# Trajectory-Guided 3D-Aware and Identity Consistent Video Generation

*Longbin Ji*


Master of Science
School of Informatics
University of Edinburgh
2024

# Abstract

In recent years, the efficient application of reliable video generation in file making, game development and other fields has gained significant attention. Built upon these, trajectory-controllable video models have been recognized for their enhanced editing potential for temporal content through incorporating motion guidance from user-interactive trajectory input. However, current state-of-the-art image-based trajectory controlling models face challenges in handling 3D-aware movements such as object rotation following large-angle curve-based trajectory. This problem stems from the insufficient 3D motions presented in training dataset containing open-domain videos and lack of 3D guidance from 2D-based trajectory. To address these issues, our research focused on incorporating 3D-aware guidance to enable our trained model to generate corresponding 3D movements like rotation following given trajectory-curves. In this project, we present a self-constructed dataset compromising animated objects with sampled rotating process and propose a novel 3D-aware two-stage fine-tuning strategy through generating 3D bounding box sequence in the first stage as additional 3D prompt and integrate a spatial-enhancement loss during training to improve consistency of object's identity with changing poses. Through extensive experiments, our designed model achieves superior performance in animating rotating motion within our constructed dataset and demonstrates potential zero-shot animating capacity on open-domain videos. This result highlights the effectiveness of our designed pre-training dataset and methodology in facilitating 3D-aware animation guided from complex trajectories while maintaining consistent object appearance, accurate object position and poses.

# Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Longbin Ji*)

# Acknowledgements

Thanks for my supervisor Changjian. Li for his enthusiastic and efficient guidance on this project, as well as inspiring coaching on research ideas. I also want to thank Phd student Lei. Zhong for his selfless and active participation along the entire project and for providing me with extremely valuable suggestions.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Background

During recent years, substantial advancements have been made for the task of video generation, resulting in more natural and stable performance across generated frames on models including SVD [3], Gen-2 [14], VideoCrafter [10] and SORA [6]. These achievements can be contributed to the success of strong visual generative basemodels based on Latent-Diffusion [32] and the utilization of large-scale video datasets. Although these realistic video generation models hold great potential for various industrial applications, such as character animation and film production, their overall controllability for generated motion tendency and content remains a significant obstacle for effective usage. Most current methods are predominantly rely on text or image-based guidance, where the conditioning signals only provide spatial information, while the overall generation lacks more precise temporal motion message. To address fine-grained motion control, previous works have focused on trajectory-based animation, inspired by its interactive and user-friendly nature and the success of trajectory-controlling for image editing on feature latent space [29].

For trajectory controllable video generation, earlier works primarily focused on the limited domain of human-specific animation [4, 11] guided by skeleton movements. DragNUWA [46] presents the first attempt to realize trajectory controlling in open-domain videos. This model employs a flow estimator with a trajectory sampler strategy to extract both dense and sparse trajectories, and subsequently injecting encoded trajectory information into the video-based Denoising UNet. Another impressive work MotionCtrl [40] endeavors to decouple object movement and camera movement through a two-stage process: with an initial pre-training stage involving a camera-moving dataset,

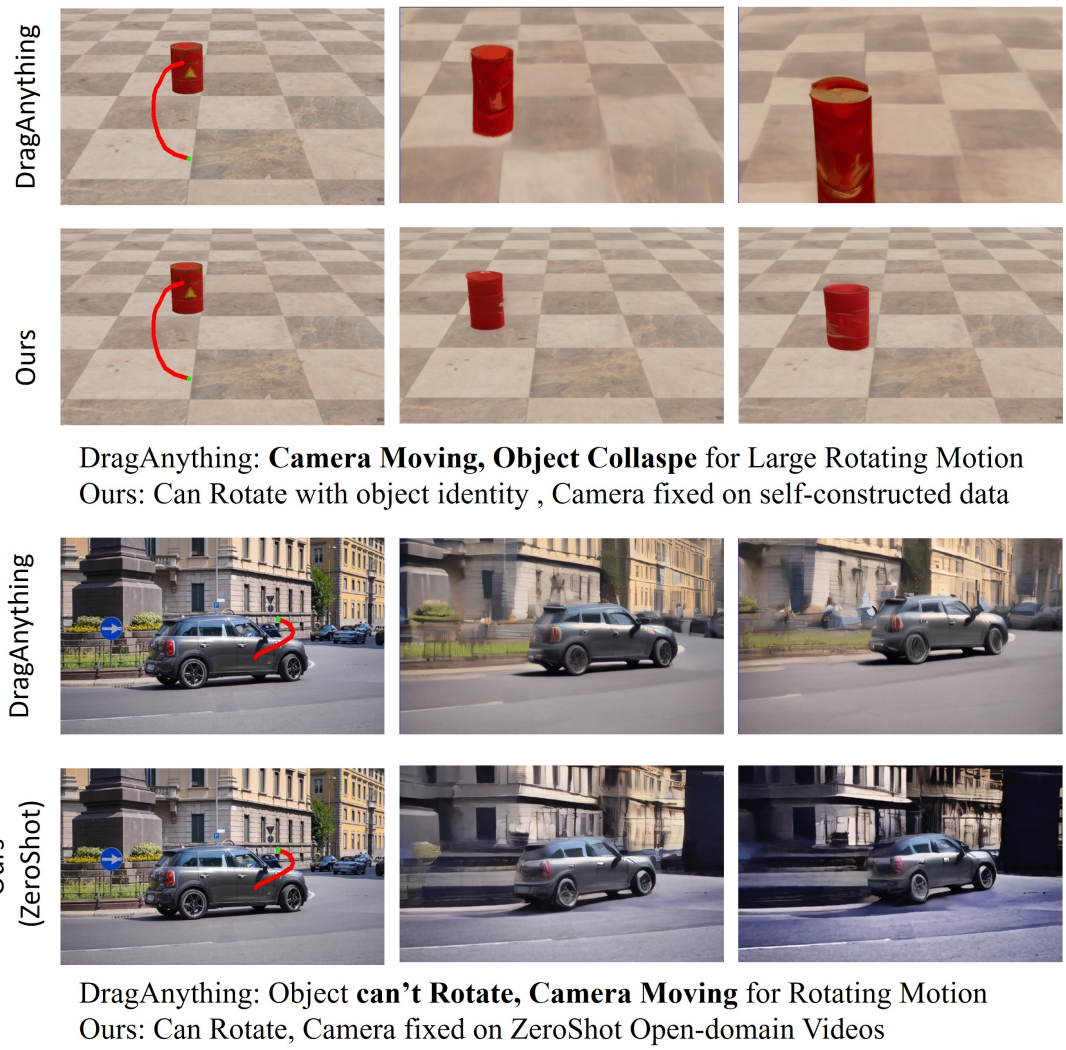


Figure 1.1: Visualization of the performance from sampled animated video given same trajectory from DragAnything and our proposed model trained in our self-constructed dataset. The objects can not rotate while the camera moves for DragAnything.

followed by second-stage finetuning with extra motion-control modules for wild videos. Built upon DragNUWA, DragAnything enhances model's trajectory awareness of object identity by extracting object embedding through Segment-Anything [25] with corresponding object mask and a pre-trained entity extractor[43]. This object-specific embedding is then fused with trajectory guidance conditioned on a Controlnet [47]. Through extra entity-guided signal, DragAnything achieves state-of-the-art trajectory animation performance with consistent object appearance and accurate controlling.

## 1.2 Motivation

Nonetheless, several challenges persist for effective trajectory controlling. As illustrated in our tested demos (figure 1.1) from DragAnything [43], current trajectory controlling ***still lacks 3D-awareness for rotating motions and are only accurate following simple trajectory, such as straight lines or basic arcs, while struggling with large and complex motions like large-angle rotation***. The primary reason of this issue of insufficient 3D understanding is that the encoded trajectory only contains 2D spatial information, while lacking 3D guidance for object pose during rotation. Furthermore, 3D movement like rotating is closely coupled with camera rotating, making effective training harder, as ***most open-domain videos do not contain sufficient scenarios for 3D movement like large-angle rotation***.

To tackle these problems, we propose the following research question: **Can we self-construct a 3D-aware trajectory-driven dataset containing only various rotating objects with complex trajectory like S-shaped arcs and pre-train a designed 3D-aware model with extra incorporated 3D guidance on it?** During following project development, we first successfully constructing a rendering dataset with rotating objects in different shapes following complex trajectory with large angle rotations. Subsequently, we select a reliable video generation basemodel, Stable-Video-Diffusion (SVD) [3], with superior trajectory controlling accuracy and temporal consistency in small-scale open-domain videos. Built upon SVD, for self-constructed dataset, we propose a 3D-aware fine-tuning strategy with trajectory-specific Controlnet by generating objects with their 3D bounding boxes during first-stage training, which significantly improves overall trajectory accuracy and spatial reconstruction quality for object appearnce during rotation.

Our project have following contributions:

- We proposed the first 3D-aware Video Generation model through two-stage fine-tuning process involving 3D bounding box generation. This model can maintain both driving trajectory accuracy and corresponding object appearance and pose during large-angle rotation, achieving superior FID, SSIM, FVD and trajectory accuracy on self-constructed 3D-rotating dataset as shown in Table 4.2 and Figure 4.4 compared to simply finetuning through trajectory conditioning as proposed in DragNUWA [46].

- We constructed the first dataset containing complex arcs or S-shaped trajectories with corresponding rotating object and realistic simulation of real-world environ-

ment. This dataset is rendered without camera-movement simplifying model's learning difficulty for camera decoupling. Moreover, we evaluate the importance of sufficient data scale for our self-constructed dataset as shown in table 4.5.

- We successfully evaluates the effectiveness of pre-training a 3D-aware video generation model on self-constructed dataset by demonstrating that our pretrained model can can generalize its trajectory animation capacity on unseen objects and even zero-shot unseen open-domain videos as shown in figure 4.4 and 4.5.

# Chapter 2

# Literature Review

## 2.1 Latent Diffusion Model

The Diffusion model for image generation, first proposed by [19], has achieved superior performance which beat Generative Advesial Networks (GAN) in terms of diversity and generation quality. Unlike GAN, which relies on a generative network with a competitive discriminator to distinguish between generated samples with true distribution, the diffusion model generates samples by predicting the added noise for each timestep during denoising process. To be specific, diffusion model contains forward and denoising stages. During the forward process, increasing noise following noise scheduler for each timestep is added to the target output, and this process is formulated following Markov Chain where each step $X_t$ is only dependent with previous step $X_{t-1}$. By repeatedly adding noise during forward process, the input distribution converges into a Gaussian Distribution, from which the denoising process can directly sample. For the denoising stage, the generator models the condition distribution through predicting the reversed noise given conditional signal. Given the predicted noise and $X_t$, sample $x_{t-1}$ in previous step can be calculated through combination of Bayes Rule. Compared to other generative model like GAN and auto-regressive models, diffusion model offer more diversity by directly modelling the generation of target distribution and recursively introducing sampled noise during each denoising step.

Although the naive diffusion model achieves impressive success in generating comprehensive and high-quality images, its denoising UNet-based generator faces intensive computational and memory cost with high-dimentional input, such as pixels in single image. Aiming at reducing the overall training cost, Latent Diffusion is proposed by compressing pixel-level information into a latent feature space with a pre-trained

Varational Auto-Encoder (VAE) [32]. Following Markov chain property, given noise scheduler $\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, the latent distribution $z_t$ in timestep $t$ can be calculated through one step equation as shown in 2.1. During denoising procedure, the conditional distribution of previous timestep $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, c)$ can be modelled by predicting reverse noise given latent target in current timestep and conditional signal $c$ following equation 2.2. By introducing a strong pre-trained VAE in large-scale unlabelled data and employing diffusion for the encoded latent feature, Latent diffusion is capable for both efficient model training with sampling and preserving satisfactory generation performance.

$$q(\mathbf{z}_t|\mathbf{z}_0) := N(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{2.1}$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, c) = N(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, c), \beta_t\mathbf{I}). \tag{2.2}$$

Further works boosts the generation performance of latent diffusion with more powerful AutoEncoders such as VQGAN [15] or denoising network like Scalable Diffusion Transformer (DIT) [30]. Concurrently, latent diffusion based generative modelling has been widely applied in various tasks and modalities like Text-to-Speech Synthesis [27], Human Motion Generation [24, 1] and Video Generation [10, 37].

## 2.2 Video Generation Models

While image generation tasks like text-to-speech synthesis has accomplished high-fedility performance for models such as DALL-E2 [31] and Imagen [33] with extensive large-scale image dataset and corresponding huge parameters in designed structure, video generation has only begun its evolution in recent years. Earlier method GODIVA [41] employs a 2D VAE with additional Sparse Attention to capture semantic guidance for text-to-video synthesis. Following it, NUWA [42] enhanced the quality of generated video by acquiring a unified representation across multi-modality learning guidance. CogVideo [20] presents the first work which finetuned a text-to-video model based on pre-trained frozen Text-to-Image weight with minimal training parameters. Make-a-Video [34] further extends CogVideo by efficiently adapting all model parameters with pesudo 3D convolutional and temporal attention layer as model backbone which significantly reduces training cost and produces superior temporal performance. Considering the intensive computational cost of pixel-level diffusion for multiple frames, several recent video diffusion models [5, 17, 23] employed latent diffusion with pre-trained

image encoder and are finetuned from text-to-image prior weights. With the collection of a significantly larger amount of high-quality videos through carefully designed data curation, more recent state-of-the-art models including VideoCrafter [10], Gen-2 [14], SVD [3] and SORA[6] have achieved realistic simulation of real world with high generation quality by fully-training their designed 3D backbone such as 3D UNet or Diffusion Transfer with inserted cross-frame attention and has supported multi-mode generation such as image-to-video extension facilitated by sufficient prior knowledge from large data.

### 2.2.1 Controllable Video Generation

Although current video diffusion models have achieved impressive performance in generating comprehensive and diverse videos, most of these base models purely rely on condition guidance of text descriptions or first-frame images. However, these conditional signals only provide spatial information during video generation, without more fine-grained control over content information especially temporal motion across frames. Consequently, many recent works are focusing on trajectory-based animation Compared to other motion representation, such as motion field or dense flow, trajectory provides straightforward object-centric motion guidance and also enable effective user-interaction by allowing users to directly hand-crafting lines on images. Early researches on trajectory-based animation such as IPoke [4] and MCDiff [11] only focused on limited-domain like human videos by extracting the movement of human skeletons and conditioning this trajectory information on generation model training from scartch. DragNUWA [46] is the first work to achieve trajectory animation in open-domain videos by employing a powerful pre-trained image-to-video generation model and fine-tuning it with extra trajectory guidance extracted from state-of-the-art flow estimator. However, for DragNUWA, objects' temporal consistency often collapses, and the trajectory guidance is not accurate for driving the whole object identity during animation. To tackle this problem, DragAnything enhances model's awareness of object's identity across frames with significantly better identity consistency through incorporating object entity by first segmenting out object's boundary with segmentor Segment-Anything [25] and then employ pre-trained spatial image encoder for entity extraction. Some other text-to-video based works like Direct-A-Video [45] and MotionCtrl [40] further attempts to disentangle between object motions from camera movement for more fine-grained and accurate controlling over for-ground and background and has successfully

integrate motion guidance along with camera poses and semantic prompt from text descripations.

Although the state-of-the-art image-based trajectory animation model DragAnything is capable of generating realistic and accurate motions with consistent object identity, it struggles with more complex and 3D-aware motions like rotation of objects or cars as shown in figure 1.1. This problem stems from the lack of 3D awareness in the pre-trained SVD model and is challenging to solve because collected wild videos usually do not contain purely rotation motions and 3D motions like object rotation is closely coupled with and influenced by frequent camera movements. Furthermore, trajectory in 2D pixel space is unable of providing extra 3D-based guidance. To solve this problem, our project aims to self-construct a rendered dataset containing rotating objects with driving trajectories and design a novel 3D-aware model structure to handle 3D complex motions, such as large-angle rotations, on this pre-training dataset.

## 2.3 Finetuning Large Pre-trained Models

Fine-tuning large pre-trained models, such as large language model, can achieve superior performance on customized data or different downstream tasks by adapting large model's understanding as prior knowledge. However, directly fine-tuning the entire network is computationally intensive and would introduce problems such as mode collapse or forgetting phenomenon for already learned knowledge. To tackle these problems, extensive researches has focused on developing more efficient and applicable fine-tuning methods. Among these, adapter is proposed in earlier methods by injecting a small number of nearly identity-initialized learnable layers into the original Transformer based network for NLP tasks [21]. Adapter fine-tuning has achieved impressive performance on transferring pre-trained NLP models into downstream tasks and has been extended successfully into Computer Vision field by adapting Vision Transfer [26] or Stable Diffusion on Controllable Image generation [16]. To further solve the catastrophic forgetting problem and limit the total amount of learning parameters, Low-Rank Adaptation (LoRA) [22] has been proposed. LoRA only optimizes weights in inserted low-rank matrices as a parameter residual to introduce new learning ability while preserving the original feature space compared to adapter finetuning. LoRA is widely used in influencing attention modules in modern text-to-image generative models with low training cost and has achieved notable controllability for tasks like style or identity controllable image generation. Although these fine-tuning methods

is capable of transferring the abilities of large-scale pre-trained model into various downstream task setting, they can not introduce precise control from new condition signals like human-poses or driving trajectories in our task setting. Aiming at enhancing spatial-consistent and task-specific condition control for pre-trained Text-to-Image (T2I) models, Controlnet [47] is proposed which clones a trainable-copy from the original network (e.g Unet's Encoder) with pre-trained parameters and integrate its output as learnable residuals upon the original network. To avoid potential negative influence from randomly initialized parameters, these residuals are connected through zero-convolutional layers where all the weights and parameters are initialized as zero. Controlnet has achieved impressive performance by adding extra sketch or pose condition during text-to-image generation and is also selected as fine-tuning method for our trajectory signal.

# Chapter 3

# Methodology

## 3.1 Dataset Construction

The majority of existing available video generation dataset [2, 7, 28], used for text-to-video, image-to-video or trajectory-controlled video generation are derived from open-domain videos. Although these datasets encompass comprehensive representation of natural object movements, the frequent carema movement introduces unexpected influences on object locations and poses. Futhermore, motions from object's movement may not be obvious for effective learning. Consequently, modelling objects' movements from trajectories for such unconstrained videos requires additional decoupling module between camera movement [40]. To address this issue, we have self-constructed a novel video dataset from rendering software blender, containing single moving object with randomly sampled trajectories, a static background and unchanged camera.
Our dataset offers several distinct advantages over open-domain videos: 1. It includes a wider range of paired complex 3D-aware movements with driving trajectory such as circular or S-shaped trajectories, as well as self-rotation process for single object. 2. All movements are purely object-centric, thus avoiding the unintended influence of camera movement and enhancing model's understanding of object-specific trajectory-based controlling. 3. The sampled trajectories and objects' bounding boxes are accurate, eliminating potential inaccuracies introduced by extra annotation tools, such as flow estimation or point tracking models. 4. The dataset size is scalable because we can sample infinite trajectories with random angle for each single object.

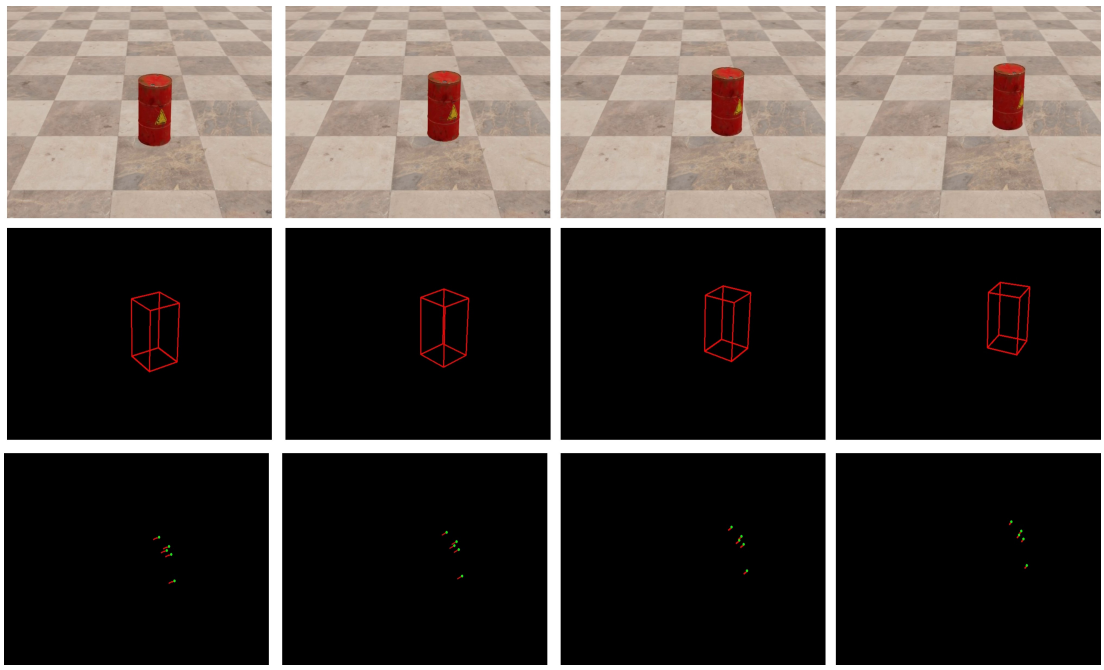

Figure 3.1: Visualization of one sample of video frames and corresponding trajectory for ”Barrel” in the self-constructed dataset with circled trajectory moving template.

### 3.1.1 Dataset Rendering

For detailed implementation, we employed Blender software for the dataset rendering pipeline as shown in 3.2. Firstly, we initailized the blender environment with floors in wooden texture and lighting. To simulate realistic lighting, we created environment light from indoor HDRI images instead of simple directional lights or sun lights. Following this, We sampled high-quality and realistic 3D models from public resource website [1] and then rendered them with textures and materials in the virtual environment. To maintain appropriate object size, each object's height is normalized below camera window's height with minimum 1 unit height. Subsequently, a random trajectory will be generated from pre-set movement templetes of ”circle” or ”S-shaped curve” with random rotating angles. Once generated, the rendered object was programmed to move according to this trajectory with fixed speed while changing its poses and facing direction towards the rotating center. At last, a camera is placed in suitable position and the moving procedure is recorded and rendered using the Cycles engine.

[1] https://polyhaven.com/

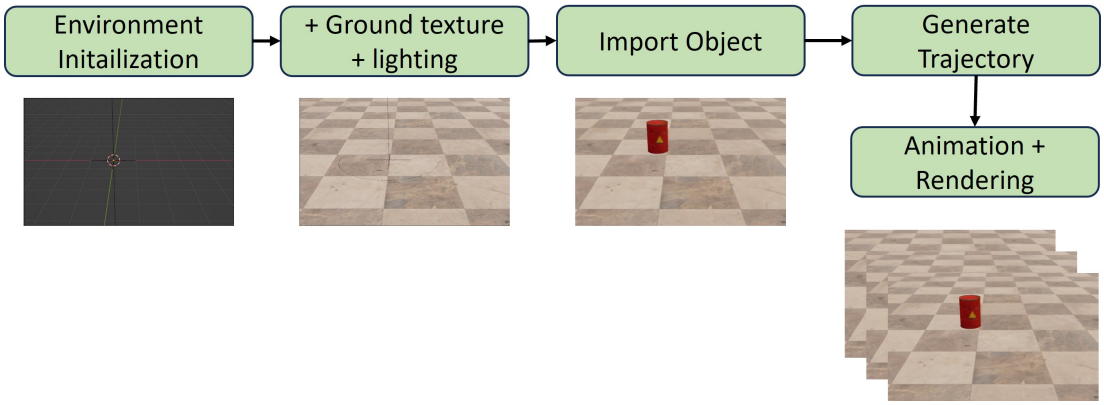


Figure 3.2: The data production pipeline of our self-constructed data through blender software.

### 3.1.2 Dataset Annotation

Along with the recorded video of objects' movements, we have saved the trajectory's parameters and object's 3D position, including location for 8 vertexs of 3D bounding boxes. All trajectories and bounding box positions are mapped to local pixel space given camera pose parameters. For trajectory representation, inspired by the annotation process of DragNuwa which sampled multiple sparse trajectories to enhance trajectory guidance and supports users' imprecise hand-crafted input rather than the precise one from object center, we employ a region-aware trajectory sampler. This sampler randomly select 1-5 points within the mapped bounding boxes and drawing the same trajectory trend from the saved trajectory parameters for these points. Through further experiments in section 4.5.4, the trajectory sampler improves the trajectory-based animation accuracy and overall generation quality.

## 3.2 Model Design

Our project's objective is to generate a sequence of video frames in length L $F = [F_1, F_2, ..., F_L]$ given starting frame $F_s$ and a corresponding trajectory sequence $J = [J_1, J_2, ..., J_L]$ that drives the object. Unlike previous methods that lack explicit controlling for more complex motion with object rotation, such as "S_shaped" trajectory, our proposed model emphasizes enhancing 3D-aware understanding, trajectory accuracy and object reconstruction quality during rotation. This is achieved through a designed two-stage 3D-aware fine-tuning method with an additional spatial enhancement loss as introduced in following sections. In this project, our focus is on training and eval-

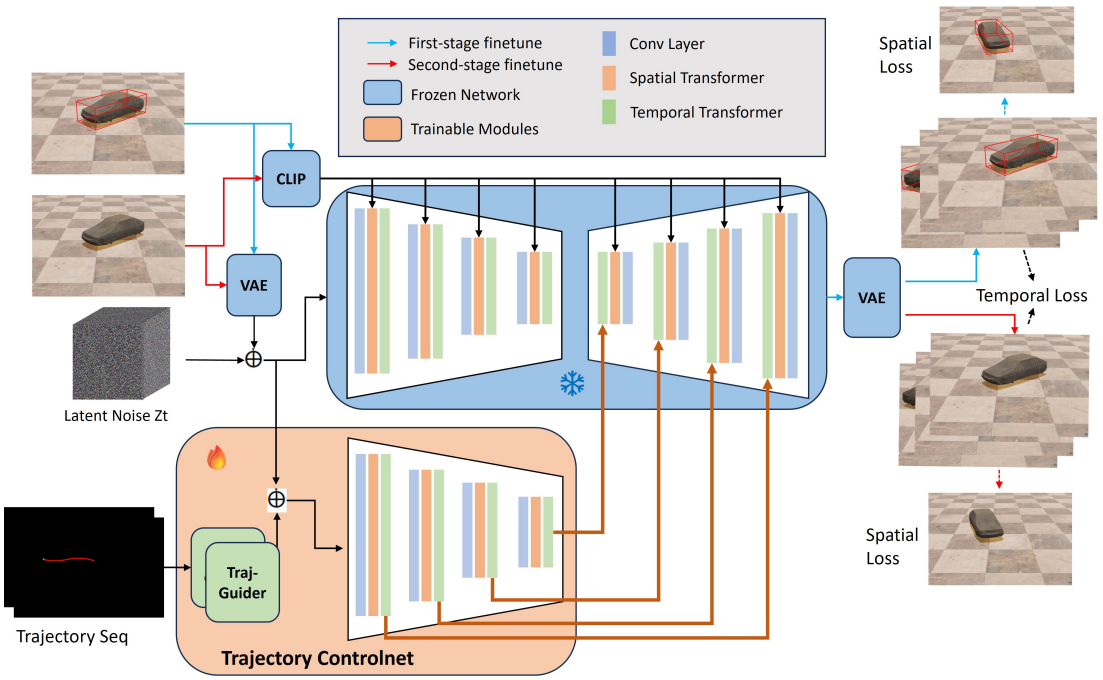


Figure 3.3: Visualization of our proposed 3D-aware model structure of trajectory-guided animation. It mainly composed of a trajectory-specific Controlnet, the designed two-stage finetuning strategy and spatial enhancement loss as explained in the paper.

uating the performance on self-reconstructed dataset containing rotating objects with corresponding trajectories. To ensure reliable video generating performance, following previous works, we utilize Controlnet for our designed 3D-aware finetuning on the state-of-the-art Image-to-Video model Stable-Video-Diffusion.

### 3.2.1 Stable Video Diffusion Model

For controllable video generation fine-tuning, it is crucial to select a robust and powerful large-scale pre-trained video generation model as baseline. Through performance comparison in section 4.3, we have chosen the state-of-the-art Image-to-Video generation model, Stable-Video-Diffusion (SVD) [3] as our base model. SVD has demonstrated superior temporal consistency and realistic generation compared to two-stage finetuing of text-to-image model. It also serves as the current basemodel for other state-of-the-art trajectory-animation model including DragNUWA and DragAnything. As shown in figure 3.3, SVD's denoiser is a 3D UNet comparmising spatial convolution layer, spatial attention layer and spatial temporal layers within each UNet block. The video is encoded by a pre-trained spatial VAE into latent feature sequences $L$. During training, the 3D UNet takes the latent feature $L_t$ with added noise from schdler in timestep t

combined with the time step embedding and encoded first frame feature from same VAE to predict the added noise $e_{t-1}$. During denoising, the condition provided to the 3D UNet is another semantic image feature from a CLIP-based image encoder, which provides direct guidance for the generated spatial information.

To achieve superior performance for reliable generation, as reported by the authors, the image-to-video based SVD is fine-tune from another pre-trained Text-to-Video model and this T2V model undergoes three-stage pre-training process. The first stage involves selecting a pre-trained Text-to-Image model - Stable Diffusion 2.1. The second stage further employs a video pre-training based on a carefully curated video dataset containing 152M samples with updated 3d UNet with cross-frame attention module for temporal modelling. During data curation, Videos with insufficient motion are filtered out and the remaining videos are divided into subsets for further performance scoring. The final stage applies high-quality fine-tuning, where the pre-trained weights are further trained at dataset in higher resolution of 320*576 with fewer high-fidelity videos. With this carefully designed training strategy and an extremely large video collection, SVD outperforms other popular video models including GEN-AI and Pikalabs and has also achieved superior performance on downstream tasks like image-to-3D generation and trajectory editing [43].

### 3.2.2 Trajectory-Specific Controlnet

In accordance with the proposed Controlnet for text-to-image finetuning [47], our Image-to-Video Controlnet creates a trainable copy of the Spatio-Temporal UNet's downsampling blocks from SVD model, incorporating zero-convolutional layers. The designed trajectory-specific Controlnet takes noised latent features encoded from image-based VAE, time step embedding, the encoded first frame and encoded Trajectory guidance to generate feature residuals derived from stacked UNet blocks. These residuals are then integrated into corresponding up-sampling blocks to inject trajectory-specific controlling signals. Through this training strategy, the fine-tuned model not only supports extra conditioning mode of trajectory animation but also retains its original video generative understanding in the frozen 3D UNet, which facilitates further zero-shot generability as shown in section 4.4.4

**Trajectory Encoding.** Through annotation and region-aware trajectory sampler, our trajectory is represented as a sequence of images in length L-1, where each image contains painted sparse sampled red directional lines or curves and green circles for

ending positions. A Trajectory-Guider composed of residual convolutional blocks with same downsampling rate as the SVD's UNet blocks, is employed for extracting trajectory guidance from the image sequences. The Trajectory Guider's convolutional only operates on image channels for spatial feature compressing without influencing for temporal cross-frame relationship and encoded frame length. Therefore, to ensure the encoded length matches the generated frames with length of L, a zero padding frame is added at the sequence end.

To save training memory requirement and accelerate the overall training speed, only the trajectory-Controlnet are trained in 32 bits, while other modules including Image-based CLIP encoder, VAE encoder and 3D UNet backbone, are frozen with 16 bits weights. Futhermore, the attention layers in each transformer block within the frozen UNet network are all replaced by flash-attention [12] layers for efficient training.

### 3.2.3 3D-Aware Finetuning of Stable-Video-Diffusion

According to our tested demos and authors' studies from DragAnything, it is evident that pre-trained SVD lacks 3D-awareness and generative ability for rotating objects and large rotating motions. Consequently, it is essential to provide extra 3D guidance for the trajectory-based fine-tuning process for SVD model. However, indicted from following experiments, merely injecting trajectory specific Controlnet for video finetuning in pixel space is uncapable of capturing 3D-aware movement like large-angle rotation and would result in serious identity collapse for target objects. This is primarily due to the fact that 2D trajectory only provides pixel-wise positional movement, while both object's pose and position undergo changes during rotation. In addition, the appearance of rotating object's also changes facing different directions, but original SVD's generation ability for continous frames lacks a clear 3D reconstruction capacity.

To enhance the Controlnet's 3D rotation-specific perceptivity, we propose a novel two-stage fine-tuning strategy by incorporating the generation of 3D bounding box as 3D guidance prompt. This is achieved as follows: First-stage fine-tuning accepts initial frame with 3D bounding box and requires model to generate 3D bounding box sequences for each animated frame along with the moving object as shown in figure 3.3. Subsequently, once the model is capable of generating accurate 3D bounding boxes, second-stage finetuning can seamlessly remove the bounding boxes and continuously improve the reconstruction quality of object's appearance. Comparing to directly modelling object-centric rotation pattern, 3D bounding bboxes not only provide more

apparent, reliable and learnable guidance during generation but also constrain the object's entity inside the bounding bbox during generation, preventing serious entity collapse especially for complex objects. By generating corresponding 3D bounding boxes, the model efficiently captures the precise rotating facing angles with the 3D object poses following given trajectory. Through following experiments 4.2, our two-stage fine-tuning significantly enhances video quality and trajectory accuracy with lower FID, FVD, Motion-Diff (reflects trajectory accuracy) and higher SSIM score as well as much better visualization performance.

### 3.2.4 Spatial Enhancement Loss

Through experiments, it has been observed that, although directly applying 3D-aware two-stage produces more accurate rotation movement following given trajectories, the reconstruction consistency for object's identity can not be maintained with potential collapse. This issue is particularly obvious for complex objects with asymmetric textures or complex geometric structures, as its identity may easily collapse for longer frames. This problem is reasonable because 1) geometric or asymmetric features are hard for high-quality reconstruction and 2) the averaged MSE loss may lack sufficient constraint for spatial reconstruction performance because it is calculated by averaging over temporal and spatial loss which couples model's temporal and spatial modelling ability together. To enhance the model's spatial awareness, as illustrated in 3.3, we employ an extra spatial loss module by randomly sampling one frame's latent noise $z_t$ with corresponding short trajectory representation between frame $0-l$ and generate single-frame object using same denoising UNet. The new loss function, as shown in equation 3.1, is calculated by adding original noise loss and spatial loss together with a weight factor $\alpha$. After inserting spatial-specific loss, model can stabilize for fixed background generation in early training stage with less frequent appearance collapse which improves overall trajectory-controlling performance as shown in experiments 4.3.

$$loss_{mse} = \alpha * loss_{spa} + loss_{temp} \tag{3.1}$$

# Chapter 4

# Experiments

In the beginning of experiments chapter, we first illustrates the limitations of current SOTA models through our tested demos. Then, we demonstrate our employed evaluation metrics for clarification purposes. Subsequently, it comes to our formal experiments, we first select a proper and powerful baseline for video generative fine-tuning in a toy open-domain video dataset and then perform our main experiment with our proposed two-stage 3D-aware fine-tuning with performance metrics. As the training code of state-of-the-art trajectory-based animation models including DragNUWA, DragAnything and MotionCtrl are not publicly avaliable (motionctrl release its training code near the end of project submission deadline). We self-implemented two methods to compare with our proposed method: Traj-Disco and Fine-tuning SVD. Traj-Disco is adapted from Disco for SOTA pose-guided video animation as shown in figure 4.2 and fine-tuning SVD directly incorporating a trajectory encoded Controlnet to finetune the pre-trained SVD similar to DragNUWA's implementation which is also our implementation baseline. Detailed explanations are provided in section 4.3.1. To further demonstrate and analysis for our choices for model and dataset design, we further conducted several ablation studies.

## 4.1 Limitations of Current Trajectory-Guided Video Models

To evaluate the controllability of current state-of-the-art Motion-Controllable models. We selected two state-of-the-art models: MotionCtrl and DragAnything with inferenced demos based on it.

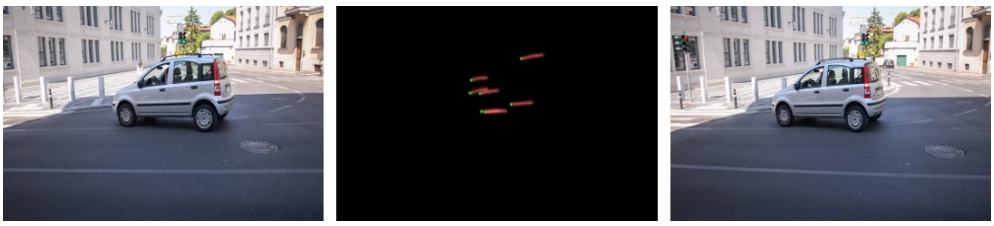

Figure 4.1: Annotation of trajectory in DAVIS dataset for baseline selection training.

**MotionCtrl**: MotionCtrl adapts separate camera encoder and trajectory encoder to distangle between camera movement and trajectory-specific object motions. However, its text guidance nature can not provide editing for fixed image and the objects in many synthesized video can not strictly follow the trajectory or presents collapsed, unnatural object appearance. This demonstrated that simply encoding trajectory into text-to-video model can not produce accurate trajectory-editing facing complex object or senario.

**DragAnything**: As reflected in the test demos from DragAnything, when facing complex or rotating trajectory, the model usually moves the camera following the angle instead of rotating the object itself. This illustrates its weakness for handling 3D-aware motions and this problem orignirates from its implicit condition of first-frame fixed object entity which constainrs the appearnce adapation in following frames.

## 4.2 Evaluation Metric on Following Experiments

To ensure a convincing and effective performance evaluation for our trajectory-based video generation experiments on following sections, we employ several metrics to measure spatial reconstruction quality, temporal consistency and overall trajectory accuracy. All metrics can be divided into image-based metrics, video-based metrics and trajectory-specific metric as follows.

**Image-based Metric**: For Image-based Metirc, we use the Frechet Inception Distance (FID) [18] and Structural Similarity Index Measure (SSIM) [39] to evaluate overall spatial generation quality. FID measures spatial reconstruction similarity and accuracy by comparing the difference between the distribution of ground truth images and generated images. Comparing to it, SSIM focuses more on overall spatial naturalness comparison between generated images and target ones. During experiments on self-reconstructed dataset, SSIM metric provides more performance insights for object-centric reconstruction quality, as we are rendering images with same background and SSIM tends to analysis more detailed features of object appearance.

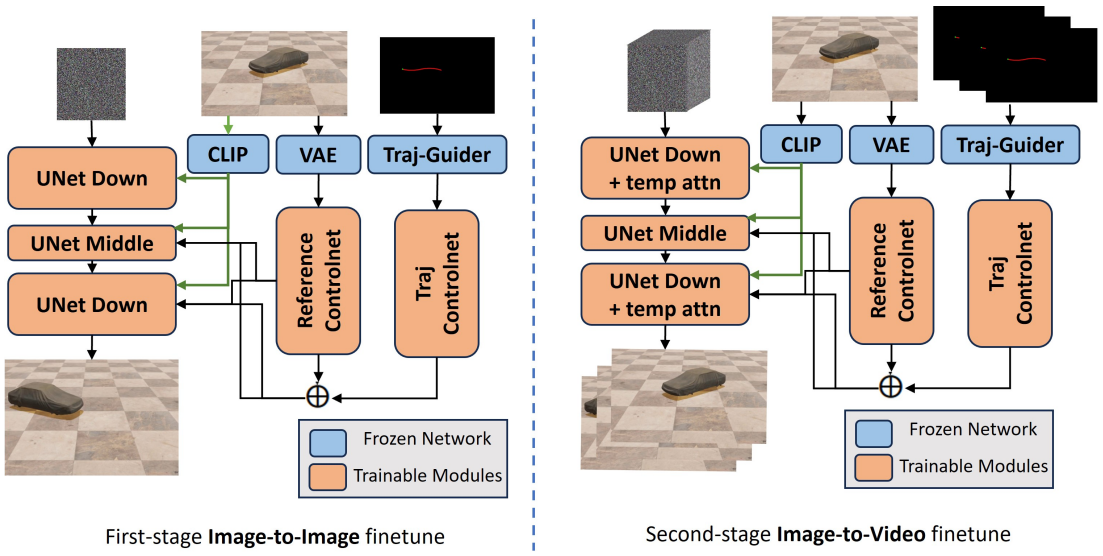


Figure 4.2: The structure of the implemented Traj-Disco adapted from original Disco for pose-guided video synthesis. It contains separate Controlnet to receive first-frane image and trajectory as generation condition.

**Video-based Metric**: To measure the generated video's quality in both spatial and temporal dimensions, we employ the Frechet Video Distance FVD [35] as video-based metric. Similar to FID, FVD first encodes video samples with pre-trained feature extractor such as I3D network [8] and then computes latent distribution difference between generated videos and ground truth videos. FVD is capable of measuring both spatial video quality and temporal consistency.

**Trajectory Metric**: Evaluating model solely with FID, SSIM and FVD is insufficient for demonstrating the accuracy of object movement following the given trajectory. Thus, we propose another metric named Motion-Diff to measure the trajectory accuracy of animated frames based on temporal motions across frames. Considering that objects rotate with changing for-ground angle, which may cause confusion for point tracker models, we employ the flow estimator UniMatch [44] as utilized in DragNUWA, to extra dense flow maps from generated and target video and than calculate the difference between them.

## 4.3 Implementation Baseline Selection

Aiming to select a proper implementation baseline for reliable video generation performance, we experiments with two popular model setting 1. Two-stage finetuning of Image Generation Model with inserted temporal modules [36, 23, 9] and 2. Single-stage


Figure 4.3: Visualization of generated video frames in DAVIS dataset, from top to bottom are: Ground Truth Video, Video from Disco Image-to-Image, Disco Image-to-Video, Finetuning SVD with 256*256 resolution and Finetuning SVD with 320*592 resolution.

Finetuning of Video diffusion model [38]. Specifically, for two-stage finetuning of Image-to-Image (I2I) model, we implemented Traj-Disco based on Disco [36] as shown in figure 4.2. The first-stage of training leverages Controlnet taken short-trajectory and initial image as input to finetune for the pre-trained UNet, and second-stage training continues finetuning with inserted trainable temporal layers with longer trajectories. For SVD-Finetuning, the structure is similar to our final model structure in figure 3.3, the Controlnet takes trajectory sequence as input to finetune for the entire 3D UNet.

### 4.3.1 Open-domain Dataset and Trajectory Annotation

Due to GPU resource and time constraint, we select a smaller dataset, DAVIS, compared to the training dataset Webvid-10M [2] of DragNUWA and VIPSeg [28] of DragAnything. DAVIS contains 90 videos containing multiple moving objects and corresponding object mask. For trajectory annotation, as shown in 4.1, we applies similar procedure with DragNUWA by employing Unimatch's [44] flow estimator to draw sparse trajectories of points sampled from each object's masks. A Gaussian Filter is further applied to enhance the strength of trajectory information.

| Model | FID | SSIM | FVD |
|---|---|---|---|
| Traj-Disco I2I | 144.65 | 0.439 | 2194.79 |
| Traj-Disco I2V | 188.59 | 0.587 | 3067.11 |
| Ft-SVD 256*256 | 67.60 | 0.782 | 1294.29 |
| Ft-SVD 320*576 | **46.60** | **0.866** | **1148.72** |

Table 4.1: Evaluation metrics for image-based FID, SSIM and video-based FVD for the baseline selection experiment. Finetune-SVD with 320*576 resolution achieves significantly better performance compared to two-stage finetuning for Traj-Disco.

### 4.3.2 Result Analysis

For detailed training implementation, we train each stage of finetuning for 20 epochs with Adam-W optimizer with initial learning rate of 1e-5. For each method, following Disco's implementation, we perform training on 256*256 resolution. To investigate the impact of finetuning resolution for SVD's Controlnet, we perform an extra training on 320*576 resolution on SVD following DragAnything.

We divided each video in DAVIS into 16-frame windows with 4 overlapping steps, finally sampling 1159 training sub-videos and 23 testset sub-videos. For Image generation based evaluation, we calculate frame-wise FID and SSIM, while for video generation, we measure FVD-VID and FVD for video of 16 frames.

As illstrurated from table 4.1 and figure 4.3, several conclusion could be drawn: 1. Finetuning SVD achieves significantly more realistic and temporal consistency comparing to Disco based two-stage finetuning of Image models. 2. Although first-stage Image-to-Image training of Disco could preserve the spatial consistency, the object's identity collapse for longer trajectory and totally broken for second-stage video finetuning. Thus, pre-trained temporal modelling plays an important role. 3. During SVD finetuning, larger resolution training produces notable better performance. This is reasonable as more pixels offers more fine-grained information for Controlnet learning.

## 4.4 Experiments on Self-constructed Blender Dataset

### 4.4.1 Dataset Details

To ensure a fair and robust evaluation of model's trajectory-specific controlling performance and its generalization ability, we constructed three sub-datasets: 1) Training dataset, 2) Test-1 dataset containing already seen objects with unseen moving trajectories and 3) Test-2 dataset containing unseen objects and trajectories. For detailed implementation, training subset contains 50 objects with 15 random sampled trajectories per object (750 video clips in total) and test-1 subset share same objects with one sampled unseen trajectory (50 video clips). Test-2 subset contains 10 unseen objects with two sampled unseen trajectory (20 video clips).

### 4.4.2 Training Implementation

As mentioned in section 3.2, to save computation memory and training time, only the Controlnet is trainable with 16 bits weights initialized from UNet's pre-trained downsampling blocks or pre-trained Controlnet from previous stage. All other modules including original UNet, VAE and image-specific CLIP are frozen in 8 bit, an accelerator package is also employed for further training speedup. Each training stage, including pre-training stage, contains 30k iteration steps. The used optimizer is Adam-W with an initial learning rate of 1e-5. Due to resource constraints, all training is operated on single 40G A100 card with a batch size of 1 where each stage of training took about 45 hours. Considering the sampling diversity originated from Diffusion architecture, we directly use model weights of 30k steps for evaluation without checkpoint selection through validation set.

### 4.4.3 Results Analysis for Self-Constructed Dataset

As illustrated in table 4.2, our method, which incorporates 3D-Aware Two-stage finetuning outperforms original version of single-stage SVD finetuning and Image-to-Video based Traj-Disco for image-based SSIM, video-based metric FVD and trajectory-based motion-diff. This further reflects the importance of extra-stage finetuning for modelling 3D-aware rotating movement facing precise trajectory guidance.
Shown from table 4.2 and figure 4.4, several conclusions can be made: 1. Similar to evaluation performance in section 4.3.2, for disco based two-stage fintuning from

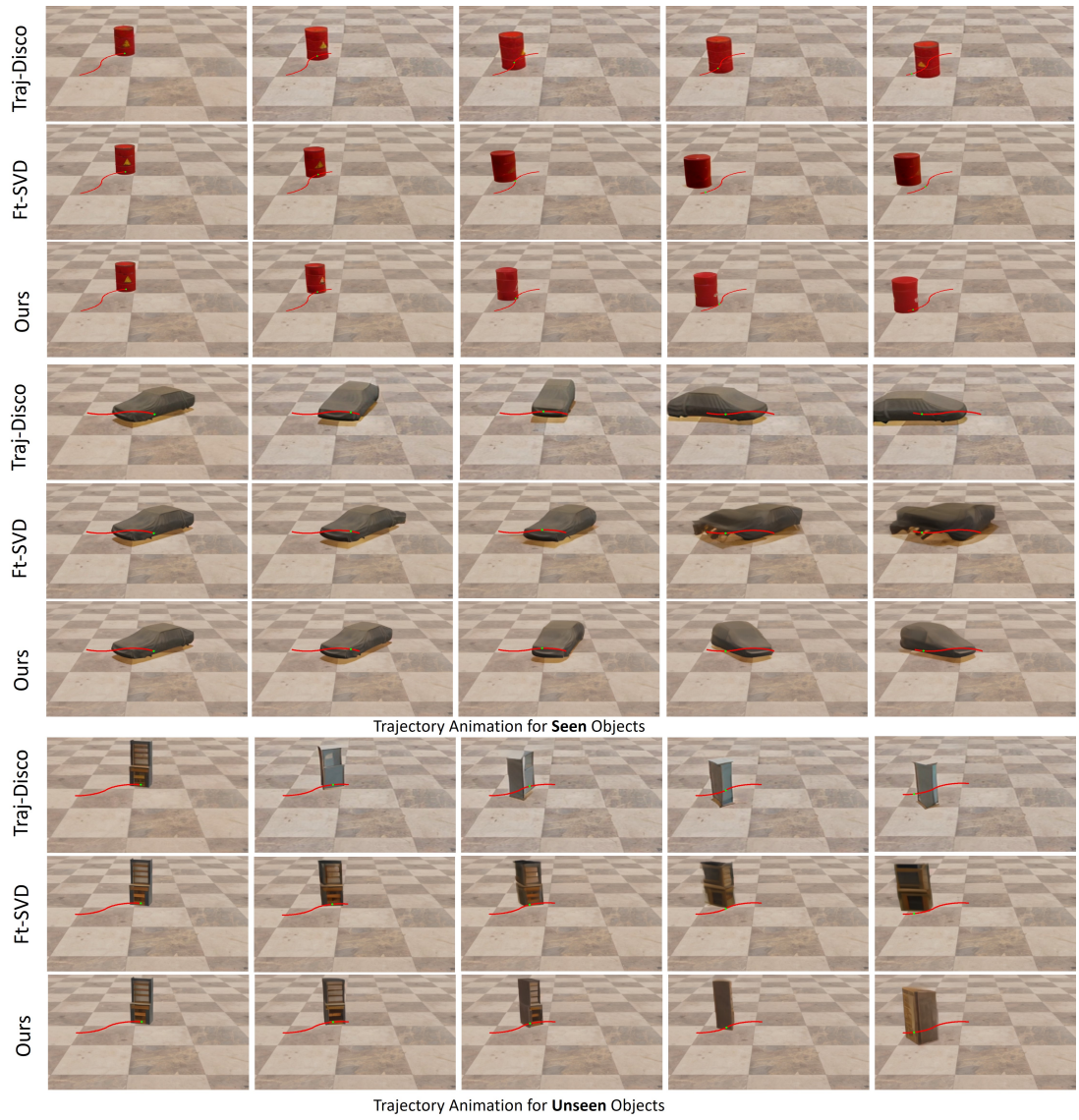


Figure 4.4: Visualization of generated video frames in self-constructed dataset, from modesls: Disco Image-to-Video, Finetuning SVD and our proposed method.

text-to-Image model, although the object and background can be reconstructed in few frames which results in higher FID, it can not ensure general reconstruction naturalness, temporal consistency and trajectory accuracy in subsequent frames with significantly lower SSIM, and higher motion-difference and FVD score. Its better single-frame reconstruction of FID is highly likely due to its first-stage image-based training. 2. For the baseline method of Controlnet-based finetuning of SVD, although its temporal consistency and trajectory accuracy outperforms Traj-disco, its overall performance for all metrics is lower than our final designed model with corresponding worse visualization result where the object usually can not rotate accurately and would collapse its

| | Model | FID | SSIM | FVD | Motion-Diff |
|---|---|---|---|---|---|
| Seen objs | Traj-Disco | **77.10** | 0.812 | 336.02 | 0.151 |
| | Ft-SVD | 108.36 | 0.892 | 212.14 | 0.143 |
| | Ours | 101.54 | **0.904** | **180.17** | **0.098** |
| UnSeen objs | Traj-Disco | 122.65 | 0.801 | 440.44 | 0.176 |
| | Ft-SVD | 122.16 | 0.880 | 265.01 | 0.186 |
| | Ours | **114.62** | **0.895** | **235.40** | **0.122** |

Table 4.2: Evaluation metrics for image-based FID, SSIM and video-based FVD-VID and FVD for 3D-aware movement animation on our self-constructed dataset for both seen and unseen objects. Our proposed model incorporating two-stage 3D-aware finetuning outperforms other baseline including Traj-Disco and directly Finetuning SVD.

identity structure following the trajectory. 3. Comparing to simply empolying trajectory Controlnet, our final model trained by two-stage 3D-aware finetuning with 3D bounding box prompt and spatial enhancement loss generates much more reliable object appearance, which the object integrity during rotation, owing to model's prior knowledge for 3D reconstruction and pose estimation, resulting in significantly improved FVD and SSIM. Moreover, the model's trajectory accuracy is much better as reflected by both a significantly lower motion difference score and the visualization results of rotating objects following complex trajectory.

We also tested our model on unseen object to reflect their generalizability and to determine whether our trajectory model overfits for already-seen training objects. We can observe from table 4.2 that the performance comparison between the three methods on unseen objects is similar to that on seen objects but with larger performance gap. Our 3D-aware finetuning model achieves better overall performance for FID, SSIM, FVD score and motion-accuracy which demonstrates its superior ability for generalize across different scenarios through injecting 3D-prior knowledge. Moreover, it is notable that Traj-Disco exhibits worse performance distance regarding all four metrics on unseen objects and can not maintain object's appearance even in first few frames. This highlights the importance of a large pre-trained model on general videos, which contributes better object-centric temporal reconstruction consistency and trajectory understanding.

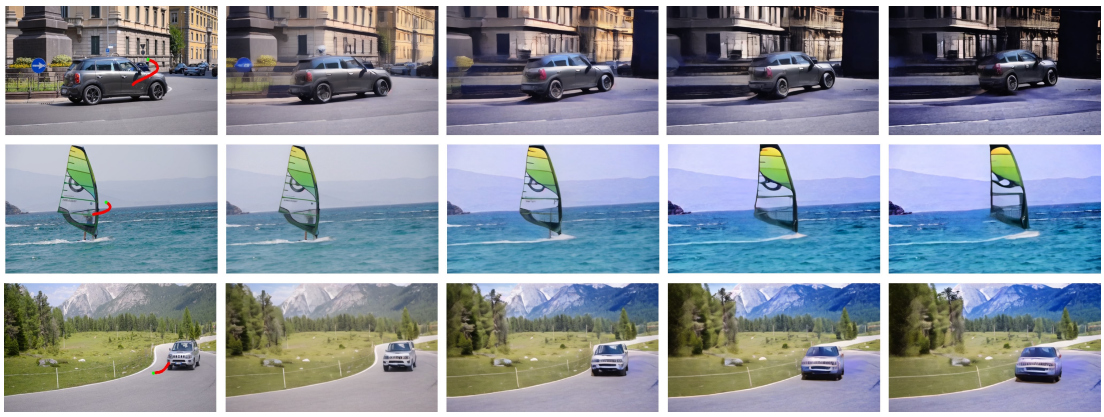

Figure 4.5: Visualization of generated video frames inferenced from the model weight finetuned in self-constructed dataset on zeroshot open-domain videos with reliable trajectory animation.

### 4.4.4 Performance on Zero-shot Controlling on Open-domain Videos

To further evaluate the effectiveness of pre-training in self-reconstructed data, we conducted an additional experiment for 3D-aware trajectory controlling on open-domain videos, which involves zero-shot data distribution compared to our unreal rendered dataset. Specifically, we select several video samples with corresponding trajectory from previous DAVIS dataset and directly apply our Controlnet weights finetuned from self-constructed dataset for animation based on its first-frame. As visualized in figure 4.5, it can be seen that our model can successfully animate target object with accurate movement including rotation following rotating trajectory while maintains the background with minor natural movement. This result demonstrated that the learned 3D-aware trajectory guided motions is decoupled with for-ground and background appearance as well as input image's distribution from finetuned dataset and can be transfered in zero-shot manner on open-domain videos. It is noteworthy that current state-of-the-art trajectory oriented works including DragNUWA, MotionCtrl and DragAnything can not produce similar 3D-aware animation performance of target object given only 2D trajectory in pixel space.

## 4.5 Ablation Study

During model development, to optimize model performance and analysis the effectiveness of different modules, we apply experiments mainly on three aspects: 1. The method of injecting 3D information during model training, 2. Effectiveness of model

| Dataset | FID | SSIM | FVD | Motion-Diff |
|---|---|---|---|---|
| Ours | **101.54** | **0.904** | **180.17** | 0.098 |
| - Spatial Loss | 102.73 | 0.901 | 182.14 | **0.095** |
| - two-stage finetuning | 118.70 | 0.897 | 261.38 | 0.140 |

Table 4.3: Evaluation metric of our ablation study to compare the performance of trained model without Spatial-enhancement loss or two-stage 3D-aware finetuning

module design and 3. The importance of dataset scale for our self-reconstructed dataset for promising performance.

### 4.5.1 Importance of Model Design

To evaluate the effectiveness of our designed modules to increase model performance, we conducted the first ablation study to compare the performance metric of models without our proposed 3D-aware two-stage finetuning pipeline. To be specific, we conducted two experiments: 1. Training our model without first-stage finetuning involving 3D bounding box generation and 2. Training our model without spatial enhancement loss during two-stage finetuning.

As shown in table 4.3 and figure 4.6, without spatial loss, although model captures comparable trajectory accuracy as the original model, it exhibited decreased spatial construction quality as reflected in corrupted object identity in visualization results and decreased FID and SSIM, as well as slight dropping in temporal consistency from FVD. For the model without two-stage finetuning, it perform much worse on all metrics, with serious identity collapse which even results in the object vanishing in last few frames with inaccurate motions. Moreover, its performance is also worse than directly finetuning SVD, which suggests that for a model with week spatial reconstruction ability, spatial loss can reversely hinder model learning by imposing too strong spatial constraints. These results further demonstrate the importance of applying both two-stage finetuning and spatial loss for reliable and accurate trajectory-guided 3D-aware video generation.

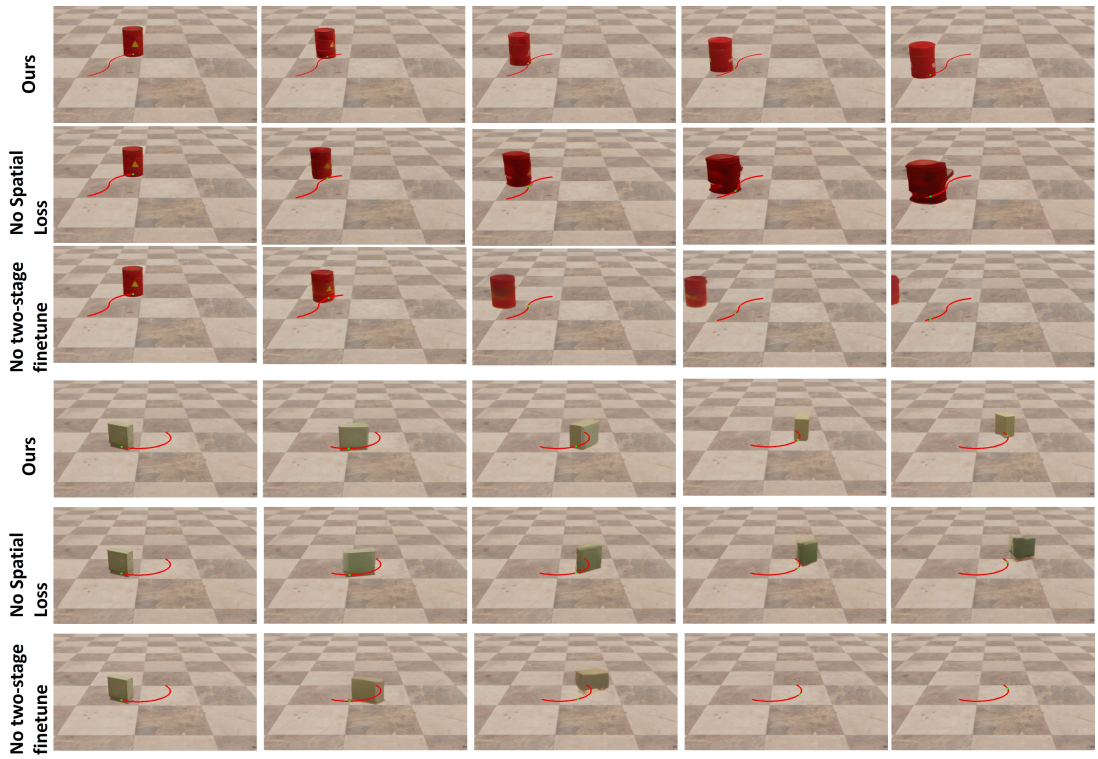


Figure 4.6: Visualization of generated animation videos for our trained model without Spatial-enhancement loss or two-stage 3D-aware finetuning

### 4.5.2 How to inject 3D Guidance Information

Despite our finally choice of the two-stage finetuning method with first-stage 3D-bbox generative training, we have experimented on another 3D conditioning mode named ”bbox-condition”. Considering the effectiveness of 3D bounding box guidance, we designed an alternative 3D-aware module (shown from figure 4.7) by providing first-frame’s bounding box image with a bbox guider. This bbox guider first encodes 3D-bbox image into feature $B_c$ through similar structure as our trajectory-guider and then concatenates it with the trajectory feature from trajectory-guider by repeating it n times into $B_{c1}, B_{c2}, ..., B_{cL}$. To investigate whether bbox-condition mode is also beneficial for model performance, we employ three experiments: 1. model trained with our two-stage finetuning only, 2. model with 3D bbox-condition only and 3. model with two-stage finetuning and bbox-condition in each training stage.

As illustrated in table 4.4, employing only two-stage bounding-box aware finetuning outperforms model with 3D-bbox condition over all metrics with frequent inaccurate animated object’s movement and object poses. Moreover, even combing first-frame bbox condition with two-stage finetuning, the model performance also drops especially for

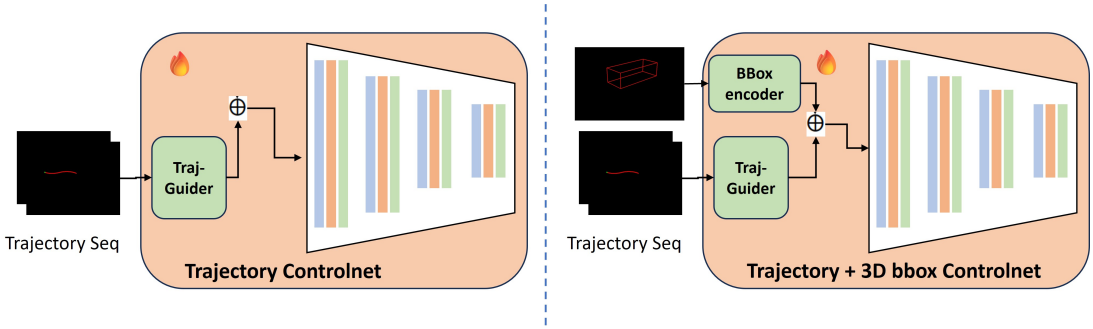


Figure 4.7: Visualization of generated video frames in self-constructed dataset, from top to bottom are: Ground Truth Video, Disco Image-to-Video, Finetuning SVD and our proposed model.

| Model | FID | SSIM | FVD | Motiond-Diff |
|---|---|---|---|---|
| Ft-SVD | 108.36 | 0.892 | 212.14 | 0.143 |
| + bbox condition | 117.56 | 0.891 | 205.43 | 0.136 |
| + 3D two-stage ft | **101.54** | **0.904** | **180.17** | **0.098** |
| + both | 110.92 | 0.898 | 196.01 | 0.105 |

Table 4.4: Evaluation metric of our ablation study to compare two potential methods for 3D-aware conditioning including two-stage finetuning with 3D bbox and employing bbox-condition through Controlnet.

generated temporal video quality. This result demonstrates that providing the 3D bounding box in first frame alone is not helpful for model's understanding of 3D knowledge for rotating poses, while it increases model's learning difficulty with extra parameters. This is reasonable because, compared to forcing the model to generate3D bounding boxes sequence, single frame condition of first frame can not provide awareness of potential temporal changing trend for 3D positional information following trajectory. In contrast, generating 3D bounding box sequences is capable of combining explicit modelling of objects' continuous 3D motions during trajectory-animated training.

### 4.5.3 Dataset Scale

For reliable and accurate trajectory animation both on seen rendered dataset and unseen dataset, it is crucial to have a large enough dataset containing corresponding trajectory and animated videos for the added Controlnet training as done in DragNUWA and DragAnything [43]. During the experiments for model developing, we also observed that

| | Model | FID | SSIM | FVD-VID | FVD |
|---|---|---|---|---|---|
| Seen objs | Small Data | 105.03 | 0.896 | 201.13 | 0.128 |
| | Medium Data | 110.84 | 0.899 | 204.52 | 0.099 |
| | Large Data | **101.54** | **0.904** | **180.17** | **0.098** |
| UnSeen objs | Small Data | 130.72 | 0.890 | 269.37 | 0.142 |
| | Medium Data | 115.63 | 0.893 | 262.12 | **0.107** |
| | Large Data | **114.62** | **0.895** | **235.40** | 0.122 |

Table 4.5: Evaluation metric of our ablation study for the performance from models trained with different data scale.

the trained model with the same structure suffered from insufficient self-reconstructed dataset. To further evaluate the importance of dataset scale and provide several training hints for following trajectory-based generation works, we conducted experiment on three different data scale: small ($S_{gen}$), medium ($M_{gen}$) and large ($L_{gen}$). $L_{gen}$ contains same data amount of 750 videos with 15 clips for 50 objects. The small set $S_{gen}$ contains 250 videos with 5 clips for each object and the medium set $M_{gen}$ have 500 videos with 10 clips per object. To ensure fair comparison, all models on each dataset are trained with two-stage 3D-aware finetuning for 30k per stage.

As illustrated in table 4.5, model trained with the largest dataset achieve better overall performance across all metrics on both seen and unseen dataset compared to model trained with medium and small dataset. This phenomenon verifies that reliable trajectory animation and 3D-aware reconstruction capacity actually requires large dataset to provide sufficient 3D-aware understanding across different scenarios facing comprehensive trajectories. Moreover, it can be noticed that the overall performance gap between largest data and medium data is more obvious than that between medium and small datasets, this phenomenon indicates the fact that larger data can further boost model's trajectory-guided generative ability at higher rate while its model's overall convergence and generalizability suffers with insufficient data.

### 4.5.4 Trajectory Representation

In this section, we conduct ablation experiment to evaluate the influence of incorporating sparse trajectories sampled from region-aware trajectory sampler during data annotation

| | Traj-sampler | FID | SSIM | FVD | Motion-Diff |
|---|---|---|---|---|---|
| Seen objs | No | 111.85 | 0.900 | 217.66 | 0.124 |
| | YES | **101.54** | **0.904** | **180.17** | **0.098** |
| UnSeen objs | No | 116.75 | 0.892 | 237.86 | 0.155 |
| | Yes | **114.62** | **0.895** | **235.40** | **0.122** |

Table 4.6: Evaluation metric of our ablation study for comparing the model performance with and without trajectory representation of sparse trajectories from region-aware sampler.

process. Specifically, we experiment on the model trained with single driving precise trajectory and the model with sparse trajectories representation to verify whether more sampled trajectories surrounding object can enhance trajectory understanding and facilitate the use of hand-crafted imprecise trajectories during inference. In detail, we employ same test sets including imprecise sampled trajectories as previous experiments and modify training time encoded trajectories into the precise one for same 3D-aware two stage fine-tuning strategy.

As illustrated in table 4.6, when compared to directly employing single driving trajectory, the application of trajectory sampler to generate more sparse and random trajectories around object yields better performance regarding all metrics on both seen and unseen objects, particularly in relation to trajectory accuracy and video quality measured by FVD score. These findings further evaluates the hypothesis that more sparse and inaccurate trajectories (those not strictly following the object center) during training enhance model's awareness for trajectory-based motion guidance and support the generalization for hand-crafting trajectory.

# Chapter 5

# Discussion

## 5.1 Contribution of Our Project

As shown from our experiments, several contributions can be observed from our self-constructed data and proposed methodology:

### 5.1.1 Importance of Self-Constructed Data

As illustrated from our experiments on DragAnything's performance for rotating and large trajectory, fine-tuning video generation model on open-domain videos cannot incorporate sufficient 3D understanding for motions like rotation. Through finetuning model on our self-constructed dataset, the model is capable of rotation following complex trajectory for target object with satisfied spatial and temporal consistency. Furthermore, by comparing between DragAnything's performance and our model's zeroshot performance on open-domain videos, we can discover that the the prior knowledge of 3D rotation and large motions can be injected by pre-training the model on our constructed dataset with generalization potential on wild videos. Although there is domain gap between constructed data in blender software and realistic videos, model finetuned from Controlnet can capture the trajectory guidance with corresponding 3D influence of object's appearance by decoupling motion information between background and data distribution of images.

### 5.1.2 Importance of Two-stage 3D-Aware Finetuning

As discussed in the ablation study, simply employing trajectory finetuning through Controlnet still cannot handle 3D rotating motions, even in our self-constructed dataset, with

serious identity collapse problem and inaccurate animated motions. This phenomenon further emphasizes the importance of two-stage 3D finetuning which successfully generated videos with consistent object identity during rotation with more accurate motions following complex trajectory. Facilitated by 3D bounding box as 3D signals containing both rotating 3D pose and temporal positions, our model achieve superior performance on all metrics including Image-based FID, SSIM, Video based FVD and Trajectory-based motion difference.

## 5.2 Limitation and Future Work

Although our current model with 3D-aware two-stage finetuning ans spatial enhancement loss has achieved superior performance for rotating animation following large trajectory, several problems still remains.

### 5.2.1 Reconstruction Quality for Complex Object

As illustrated in section 4.5.1, incorporating spatial loss has successfully stablized the reconstruction of rotating object's appearance. However, the appearance identity still collapse in special situations if objects have too complex geometrical structure like armchair with multiple legs. This problem is highly due to the pre-trained SVD does not involve enough awareness for 3D reconstruction on complex structure and our current dataset not being large enough to involve sufficient prior knowledge during finetuning. For future work, to achieve more realistic reconstruction for rotating objects, we will construct a much-larger rendering dataset to provide 3D prior information with more diverse objects sampled from objverse-xl [13] which is the largest high-quality 3D objects dataset containg millions of high-quality 3D models.

### 5.2.2 Domain Gap between Rendered Dataset and Real Videos

In this project, we have demonstrated that our proposed model is capable of 3D-aware trajectory animation for in self-rendered dataset and exhibits potential zero-shot animation capacity on open-domain videos. However, as shown in figure 4.5, the animation for the object movement produces unnatural synthesized frames with darker background and blurred object identity in the final few frames. For example, the background for the black car turns darker with blurry car's rear-end appearance. Consequently, for our future work, we are attempting to bridging the domain gap

between rendered data and real videos through several potential methods including 1. Constructing more realistic environment for blender rendering with larger realistic 3D models and 2. Collecting a diverse-set high-quality videos containing obvious 3D motions like rotating through designed data filtering pipeline to ensure our selected videos do not contain obvious camera movement. Subsequently, we plan to fine-tune our pre-trained model with a lower learning rate or extra Controlnet to adapt our pre-trained model's 3D-awareness from limited domain of rendering data to general open-domain scenarios.

# Chapter 6

# Conclusions

In conclusion, our project self-constructed a dataset containing object's 3D motions like rotating with corresponding trajectory and successfully introduced a 3D-Aware Video Generation model with a novel two-stage 3D-aware fine-tuning process which can generate identity-consistent 3D motions like large-angle rotation following complex trajectory in our self-constructed dataset. Our proposed model exhibits corresponding enhanced metric and visualization performance and further demonstrates its potential 3D controlling generalizability on zero-shot open-domain videos.

Our primary motivation lies in tackling the main challenge of video generation for generating accurate 3D rotation motions, especially when facing large-angle and complex trajectory. Our experimental findings indicate that this challenge arises from two main problems 1. open-domain videos lack sufficient 3D rotating scenarios and 2. Simply injecting 2D trajectory controlling can not provide extra 3D conditioning prompt necessary for changing object pose and appearance during rotation. To address these issues, in our project, we first self-constructed a realistic rendering dataset containing object's large-angle rotation driving by randomly sampled complex arc or S-trajectory in blender software. Built upon this, we designed a two-stage 3D-aware fine-tuning strategy by forcing model to generate object with 3D bounding box in first stage as 3D prompt to facilitate the capture the change of object's position, pose and appearance during rotation following given trajectory. After generating accurate 3D bbox sequences, in second stage, model can easily learn to eliminate the 3D bounding box and thereby enhancing the reconstruction quality of the object's identity. Additionally, we employed spatial loss to improve model's reconstruction quality for more complex object's appearance across longer frames by randomly passing single-frame noise with corresponding short trajectory to generate the respective single-frame object.

Compared to simply encoding trajectory for Controlnet's condition signal, our proposed method achieves superior performance regarding all image-based metric, video-based metric and motion difference which represents significantly higher temporal and spatial consistency as well as improved trajectory accuracy. Through ablation study, we further validates the effectiveness of our model design, the appropriateness of our method of injecting 3D conditional guidance and the usage of trajectory sampler. Furthermore, leveraging the scalability of our self-generated dataset, we investigate the impact of dataset scale for achieving stable rotating motions and discover that our model suffers from insufficient smaller data.

In our future work, we are planning on two primary objectives 1. constructing a larger-scale data to further improve model's 3D-awareness facing complex objects and 2. collecting and filtering a high-quality open-domain videos to fully adapt the model's 3D-awareness from our self-constructed data into realistic videos, achieving state-of-the-art trajectory controlling performance.

# Bibliography


[1] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023.

[2] 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*, 2021.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14707–14717, 2021.

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024.

[7] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[9] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2023.

[10] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.

[11] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023.

[12] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[15] 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.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[17] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[21] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[23] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.

[24] Longbin Ji, Pengfei Wei, Yi Ren, Jinglin Liu, Chen Zhang, and Xiang Yin. C2g2: Controllable co-speech gesture generation with latent diffusion model. *arXiv preprint arXiv:2308.15016*, 2023.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[26] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022.

[27] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc., 2023.

[28] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[29] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, October 2023.

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[32] 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 (CVPR)*, pages 10684–10695, June 2022.

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 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] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[36] 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. *arXiv preprint arXiv:2307.00040*, 2023.

[37] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[40] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023.

[41] 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.

[42] 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*, pages 720–736. Springer, 2022.

[43] Wejia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion

control for anything using entity representation. *arXiv preprint arXiv:2403.07420*, 2024.

[44] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[45] 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 SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[46] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models.