object from various angles. Figure **??** illustrates the video generation effects during camera movements including translation, rotation, and zooming, demonstrating the target object's strong consistency.

Comparison with Closed-Source Video Generation Models.

5.2.2 Quantitative Comparison

We compare DropletVideo with state-of-the-art image-to-video models, and quantitative results are

Figure 6: Results of Spatio-Temporal consistent video generation.

Figure 7: Illustration of the speed controllability of DropletVideoGen.

6 Conclusion

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *arXiv preprint arXiv:2411.18673*, 2024.

[2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al.

Figure 8: Illustration of the 3D consistency of DropletVideoGen.

Table 3: Comparison of DropletVideo with state-of-the-art image-to-video models.

Models	I2V Subject	I2V Background	Camera Motion	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
DropletVideoGen	98.51%	96.74%	37.93%	96.54%	97.02%	97.68%	98.94%	27.97%	60.94%	70.35%
I2VGen-XL	96.08%	94.67%	12.95%	95.76%	97.67%	97.40%	98.27%	24.80%	65.26%	69.21%
Animate-Anything	98.13%	96.05%	10.64%	98.18%	97.46%	98.15%	98.52%	2.52%	66.42%	71.89%
DropletVideo	98.55%	96.82%	36.72%	96.48%	96.88%	97.94%	99.06%	23.82%	60.08%	70.11%
I2VGen-XL	96.74%	95.44%	13.32%	96.36%	97.93%	98.48%	98.31%	24.96%	65.33%	69.85%
Animate-Anything	98.54%	96.88%	12.56%	98.90%	98.19%	98.14%	98.61%	2.68%	67.12%	72.09%
DynamiCrafter-1024	96.71%	96.05%	35.44%	95.69%	97.38%	97.63%	97.38%	47.40%	66.46%	69.34%
SEINE-512x320	94.85%	94.02%	23.36%	94.20%	97.26%	96.72%	96.68%	34.31%	58.42%	70.97%
ConsistI2V	94.69%	94.57%	33.60%	95.27%	98.28%	97.56%	97.38%	18.62%	59.00%	66.92%
VideoCrafter-I2V	90.97%	90.51%	33.58%	97.86%	98.79%	98.19%	98.00%	22.60%	60.78%	71.68%
SVD-XT-1.1	97.51%	97.62%	-	95.42%	96.77%	99.17%	98.12%	43.17%	60.23%	70.23%

Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021.

[4] A. 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023.

[5] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[6] Xuweiyi Chen, Tian Xia, and Sihan Xu. Unictrl: Improving the spatiotemporal consistency of text-to-video diffusion models via training-free unified attention control, 2024.

[7] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024.

[8] PySceneDetect Developers. Pyscenedetect. *https://www.scenedetect.com*, 2024.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[11] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 393–411, Cham, 2025. Springer Nature Switzerland.

[12] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2024.

[13] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

[14] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

[15] Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398*, 2024.

[16] Yanqin Jiang, Li Zhang, Jin Gao, Weiming Hu, and Yao Yao. Consistent4d: Consistent 360° dynamic object generation from monocular video. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024.

[18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[19] kuaishou. kuaishou-klingai. *https://klingai.kuaishou.com*, 2024.

[20] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan. *apr*, 2024.

[21] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024.

[22] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[23] Tianrui Liu, Qingjie Meng, Jun-Jie Huang, Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. Video summarization through reinforcement learning with a 3d spatio-temporal u-net. *IEEE transactions on image processing*, 31:1573–1586, 2022.

[24] lumalabs.ai. lumalabs.ai-dream-machine. *https://lumalabs.ai/dream-machine*, 2024.

[25] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liangsheng Wang, Yujun Shen, Deli Zhao, Jinren Zhou, and Tien-Ping Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, 2023.

[26] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[27] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[28] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv 2401.08742*, 2024.

[29] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to- video generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6635–6645, 2023.

[30] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[32] runwayml. runwayml. *https://app.runwayml.com/login*, 2024.

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[34] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[35] Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue 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. *ArXiv*, abs/2209.14792, 2022.

[36] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024.

[37] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *ArXiv*, abs/2308.06571, 2023.

[38] 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. 2023.

[39] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[40] Yuelei Wang, Jian Zhang, Pengtao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Cpa: Camera-pose-awareness diffusion transformer for video generation. *arXiv preprint arXiv:2412.01429*, 2024.

[41] Zhouxia Wang, Yushi Lan, Shangchen Zhou, and Chen Change Loy. Objctrl-2.5 d: Training-free object control with camera poses. *arXiv preprint arXiv:2412.07721*, 2024.

[42] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.

[43] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022.

[44] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion²: Dynamic 3d content generation via score composition of video and multi-view diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

[45] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[46] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[47] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023.

[48] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.