Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation

YINGQING HE*, Hong Kong University of Science and Technology, China
MENGHAN XIA*, Tencent AI Lab, China
HAOXIN CHEN*, Tencent AI Lab, China
XIAODONG CUN, Tencent AI Lab, China
YUAN GONG, Tsinghua Shenzhen International Graduate School, Tsinghua University, China
JINBO XING, The Chinese University of Hong Kong, China
YONG ZHANG†, Tencent AI Lab, China
XINTAO WANG, Tencent AI Lab, China
CHAO WENG, Tencent AI Lab, China
YING SHAN, Tencent AI Lab, China
QIFENG CHEN†, Hong Kong University of Science and Technology, China

Fig. 1. Our synthesized videos with a consistent character (user-provided) and storyline. By utilizing text-retrieved video clips as structure guidance, our system manages to generate high-quality videos of similar structures, while also accommodating arbitrary scene appearances based on text prompts. Furthermore, our proposed concept personalization method ensures consistent character rendering across different plot scenarios. Each clip is visualized with three keyframes.

Generating videos for visual storytelling can be a tedious and complex process that typically requires either live-action filming or graphics animation rendering. To bypass these challenges, our key idea is to utilize the abundance of existing video clips and synthesize a coherent storytelling video by customizing their appearances. We achieve this by developing a framework comprised of two functional modules: (i) Motion Structure Retrieval, which provides video candidates with desired scene or motion context described by query texts, and (ii) Structure-Guided Text-to-Video Synthesis, which generates plot-aligned videos under the guidance of motion structure and text prompts. For the first module, we leverage an off-the-shelf video retrieval system and extract video depths as motion structure. For the second module, we propose a controllable video generation model that offers flexible controls over structure and characters. The videos are synthesized by following the structural guidance and appearance instruction. To ensure visual consistency across clips, we propose an effective concept personalization approach, which allows the specification of the desired character identities through text prompts. Extensive experiments demonstrate that our approach exhibits significant advantages over various existing baselines.

*First authors
†Corresponding authors
Project page : https://videocrafter.github.io/Animate-A-Story

Additional Key Words and Phrases: Story Visualization, Video Diffusion Models, Retrieval-augmented Generation, Personalized Generation

1 INTRODUCTION

Creating engaging storytelling videos is a complex and laborious process that typically involves live-action filming or CG animation production. This technical nature not only demands significant resources from professional content creators but also creates barriers for the general public in effectively utilizing this powerful medium. Recently, significant progress has been made in text-to-video (T2V) generation, allowing for the automatic generation of videos based on textual descriptions [He et al. 2022; Ho et al. 2022a; Singer et al. 2022; Zhou et al. 2022].

However, the effectiveness of these video generation techniques is still limited, yielding results that fall short of expectations and hinder their practical application. Additionally, the layout and composition of the generated video cannot be controlled through text, which is crucial for visualizing an appealing story and filming a movie. For example, close-ups, long shots, and composition can assist directors in conveying implicit information to the audience. Current text-to-video generation models can hardly generate proper motions and layouts that meet the requirement of film.

To overcome these challenges, we propose a novel video generation approach incorporating the abundance of existing video content into the T2V generation process, which we refer to as *retrieval-augmented video generation*. Specifically, our approach retrieves videos from external databases based on text prompts and utilizes them as a guidance signal for the T2V generation. Building on this idea, our approach also enables users to have greater control over the layout and composition of the generated videos when animating a story, by utilizing the input retrieved videos as a structure reference. Additionally, the quality of the generated videos is enhanced by leveraging the rich knowledge and information contained in the retrieved videos, which can be used to improve the realism and coherence of the generated scenes. The text prompt is now responsible for rendering the appearances of scenes and objects to generate novel videos.

However, such retrieval-guided video generation processes still suffer from the inconsistency problem of the character across different video clips. Besides, the character's appearance is controlled by the text prompt and generated in a stochastic way which lacks user control. To further tackle this issue, we study existing literature on personalization [Ruiz et al. 2023] to finetune the generation model to re-render the appearance of the character. and propose a novel approach (TimeInv) to better represent personalized concepts and improve performance.

By incorporating the two core modules: retrieval-enhanced T2V generation and video character rerendering, our approach provides a more efficient and accessible way for content creators to produce high-quality animated videos. To justify its effectiveness, we evaluate our method from the following perspectives: First, our retrieval-enhanced T2V generation model is compared against existing baselines, demonstrating notable superiority in video generation performance. Second, we justify the advantage of our proposed personalization method in relation to existing competitors. Furthermore, we conduct comprehensive experiments of the overall effectiveness of our proposed storytelling video synthesis framework, suggesting its potential for practical applications.

Our contribution is summarized as the following:

- We present a novel retrieval-augmented paradigm for storytelling video synthesis, which enables the use of existing diverse videos for storytelling purposes for the first time. Experimental results demonstrate the framework's effectiveness, positioning it as a novel video-making tool with remarkable convenience.
- We propose an adjustable structure-guided text-to-video model, which effectively resolves the conflict between structure guidance and character generation. ?
- We propose TimeInv, a new concept personalization approach that outperforms existing competitors notably.

2 RELATED WORK

2.1 Video Generation

Numerous earlier works [Saito et al. 2017; Skorokhodov et al. 2022; Tulyakov et al. 2018; Vondrick et al. 2016; Wu et al. 2021] focus on unconditional video generation methods, employing generative adversarial networks (GANs) or variational auto-encoder (VAE) to model video distributions. For instance, VGAN [Vondrick et al. 2016] separately models the foreground and background by mapping random noise to space-time cuboid feature maps. The fused feature map signifies the generated video and is input into a discriminator. TGAN [Saito et al. 2017] generates a noise sequence with a temporal generator and then uses an image generator to transform it into images. These images are concatenated and fed into a video discriminator. StyleGAN-V [Skorokhodov et al. 2022] leverages the capabilities of StyleGAN. It uses multiple random noises to control motion and an additional noise to manage appearance.

Several methods [Ge et al. 2022; Yan et al. 2021; Yu et al. 2023a] aim to capture spatio-temporal dependencies using transformers in the latent space. Initially, they project videos into a latent space by learning a VAE or VQGAN [Esser et al. 2021], followed by training a transformer to model the latent distribution. For example, TATS [Ge et al. 2022] trains a time-agnostic VQGAN and subsequently learns a time-sensitive transformer based on the latent features. VideoGPT [Yan et al. 2021] follows a similar pipeline. To enable text control over generated content, some works [Hong et al. 2022; Villegas et al. 2022] extract visual and text tokens, projecting them into the same latent space. A transformer is consistently used to model the interdependencies among them.

Recently, text-to-image (T2I) generation [Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022] has achieved significant advancements in generating high-quality images, primarily due to diffusion-based models. Capitalizing on the progress of T2I generation, the field of text-to-video generation has experienced breakthroughs as well. VDM [Ho et al. 2022b] is the first work employing diffusion models for video generation. Make-a-video [Singer et al. 2022] and Imagen Video [Ho et al. 2022a] are cascade models that initially model video distribution at low-resolution and then apply spatio-temporal interpolation to increase the resolution and time duration. Inspired by LDM [Rombach et al. 2022], several works [Blattmann et al. 2023; He et al. 2022; Luo et al. 2023; Mei and Patel 2022; Yu et al. 2023b; Zhou et al. 2022] extend LDM for video generation. For example, LVDM [He et al. 2022] inflates LDM into a video version

by introducing temporal attention layers. It employs the pretrained LDM as initialization and trains the learnable parameters with video data. Similar to LDM, the text embedding is injected into the UNet using the cross-attention mechanism. Video LDM [Blattmann et al. 2023] shares the same concept as LVDM, but with the distinction that it fixes the spatial weights of LDM.

?? 2.2 Structure-guided Video Generation

Mirroring the evolution of T2I, numerous works [Wang et al. 2023a,b; Xing et al. 2023; Yang et al. 2023; Zhang et al. 2023] investigate the capability of conditional video generation based on pretrained text-to-image or text-to-video models. For instance, Make-Your-Video [Xing et al. 2023] uses depth as an additional condition besides text. The spatial weights of Stable Diffusion are fixed, while the newly added temporal weights are learned on video data. Since depth is extracted from the video, it can re-render the appearance of the source video. Follow-Your-Pose [Ma et al. 2023] utilize pose as a condition to guide the human-like character video synthesis process. VideoComposer [Wang et al. 2023b], an extension of Composer [Huang et al. 2023], takes multiple types of images as conditions, such as RGB images, sketches, depths, etc. These conditions are fused in the latent space and interact with the UNet via cross attention.

2.3 Concept Customization

Generating an image with a specified object is referred to as customization or personalization. Numerous works [Alaluf et al. 2023; Gal et al. 2022; Kumari et al. 2023; Ruiz et al. 2023; Wei et al. 2023] explore this task from various perspectives. Textual inversion [Gal et al. 2022], the first inversion work on Stable Diffusion, optimizes a token without tuning the model for a given object's images. In contrast, Dreambooth [Gal et al. 2022] learns a token and fine-tunes the entire model parameters. Multi-concept [Kumari et al. 2023] is the first to propose a method for inverting multiple concepts. ELITE [Wei et al. 2023] trains an encoder to map visual images to text embeddings for customized text-to-image generation, rather than optimization. NeTI [Alaluf et al. 2023] introduces a novel text-conditioning latent space that depends on both the timestep and UNet layers. It learns a mapping that projects timestep and layer index into the embedding space.

3 METHOD

Our goal is to develop a framework that can automatically generate high-quality storytelling videos based on storyline scripts or with minimal interactive effort. To achieve this, we propose to retrieve existing video assets to enhance the performance of T2V generation (see Sec. 3.1). Specifically, we extract the structures from retrieved videos, which will then serve as guidance signals provided to the T2V process (see Sec. 3.2). Additionally, we perform video character rerendering based on the proposed TimeInv approach to synthesize consistent characters across different video clips (see Sec. 3.3). In the following sections, we will delve into the key technical designs that enable its functionality.

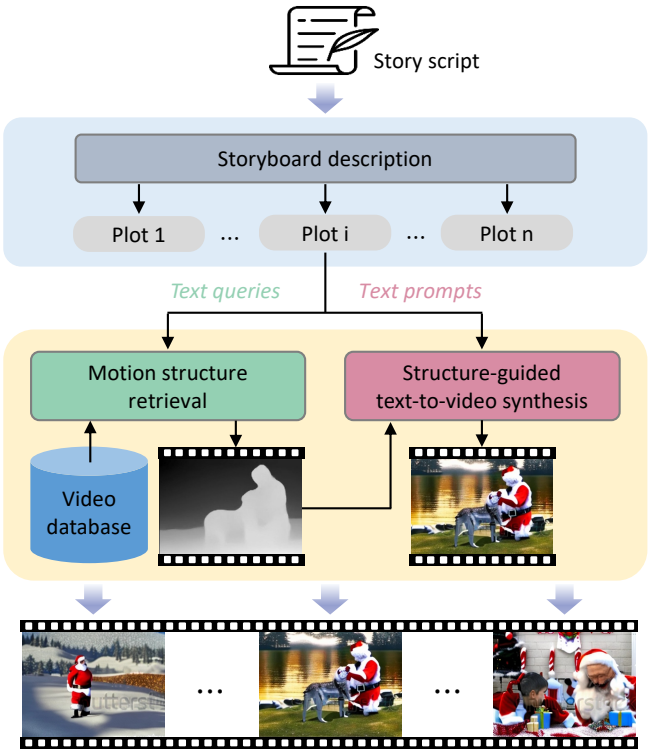

Fig. 2. Flowchart of our retrieval-augmented video synthesis framework. Given a textual story script, we first extract the key plots and modulate their descriptions as text queries and prompts. Each plot is transformed into a generated video clip through two modules: a video retrieval system and a structure-guided text-to-video model.

3.1 Retrieval-augmented Text-to-Video Generation

As illustrated in Fig. 2, our video generation framework involves three procedures: text processing, video retrieval, and video synthesis. In the text processing stage, we extract the key plots from the story script through storyboard analysis. To simplify the problem, we regulate an individual plot as a single event without a shot transition. For instance, "a boy ran into a wolf in the forest" is a single plot, while "a boy ran into a wolf in the forest and he killed the wolf with a gun" should be separated into two plots. For each plot, we further adjust and decorate the description so that they can serve as effective text queries and text prompts respectively. This stage is completed manually or using the assistance of large language models (LLMs) like GPT-4 [OpenAI 2023].

After that, we process each plot separately using two sequentially conducted modules. Given the text query, we can obtain video candidates showing desired scenario through an off-the-shelf text-based video retrieval engine [Bain et al. 2021] that associates with a database with about 10M open-world videos collected from the Internet. Since the video appearance may not match the plot precisely, we only take the motion structure of it by applying a depth estimation algorithm to it. This extends the usability of existing videos. As the example illustrated in Fig. 3, to synthesize the video of "Santa Claus playing with a *wolf* in the forest", we can use the

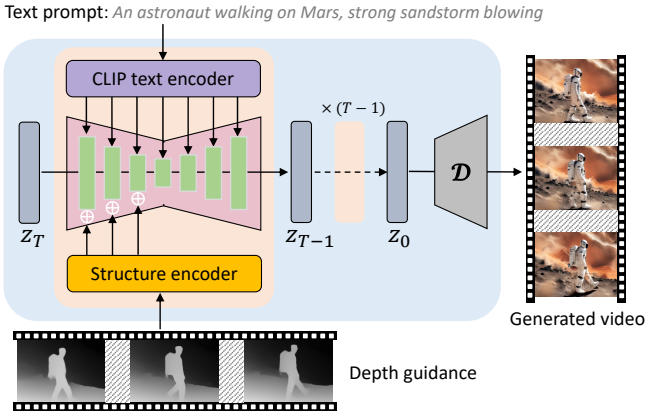

Fig. 3. Overview of our adjustable structure-guided text-to-video model. We use the depth information from source videos to guide the video synthesis process. The model consists of two branches: a general text-to-video synthesis branch which is a video diffusion model in the latent space, and a side branch for encoding and imposing structure control. The controlling mechanism is elementwise feature addition. Notably, the depth control is *adjustable* and this property is crucial for further character rerendering, which we will illustrate in Sec. 3.3.

motion structure in the video of "a man playing with a *dog* in the park", which is quite common in the video database. Utilizing the motion structure as guidance, we can synthesize plot-aligned videos through text prompts. Next, we describe the structure-guided T2V model in detail.

3.2 Structure-Guided Text-to-Video Synthesis

Preliminary. Denoising Diffusion Probabilistic Models (DDPM) [Ho et al. 2020], also called Diffusion Models (DM) for short, learn to model an empirical data distribution $p_{data}(\mathbf{x})$ by building the mapping from a standard Gaussian distribution to the target distribution. Particularly, the forward process is formulated as a fixed diffusion process that is denoted as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t, \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I). \tag{1}$$

This is a Markov chain that adds noise into each sample $\mathbf{x}_0$ gradually with the variance schedule $\beta_t \in (0,1)$ where $t \in \{1,...,T\}$ with T being the total steps of the diffusion chain. The reverse process is realized through a denoiser model $\mathbf{f}_\theta$ that takes the diffused $\mathbf{x}_t$ together with the current time step t as input and it is optimized by minimizing the denoising score matching objective:

$$L(\theta) = \mathbb{E}_{\mathbf{x}_0 \sim p_{data}, t} \|\epsilon_t - \mathbf{f}_\theta(\mathbf{x}_t; \mathbf{c}, t)\|_2^2, \tag{2}$$

where $\mathbf{c}$ is optional conditioning information (e.g. text embedding) and the supervision ϵ_t is a randomly sampled noise used in the forward diffusion process: $\mathbf{x}_0 \rightarrow \mathbf{x}_t$. Extended from DM, LDM formulates the diffusion and denoising process at a latent space. That is usually a compressive representation space that is separated and learned through a Variational Autoencoder(VAE) [Kingma and Welling 2014] comprised of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$.

We employ a conditional LDM to learn controllable video synthesis, as the overview depicted in Fig. 3. In our approach, the videos are transformed (or reconstructed) into (or from) the latent space in a frame-wise manner. Specifically, our encoder $\mathcal{E}$ downsamples the input RGB-image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ with a downsampling factor of 8 and outputs a latent representation $\mathbf{z} \in \mathbb{R}^{4\times H'\times W'}$, where $H' = H/8, W' = W/8$, enabling the denoiser model to work on a much lower dimensional data and thus improves the running time and memory efficiency. The latent diffusion model is conditioned on both the motion structure and text prompt. Given a video clip $\mathbf{x} \in \mathbb{R}^{L\times 3\times H\times W}$ with L frames, we obtain its latent representation $\mathbf{z}$ through $\mathbf{z} = \mathcal{E}(\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^{L\times 4\times H'\times W'}$. In the fixed forward process, $\mathbf{z} \coloneqq \mathbf{z}_0$ is diffused into a pure noise tensor $\mathbf{z}_T$ by T steps. In the reverse process (namely the denoising process), the denoiser model predicts the previous-step data $\mathbf{z}_{t-1}$ from current noisy data $\mathbf{z}_t$ by taking the embedded text prompt and frame-wise depth maps as conditions, and a clean data $\mathbf{z}'_0$ can be sampled from random noise $\mathbf{z}_T$ in a recurrent manner. Specifically, the denoiser model is a 3D U-Net which we adopt the architecture from [He et al. 2022]. We adopt CLIP [Radford et al. 2021] as the textual encoder to extract visual-aligned tokens from the input text prompts. For the depth estimator, we choose the Midas depth estimation model [Ranftl et al. 2022] due to its robustness on various videos. The depth maps are encoded through a CNN-based structure encoder and the multi-scale features are added to the feature maps of the denoiser U-Net for structural modulation. Different from structure control, semantic control via textual prompt affects the backbone features via a cross-attention module[Rombach et al. 2022].

3.3 Video Character Rerendering

Above mentioned video synthesis framework manages to provide videos with high-quality and diverse motion. However, the generated character appearance controlled by text prompts varies in different video clips. To overcome this challenge, we formulate this problem with the following objective: Given a pre-trained video generation model, and user-specified characters, our goal is to generate consistent characters across different video clips; we referred to this task as *video character rerendering*. To do so, we examined existing literature on personalization approaches of image diffusion models. However, there are several challenges associated with directly applying these methods for video personalization. 1) How to leverage image data for personalizing video models? One straightforward approach for video personalization is to utilize video data that portrays a specific character. However, the video data of a consistent character is much harder to collect than images. 2) How to adjust the tradeoff between the concept compositionality and the character fidelity? This challenge also exhibits in image personalization literature. In this section, we will explain preliminary approaches and our method in detail.

Preliminary: Textual Inversion. Textual Inversion is an image personalization approach that aims to represent a new concept to a new token S_* and learns a corresponding new token embedding vector v_* in the CLIP text encoder c_θ. The v_* is directly optimized with 3-10 images depicting a specific concept. The training objective is the same as the original loss of diffusion models, which can be defined as:

$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(\mathbf{x}),\mathbf{y},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon - \mathbf{f}_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{3}$$

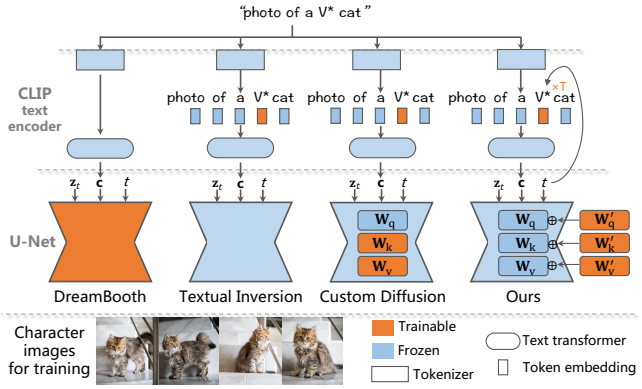

Fig. 4. Concept diagram of different approaches for personalization. To overcome the inconsistency problem of generated characters, we study existing personalization approaches and propose a new method to rerender the appearance of the target character. We keep all parameters from the CLIP text encoder and the denoiser U-Net frozen and learn timestep-dependent token embeddings to represent the semantic features of target characters. Additionally, we insert a new branch to the projection layers of $\mathbf{q}, \mathbf{k}, \mathbf{v}$ in attention modules and modulate the pre-trained weight to better represent the character.

After training, the new token can be combined with other word tokens to form a sentence. This token sequence can then be passed through the text encoder to obtain conditional text token embeddings that facilitate the control of image generation for producing the desired concept.

Timestep-variable textual inversion (TimeInv). However, optimizing the single token embedding vector has limited expressive capacity because of its limited optimized parameter size. In addition, using one word to describe concepts with rich visual features and details is very hard and insufficient. Hence, it tends to suffer from unsatisfactory results regarding the fidelity of concepts. To tackle this problem, we propose timestep-variable textual inversion (TimeInv). TimeInv is based on the observation that different timesteps control the rendering of different image attributes during the inference stage. For example, the previous timesteps of the denoising process control the global layout and object shape, and the later timesteps of the denoising process control the low-level details like texture and color [Voynov et al. 2022]. To better learn the token depicting the target concept, we design a timestep-dependent token embedding table to store the controlling token embedding at all timesteps. During training, we sample random timesteps among all ddpm timesteps to directly optimize the timestep-embedding mapping table $\mathbf{V} \in \mathbb{R}^{T \times d}$ where T is the total timesteps of diffusion process and d is the dimension of token embeddings. The training objective can be defined as:

$$\mathbf{V} := \arg\min_{v_{1:T}} \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - f_\theta(\mathbf{z}_t, t, \mathbf{c}_\theta(y,t))\|_2^2\right]. \quad (4)$$

During inference, the token embedding is retrieved based on the current denoising timestep and then composite to a sequence of token embeddings, defined as $v_*^t = V_t$.

Video customization with image data. Another challenge for video personalization is how to leverage image data to optimize video generation models. Directly repeating images to videos and then optimizing the tokens will lead to the motion omission problem. Since the static motion tends to bind with the target concept and hard to generate concepts with diverse motions. Thanks to the previously introduced structure-guided module, now we can learn the concept using static structure guidance. Specifically, we repeat the concept image to a pseudo video with L frames and extract framewise depth signals to control the video generation model to synthesize static concept videos. During inference, it can easily combine the target concept with other motion guidance to generate a concept with diverse actions.

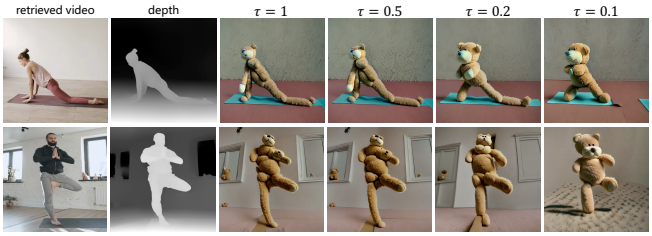

Fig. 5. Effectiveness of adjusting the τ. Small τ can relax the depth control to make the shape render towards the character shape while maintaining a coarse layout and action control from the depth. This technique can generate videos of teddy bears without the need to retrieve the motion video of teddy bears, which is very hard to collect since there is a lack of real videos of teddy bears with diverse motions (e.g., doing yoga).

Low-rank weight modulation. Using textual inversion only is still hard to capture the appearance details of the given character. Instead of previous approaches that directly optimize model parameters, we add additional low-rank [Hu et al. 2021] matrices to the pre-trained linear layers in attention modules, without hurting the concept generation and composition ability in the pre-trained model. The low-rank matrices comprise two trainable linear layers. We insert these matrices in the cross and spatial self-attention modules in our model.

Conflict between structure guidance and concept generation. Although the concept can be successfully injected into video generation with our tailored design, there still exists a severe concept-guidance conflict issue. Specifically, if we want to learn a personalized teddy bear and then use a source video to provide motion guidance, it is challenging and time-consuming to collect teddy bear moving videos. Besides, the shape provided by the depth will severely affect the id similarity because the generated shape needs to follow the id shape. Hence it is crucial for a depth-guidance model to have the ability to relax its depth control. To solve this, we make our depth-guidance module to be adjustable via timestep clamping during sampling. Concretely, we apply depth guidance on the feature only for the timesteps $t = T, ..., \tau$ and drop the depth feature after timestep τ. We also experimented with feature rescaling during inference in early attempts, which shows worse depth adjustment than the timestep clamping.

4 EXPERIMENT

As the storyboard splitting is conducted manually or assisted by LLMs, we mainly evaluate the retrieval-augmented video generation,

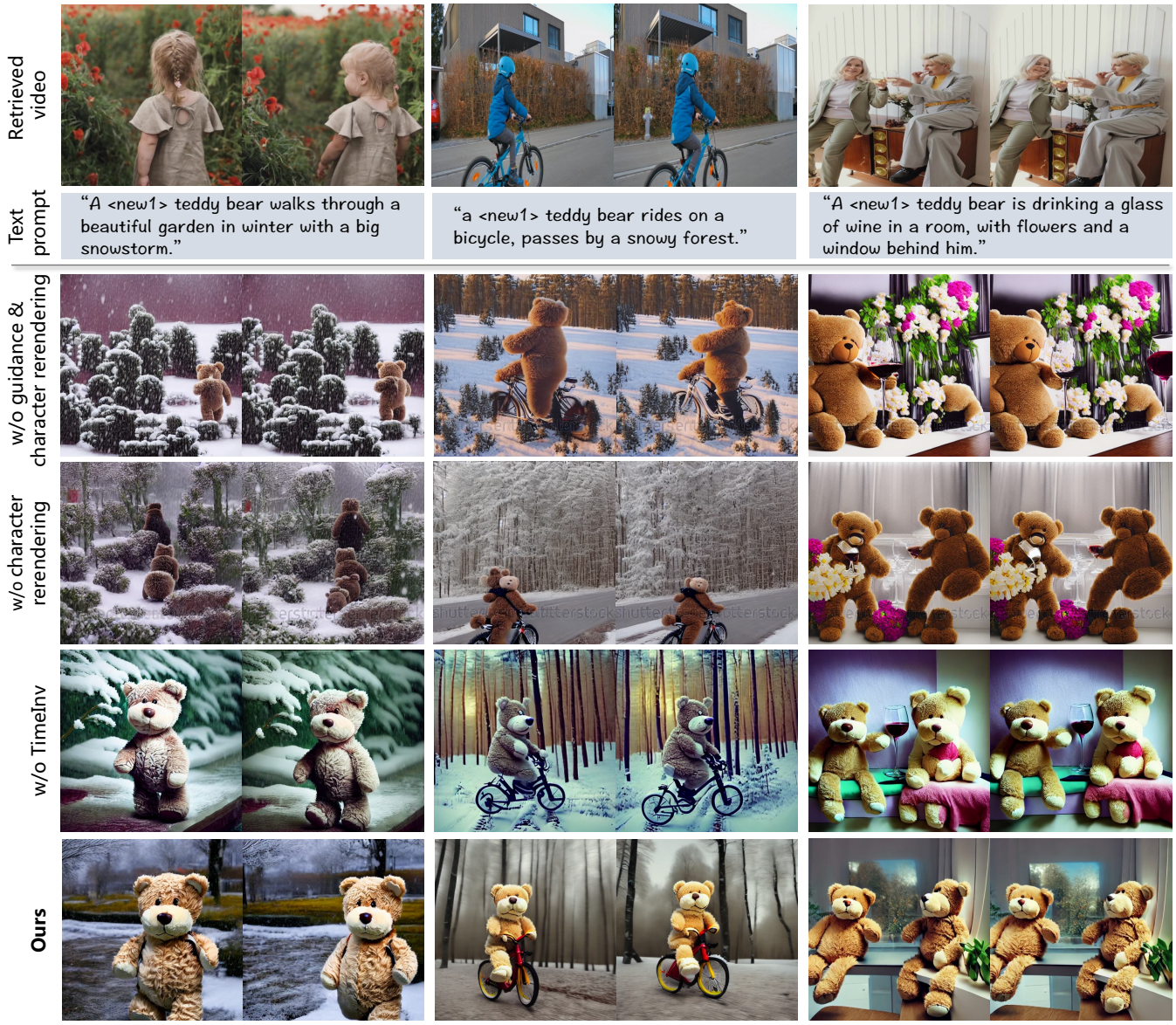

Fig. 6. Abalation results of the core components in our pipeline, including structure guidance, character rerendering, and TimeInv.

as the major technical innovation of this paper. Concretely, we first validate the effectiveness of our overall pipeline for storytelling video synthesis in Sec. 4.2, and then evaluate the video synthesis quality, and the concept customization performance in Sec. 4.3 and Sec. 4.4 respectively.

4.1 Implementation Details

Our video generation model is trained in three stages. Firstly, we train the base text-to-video model on WebVid-10M [Bain et al. 2021] dataset, with the spatial parameters initialized with the publicly available pre-trained Stable Diffusion Image LDM [Rombach et al. 2022]. WebVid-10M consists of 10.7M video-caption pairs with a total of 52K video hours. For training, we resize the videos into resolution 256×256 and sample 16 frames with a frame stride of 8. Secondly, to equip the model with depth guidance, we train the structure encoder on the same training set with the pre-trained base model frozen. At last, for concept customization, the depth-guided text-to-video model is fintuned along with specific textural tokens optimization, which is performed on task-dependent small dataset. Although our model is trained with fixed resolution, we find it support other resolutions well in inference phase.

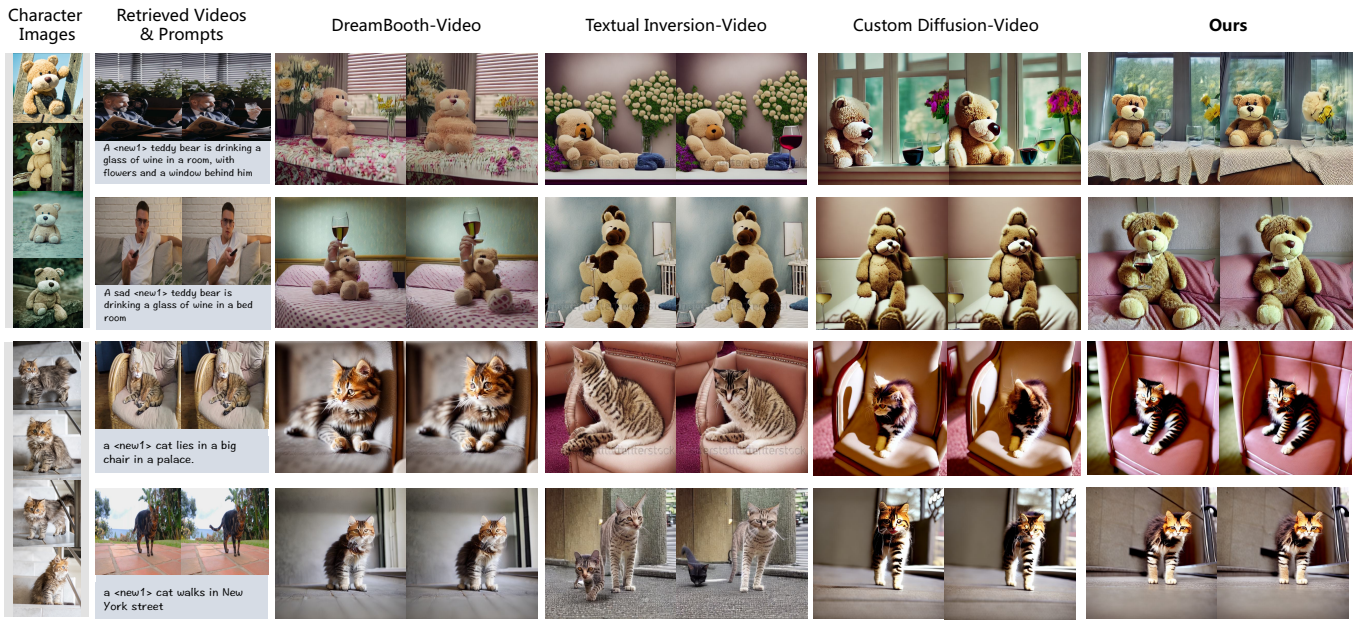

Fig. 7. Quantitative comparisons with previous personalization approaches. We show the results of two characters using four different approaches. For each approach, we show one video clip and fix the same random seed. Each video clip shows two frames with a frame sampling stride of 8. Readers can zoom in for a better view.

4.2 Evaluation on Storytelling Video Synthesis

Since we are the first work to tackle the task of storytelling video synthesis task, there are no existing baselines for comparison. So we conduct ablation experiments to justify the effectiveness of our design choices, as shown in Fig. 6. We can see that the generation quality of the target character deteriorates without using TimeInv. In addition, the concept compositional ability is been damaged without TimeInv. For example, the model fails to generate the flowers and window in the third clip. The absence of a personalization process makes it difficult to maintain consistency of character across various clips. Without the inclusion of video retrieval, it becomes challenging to control the character position and the overall layout. Additionally, the generation performance is inferior compared to retrieval-augmented generation, as evidenced by the quantitative results presented in Table 1. Video results can be checked in supplementary results.

Method	Condition	FVD↓	KVD↓
T2V-Zero + ControlNet [Khachatryan et al. 2023]	text + depth	4685.27	168.20
ModelScope [damo vilab 2023]	text	2616.06	2052.84
LVDM [He et al. 2022]	text	917.63	116.63
Ours	text + depth	**516.15**	**47.78**

Table 1. Quantitative comparison with open-sourcing video generation models on UCF-101 under zero-shot setting.

4.3 Evaluation on Text-to-Video Synthesis

Baselines and evaluation metrics. We compare our approach with existing available (depth-guided) text-to-video models, including T2V-Zero combined with ControlNet to inject depth control [Khachatryan et al. 2023], and two open-sourcing text2video generation models ModelScope [damo vilab 2023] and LVDM [He et al. 2022]. We measure the video generation performance via FVD and KVD. The used real video dataset is UCF-101 [Soomro et al. 2012] and we sample 2048 real videos and their corresponding class names. We use the class name as a text prompt and then sample one fake video for each prompt. For all three models, we use DDIM 50 steps and their default classifier-free guidance scales, which are 9 for ModelScope and T2V-Zero, and 10 for LVDM.

Results. In Tab. 1, we show quantitative results of the performance of video synthesis. As can be seen, equipped with depth structure guidance, text-to-video synthesis achieves significantly better performance than video synthesis from pure text. In addition, our approach also surpasses the existing depth-guided text-to-video generation method, T2V-Zero combined with ControlNet, demonstrating the superiority of our model.

4.4 Evaluation on Personalization

Baselines and evaluation metrics. To evaluate the effectiveness of our personalization module, we compare our approach with the previous three baseline approaches: Dreambooth[Ruiz et al. 2023], Textual inversion [Gal et al. 2022], and Custom Diffusion [Kumari et al. 2023]. For quantitative evaluation, we measure the semantic alignment between generated videos and text and the concept fidelity between generated videos and the user-provided concept images. For each approach, we sample 10 random videos using 20 prompt-video pairs constructing 200 generated videos in total. The

prompts range from different backgrounds and different compositions with other objects. Then the semantic alignment is computed by the cosine similarity between the CLIP text embedding and the CLIP image embedding of each frame, then we average the scores from all frames to obtain the alignment score between a text and a generated video clip. The concept fidelity is measured by the average alignment between generated video clips and corresponding concept images. For each pair, we compute the cosine similarity between each frame of the clip and the target concept image, and then we average the scores from all frames. Different approaches share the same set of random seeds for a fair comparison.

Implementation details. We adopt character images from Custom Diffusion, and we experiment with a teddy bear and a cat, containing 7 and 5 images respectively. We use <new1> as the pseudo word to represent the given character in the input text prompt. For the character of the teddy bear, we train our approach and baseline approaches to 1,000 steps. The learning rate we used for our approach and textual inversion is 1.0e-04, while we use the learning rate of 1.0e-5 for dreambooth and custom diffusion since they directly optimize the pre-trained model parameters. For all these approaches, we use real videos of the same category as the regularization dataset to prevent the overfitting issue. The regularization data are retrieved from the WebVid dataset. For each character, we retrieved 200 real videos. We also use data augmentation of target character images in all approaches to enrich the diversity of the training dataset following Custom Diffusion. During inference, we use DDIM sampling with 50 sampling timesteps and the classifier-free guidance of 15 for all approaches.

Results. In Fig. 7, we present qualitative results of comparisons with baseline personalization approaches in the video generation setting. As can be seen, DreamBooth updates the whole pre-trained parameters thus it tend to suffer from the overfitting problem, hindering its generation diversity (e.g., the background of the third and fourth row is very similar to the training character images). Textual Inversion only optimizes a single token embedding to represent the target character's appearance, so it is hard to capture the character details and exhibits poor character fidelity. Custom Diffusion updates the linear layer for computing k and v in attention modules inside the pre-trained network, combined together with textual inversion. Although it achieves better character fidelity, it frequently shows artifacts in the generated character appearance.

In Tab. 2, we provide quantitative results of comparisons. Our proposed TimeInv can serve as a replacement for Textual Inversion and can be combined with custom diffusion, achieving better semantic alignment.

Besides the video generation setting, we also evaluate the proposed timestep-variable textual inversion in the common image personalization setting. We use the Custom Diffusion codebase and compare the performance between TimeInv and Textual Inversion. The results are shown in Fig. 8. We can see that combining TimeInv with Custom Diffusion shows better background diversity, and concept compositionally (e.g., the ball appears more frequently than the Custom Diffusion + Textual Inversion). Comparing TimeInv with Textual Inversion directly without updating model parameters shows that TimeInv has better character similarity (i.e., the unique texture of the cat).

Method	Teddy Bear		Cat	
	Sem.	ID	Sem.	ID
DreamBooth [Ruiz et al. 2023]-Video	0.272	0.778	0.255	0.869
Textual Inversion [Gal et al. 2022]-Video	0.263	0.666	0.252	0.743
Custom diffusion [Kumari et al. 2023]-Video	0.275	0.849	0.256	0.841
Ours	**0.295**	**0.853**	**0.257**	**0.902**

Table 2. Quantitative comparison with previous personalization approaches.

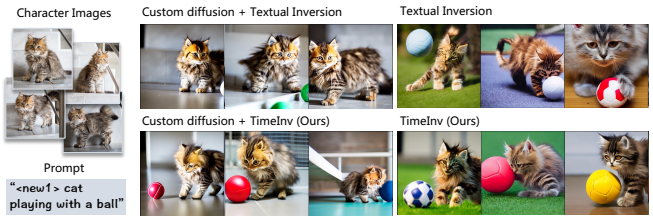

Fig. 8. Effectiveness of the proposed Timestep-variable Textual Inversion (TimeInv) on image personalization using the pre-trained Stable Diffusion. Results in the same column are compared under the same training step and random seeds. This demonstrates that our approach can serve as a general approach for personalization on both image and video generation tasks.

5 CONCLUSION

We introduce a novel retrieval-based pipeline for storytelling video synthesis. This system enables better video synthesis quality, layout, motion control, and character personalization for producing a customized and character-consistent storytelling video. We incorporate a structure-guided video generation module to a base text-to-video model, and we devise a new personalization method to boost the character control performance. We also solve the character-depth confliction problem with the adjustable depth controlling module. While we have shown the ability of our system for the challenging storyline synthesis task, there is much room for future improvement from multiple aspects. For example, a general character control mechanism without finetuning and a better cooperation strategy between character control and structure control can be potential directions.

REFERENCES

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *arXiv preprint arXiv:2305.15391* (2023).

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision (ICCV)*.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.

damo vilab. 2023. modelscope-text-to-video-synthesis. https://huggingface.co/damo-vilab/modelscope-damo-text-to-video-synthesis

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

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

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. 2022. Long video generation with time-agnostic vqgan

and time-sensitive transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer, 102–118.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent Video Diffusion Models for High-Fidelity Video Generation with Arbitrary Lengths. *arXiv preprint arXiv:2211.13221* (2022).

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022b. Video diffusion models. *arXiv preprint arXiv:2204.03458* (2022).

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

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

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778* (2023).

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439* (2023).

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jinren Zhou, and Tieniu Tan. 2023. Decomposed Diffusion Models for High-Quality Video Generation. *arXiv preprint arXiv:2303.08320* (2023).

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. 2023. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186* (2023).

Kangfu Mei and Vishal M Patel. 2022. VIDM: Video Implicit Diffusion Models. *arXiv preprint arXiv:2212.00235* (2022).

OpenAI. 2023. GPT-4 Technical Report. *arXiv* (2023). arXiv:2303.08774

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 3 (2022), 1623–1637.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Masaki Saito, Eiichi Matsumoto, and Shunta Saito. 2017. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*. 2830–2839.

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

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. 2022. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3626–3636.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1526–1535.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399* (2022).

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in neural information processing systems* 29 (2016).

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2022. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752* (2022).

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. 2023a. Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising. arXiv:2305.18264 [cs.CV]

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023b. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018* (2023).

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848* (2023).

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806* (2021).

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. 2023. Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance. *arXiv preprint arXiv:2306.00943* (2023).

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

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *arXiv preprint arXiv:2306.07954* (2023).

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. 2023a. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10459–10469.

Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. 2023b. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18456–18466.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. 2023. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077* (2023).

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018* (2022).