International Journal of Computer Vision (2025) 133:4177–4195
https://doi.org/10.1007/s11263-025-02349-y

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation

Wenjing Wang[1] · Huan Yang[2] · Zixi Tuo[2] · Huiguo He[2] · Junchen Zhu[2] · Jianlong Fu[2] · Jiaying Liu[1]

Received: 1 April 2024 / Accepted: 6 January 2025 / Published online: 24 February 2025

Abstract

With the explosive popularity of AI-generated content (AIGC), video generation has recently received a lot of attention. Generating videos guided by text instructions poses significant challenges, such as modeling the complex relationship between space and time, and the lack of large-scale text-video paired data. Existing text-video datasets suffer from limitations in both content quality and scale, or they are not open-source, rendering them inaccessible for study and use. For model design, previous approaches extend pretrained text-to-image generation models by adding temporal 1D convolution/attention modules for video generation. However, these approaches overlook the importance of jointly modeling space and time, inevitably leading to temporal distortions and misalignment between texts and videos. In this paper, we propose a novel approach that strengthens the interaction between spatial and temporal perceptions. In particular, we utilize a swapped cross-attention mechanism in 3D windows that alternates the "query" role between spatial and temporal blocks, enabling mutual reinforcement for each other. Moreover, to fully unlock model capabilities for high-quality video generation and promote the development of the field, we curate a large-scale and open-source video dataset called HD-VG-130M. This dataset comprises 130 million text-video pairs from the open-domain, ensuring high-definition, widescreen and watermark-free characters. A smaller-scale yet more meticulously cleaned subset further enhances the data quality, aiding models in achieving superior performance. Experimental quantitative and qualitative results demonstrate the superiority of our approach in terms of per-frame quality, temporal correlation, and text-video alignment, with clear margins.

Keywords Text-to-video generation · Diffusion model · Dataset · Large-scale generative model · Video synthesis

1 Introduction

Automated video production is experiencing a surge in demand across various industries, including media, gaming, film, and television (Joshi et al., 2017; Menapace et al., 2021). This increased demand has propelled video generation research to the forefront of deep generative modeling, leading to rapid advancements in the field (Ho et al., 2022b; Mathieu et al., 2016; Saito et al., 2017; Tulyakov et al., 2018; Vondrick et al., 2016). In recent years, diffusion models (Ho et al., 2020) have demonstrated remarkable success in generating visually appealing images in open-domains (Esser et al., 2024; Podell et al., 2024; Rombach et al., 2022; Ramesh et al., 2022). Building upon such success, in this paper, we take one step further and aim to extend their capabilities to high-quality text-to-video generation.

Communicated by Shengfeng He.

✉ Huan Yang
hyang@fastmail.com

✉ Jianlong Fu
jianf@microsoft.com

Wenjing Wang
daooshee@pku.edu.cn

Zixi Tuo
v-zixituo@microsoft.com

Huiguo He
v-huiguohe@microsoft.com

Junchen Zhu
v-junchenzhu@microsoft.com

Jiaying Liu
liujiaying@pku.edu.cn

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing, China

[2] Microsoft Research Asia, Beijing, China

As is widely known, the development of open-domain text-to-video models poses grand challenges, due to the limited availability of large-scale text-video paired data and the complexity of constructing space-time models from scratch. To solve the challenges, current approaches are primarily built on pretrained image generation models. These approaches typically adopt space-time separable architectures, where spatial operations are inherited from the image generation model (Ho et al., 2022b; Hong et al., 2022). To further incorporate temporal modeling, various strategies have been employed, including pseudo-3D modules (Singer et al., 2022; Zhou et al., 2022), serial 2D and 1D blocks (Blattmann et al., 2023b; Ho et al., 2022a), and parameter-free techniques like temporal shift (An et al., 2023) or tailored spatiotemporal attention (Khachatryan et al., 2023a; Wu et al., 2023). However, these approaches overlook the crucial interplay between time and space for visually engaging text-to-video generation. On one hand, parameter-free approaches rely on manually designed rules that fail to capture the intrinsic nature of videos and often lead to the generation of unnatural motions. On the other hand, learnable 2D+1D modules and blocks primarily focus on temporal modeling, either directly feeding temporal features to spatial features, or combining them through simplistic element-wise additions. This limited interactivity usually results in temporal distortions and discrepancies between the input texts and the generated videos, thereby hindering the overall quality and coherence of the generated content.

To address the above issues, we take one step further in this paper which highlights the complementary nature of both spatial and temporal features in videos. Specifically, we propose a novel Swapped spatiotemporal Cross-Attention (Swap-CA) for text-to-video generation. Instead of solely relying on separable 2D+1D self-attention (Bertasius et al., 2021) or 3D window self-attention (Liu et al., 2022) that replace computationally expensive 3D self-attention, we aim to further enhance the interaction between spatial and temporal features. Our swap attention mechanism facilitates bidirectional guidance between spatial and temporal features by considering one feature as the query and the other as the key/value. To ensure the reciprocity of information flow, we also swap the role of the "query" in adjacent layers.

By deeply interplaying spatial and temporal features through the proposed swap attention, we present a holistic VideoFactory framework for text-to-video generation. In particular, we adopt the latent diffusion framework and design a spatiotemporal U-Net for 3D noise prediction. To unlock the full potential of the proposed model and fulfill high-quality video generation, we construct a large video generation dataset, named HD-VG-130M. This dataset consists of 130 million text-video pairs from open-domains, encompassing high-definition, widescreen, and watermark-free characters. We conduct additional data processing, taking into account text, motion, and aesthetics, to create a higher-quality subset. This subset has been shown to effectively enhance video generation performance further. Additionally, our spatial super-resolution model can effectively upsample videos to a resolution of 1376×768, thus ensuring engaging visual experience. We conduct comprehensive experiments and show that our approach outperforms existing methods in terms of both quantitative and qualitative comparisons. In summary, our paper makes the following significant contributions:

- We reveal the significance of learning joint spatial and temporal features for video generation, and introduce a novel Swapped spatiotemporal Cross-Attention (Swap-CA) mechanism to reinforce both space and time interactions. It significantly improves the generation quality, while ensuring precisely semantic alignment between the input text and the generated videos.
- We curate the first open-source[1] dataset comprising 130 million text-video pairs to-date, supporting high-quality video generation with high-definition, widescreen, and watermark-free characters. We proceed with additional processing to extract a higher quality subset and delve into the impact of data processing on video generation. We believe this dataset and corresponding analysis will greatly benefit fellow researchers and advance the field of video generation.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related works. Section 3 introduces our proposed HD-VG-130M dataset, analyzes its properties, and introduces the process of constructing a higher-quality subset. Section 4 presents the proposed text-to-video generation model Video Factory and the Swap-CA design. Experimental results and concluding remarks are provided in Sects. 5 and 6, respectively.

2 Related Works

2.1 Text-to-Image Generation

Generating realistic images from corresponding descriptions combines the challenging components of language modeling and image generation. Traditional text-to-image generation methods (Mansimov et al., 2016; Reed et al., 2016; Xu et al., 2018; Zhang et al., 2017) are mainly based on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and are only able to model simple scenes such as birds (Wah et al., 2011). Later work (Ramesh et al., 2021; Ding et al., 2021) extends the scope of text-to-image generation to open

[1] Project: https://github.com/daooshee/HD-VG-130M. Our dataset was released in Jan. 2024.

domains with better modeling techniques and training data on much larger scales. In recent years, diffusion models have shown great ability in visual generation (Dhariwal & Nichol, 2021). For text-to-image multi-modality generation, GLIDE (Nichol et al., 2021), Imagen (Saharia et al., 2022), DALL·E series (Betker et al., 2023; Ramesh et al., 2022), and Stable Diffusion series (Esser et al., 2024; Podell et al., 2024; Rombach et al., 2022) leverage diffusion models to achieve impressive results. Based on these successes, some work extends customization (Ruiz et al., 2023), image guidance (Yang et al., 2023; Zhang & Agrawala, 2023), and precise control (Balaji et al., 2022). This paper further extends diffusion models for video generation.

2.2 Text-to-Video Generation

Additional controls are often added to make the generated videos more responsive to demand (Mathieu et al., 2016; Pan et al., 2017; Wang et al., 2018), and this paper focuses on the controlling mode of texts.

Early text-to-video generation models (Li et al., 2018; Pan et al., 2017) mainly use convolutional GAN models with recurrent neural networks to model temporal motions. Although complex architectures and auxiliary losses are introduced, GAN-based models cannot generate videos beyond simple scenes like moving digits and close-up actions. Recent works extend text-to-video to open domains with large-scale transformers (Yu et al., 2022a) or diffusion models (Ho et al., 2022a). Considering the difficulty of high-dimensional video modeling and the scarcity of text-video datasets, training text-to-video generation from scratch is unaffordable. As a result, most works acquire knowledge from pretrained text-to-image models. CogVideo (Hong et al., 2022) inherits from a pretrained text-to-image model CogView2 (Ding et al., 2022). Imagen Video (Ho et al., 2022a) and Phenaki (Villegas et al., 2022) adopt joint image-video training. Make-A-Video (Singer et al., 2022) learns motion on video data alone, eliminating the dependency on text-video data. To reduce the high cost of video generation, latent diffusion (Rombach et al., 2022) has been widely utilized for video generation (An et al., 2023; Blattmann et al., 2023b; Esser et al., 2023; He et al., 2022a, b; Khachatryan et al., 2023b; Ma et al., 2023; Wu et al., 2022, 2023; Yu et al., 2023; Zhou et al., 2022). MagicVideo (Zhou et al., 2022) inserts a simple adaptor after the 2D convolution layer. Latent-Shift (An et al., 2023) adopts a parameter-free temporal shift module to exchange information across different frames. PDVM (Yu et al., 2023) projects the 3D video latent into three 2D image-like latent spaces. Show-1 (Zhang et al., 2023) combines pixel and latent diffusion. Although the research on text-to-video generation is very active, existing research ignores the inter and inner correlation between spatial and temporal modules. In this paper, we revisit the design of text-driven video generation.

Dataset takes an important role in training text-to-image generative models. Nonetheless, current datasets either lack the necessary scale or quality (Bain et al., 2021), or are inaccessible to the research community (Blattmann et al., 2023a). In this paper, we provide the first open-source high-quality and large-scale dataset.

3 High-Definition Video Generation Dataset

In this section, we construct a large-scale text-video dataset tailored for high-definition, widescreen, and watermark-free video generation. Additionally, We refine the dataset by considering text, motion, and aesthetic factors to create a higher-quality subset.

3.1 Data Collection, Processing and Annotation

Datasets of diverse text-video pairs are the prerequisite for training open-domain text-to-video generation models. However, most of existing text-video datasets are limited in either scale or quality, thus hindering the upper bound of high-quality video generation. Referring to Table 1, MSR-VTT (Xu et al., 2016) and UCF101 (Soomro et al., 2012) only have 10K and 13K video clips respectively. Although large in scale, HowTo100M (Miech et al., 2019) is specified for instructional videos, which has limited diversity for open-domain generation tasks. Despite being appropriate in both scale and domain, the formats of textual annotations in HD-VILA-100M (Xue et al., 2022) are subtitle transcripts, which lack visual contents related descriptions for high-quality video generation. Additionally, the videos in HD-VILA-100M have complex scene transitions, which are disadvantageous for models to learn temporal correlations. WebVid-10M (Bain et al., 2021) has been used in some previous video generation works (Ho et al., 2022a; Singer et al., 2022), considering its relatively large-scale (10M) and descriptive captions. Nevertheless, videos in WebVid-10M are of low resolution and have poor visual qualities with watermarks in the center.

Recently, video generation has attracted considerable attention particularly in the industry, leading to the emergence of several new large-scale text-to-video datasets (Blattmann et al., 2023a; Kondratyuk et al., 2023; Wang et al., 2023). The LVD (Blattmann et al., 2023a) dataset provides 577M annotated video clip pairs and demonstrates the importance of large-scale datasets for video generation. However, as of now, none of these datasets are open source, hindering their use and analysis by other researchers. Recently released, Panda-70M (Chen et al., 2024) is a text-to-video dataset containing 70 million video clips with text annotation. Despite

Table 1 Comparison of different open-source datasets with text-video pairs. Captions are premium-quality text labels for videos. In contrast, class labels tend to be overly simplistic, and subtitles do not synchronize with the visual contents of the video

Dataset	Video clips	Resolution	Domain	Text	Visual filtering
UCF101 (2012)	13K	240p	Human action	Class label	✗
ActivityNet 200 (2015)	23K	–	Human action	Class label	✗
ACAV100M (2021)	100M	360p	Open	Subtitle	✗
HD-VILA-100M (2022)	103M	720p	Open	Subtitle	✗
HowTo100M (2019)	136M	240p	Instructional	Subtitle	✗
YT-Temporal-180M (2021)	180M	–	Open	Subtitle	Motion
MSVD (2011)	2K	–	Open	Caption	Visual text
YouCook2 (2018)	15K	–	Cooking	Caption	✗
MSR-VTT (2016)	10K	240p	Open	Caption	✗
VATEX (2019)	41K	–	Open	Caption	✗
LSMDC (2015)	118K	1080p	Movie	Caption	✗
WebVid-10M (2021)	10M	360p	Open	Caption	✗
Panda-70M (2024)	70M	720p	Open	Caption	✗
HD-VG-130M (ours)	130M	720p	Open	Caption	✗
HD-VG-40M higher-quality subset (ours)	40M	720p	Open	Caption	Visual text, motion, and aesthetics

Of all the open-source datasets available, our HD-VG-130M dataset stands out for its expansive scale, and its labels fulfill the requirements of video generation. Furthermore, while many internet videos are unsuitable for training video generation models, most existing datasets fail to adequately filter visual content. Our 40M subset enjoys higher quality (in aspects of visual text, motion, and aesthetics) and offers videos that meet stricter criteria

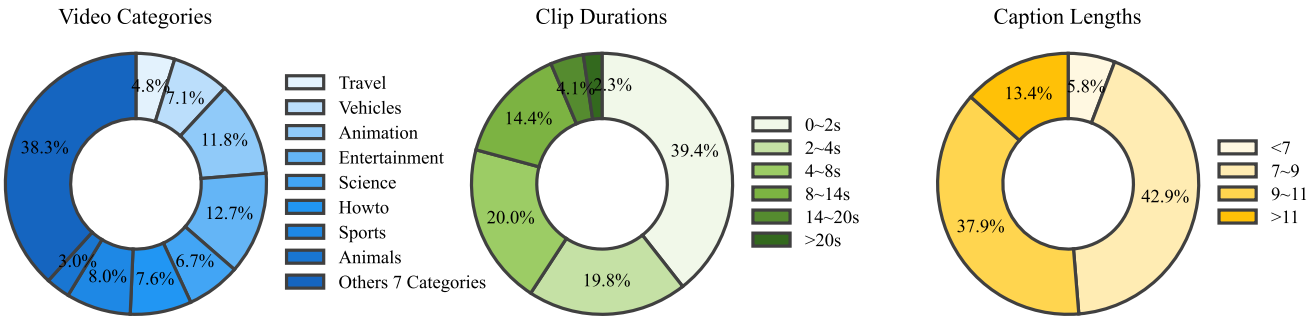


Fig. 1 Statistics of video categories, clip durations, and caption word lengths in HD-VG-130M. HD-VG-130M covers a wide range of video categories

its larger scale compared to existing open-domain datasets, Panda-70M focuses less on data processing, resulting in inappropriate content and limited performance for models trained on it. In Sect. 3.2.4, more detailed discussions are provided.

To tackle the problems above and achieve high-quality video generation, we propose a large-scale text-video dataset, namely HD-VG-130M, including 130M text-video pairs from open-domain in high-definition (720p), widescreen and watermark-free formats. We first collect high-definition videos from YouTube. The challenge lies in converting raw high-definition videos into video-caption pairs, which is far from straightforward. As the original videos have complex scene transitions which are adverse for models to learn temporal correlations, we detect and split scenes in these original videos,[2] resulting in 130M single scene video clips. Finally, we caption video clips with BLIP-2 (Li et al., 2023), in view of its large vision-language pre-training knowledge. To be specific, we extract the central frame in each clip as the keyframe, and get the annotation for each clip by captioning the keyframe with BLIP-2 (Li et al., 2023). Note that the video clips in HD-VG-130M are in single scenes, which ensures that the keyframe captions are representative enough to describe the content of the whole clips in most circumstances. Another method of annotation involves using video captioning techniques. However, we have observed that existing video captioning methods (Xu et al., 2023) often inaccurately describe the visual content, leading to less effective results compared to BLIP-2. We will delve deeper into this issue in Sect. 5.2.2.

The statistics of HD-VG-130M are shown in Fig. 1. The videos in HD-VG-130M cover 15 categories. The wide range of domains is beneficial for training the models to generate diverse content. After scene detection, the video clips are mostly in single scenes with duration less than 20 s. The textual annotations are visual contents related to descriptive captions, which are mostly around 10 words.

[2] We use the open source tool: https://github.com/Breakthrough/PySceneDetect.

3.2 Further Data Processing

Despite detecting and splitting scenes, numerous videos remain unsuitable for training high-quality text-to-video generation models. Given that the videos are sourced from YouTube, a subset of them displays YouTube channel names or subtitles. Additionally, some videos are entirely static or consist of images with simple transformation animations. Such videos negatively impact the training of text-to-video generation models. However, existing text-video datasets overlook the significance of filtering visual content. MSVD (Chen & Dolan, 2011) manually removes videos containing subtitles or overlaid text, but manual processing is impractical for handling large-scale data. YT-Temporal-180M (Zellers et al., 2021) adopts a basic strategy to remove static videos based on their four thumbnails, which is of low precision. Additionally, aesthetic quality is rarely taken into account. HowTo100M (Miech et al., 2019) and HD-VILA-100M (Xue et al., 2022) employ a simplistic approach by retaining videos with high view counts; however, a high view count does not guarantee video quality. In the following section, we provide a detailed discussion on visual filtering and propose methods to address these issues and create a higher-quality subset.

3.2.1 Text Detection

As illustrated in Fig. 2a, since videos are collected from YouTube, some of them contain channel names in the corners or subtitles at the bottom half. These videos may lead the text-to-video model to generate texts that have nothing to do with the video content, which goes against the intended purpose of users.

We utilize optical character recognition to identify and filter these videos. Specifically, we employ the text detector CRAFT (Baek et al., 2019) to locate textual elements. Note that while we want to remove channel names and subtitles, we do not want to remove all videos that contain text, which would reduce the diversity of the dataset. As shown in Fig. 2b,

(a) Videos filtered out (b) Video not filtered out

Fig. 2 Results of our filtering strategy on visual texts. The first row shows the frame contents, and the second row shows the predictions of the text detector. On the right, videos containing text but no channel names in the corner nor subtitles are not filtered out, supporting the diversity of our dataset

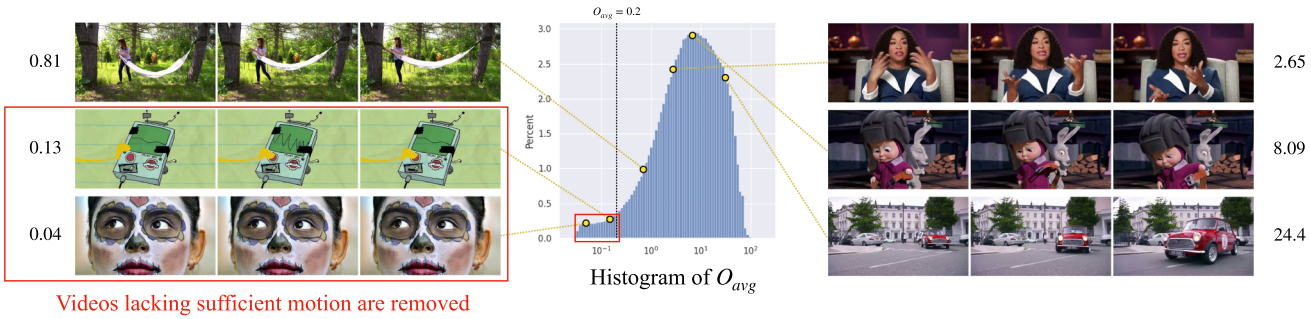
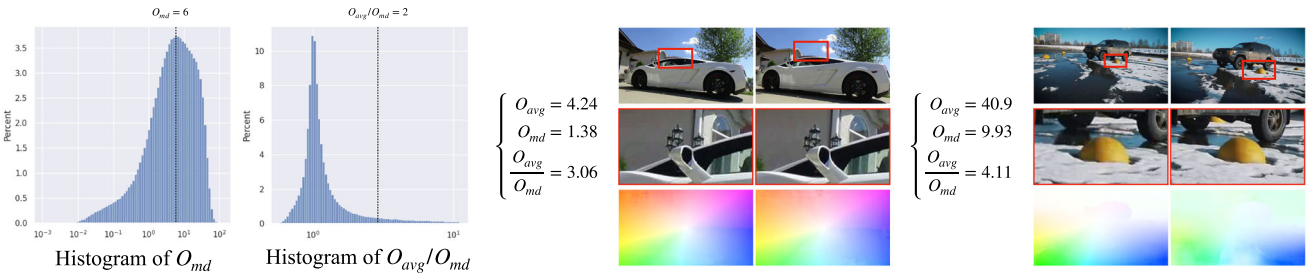


Fig. 3 The distribution of the average optical flow magnitude O_{avg} and representative samples across different O_{avg} values. Videos with $O_{avg} < 0.2$ demonstrate minimal motion, which is unsuitable for training text-to-video models effectively. Hence, we exclude these videos and retain only those with $O_{avg} > 0.2$, which indicate significant motion



(a) Distributions of O_{md} and O_{avg}/O_{md} (b) Image zooming transformation (c) Real-world zooming

Fig. 4 The distribution of the mean deviation of optical flow O_{md}, and the ratio of O_{avg}/O_{md}. Videos with high O_{avg}/O_{md} tend to exhibit consistent optical flows overall, often indicating global translation or scaling. Additionally, among these videos, real-world scenes typically have higher O_{md} values. We show two samples for static image transformation in (**b**) and real-world camera transformation in (**c**). In (**b**), the zoomed details indicate that the scene is static, while in (**c**), the zoomed details show that the relative position of objects has changed. We also show optical flows for better illustration

text can be found on various goods and clothing items, which are quite common in the real world. Therefore, we only consider text within the H_{text} pixel range from the upper, lower, left, and right edges. The keyframes selected are identical to those employed in the aforementioned captioning process. Considering speed and precision, videos are resized to 640 pixels width and H_{text} is set to 60. This strategy results in the removal of 37.33% of videos. Among the remaining videos, 73.36% still contain text, which supports the diversity of our dataset.

3.2.2 Motion Detection

We employ the PWC-Net optical flow estimator (Sun et al., 2018) to analyze video motion. To minimize computational demands, videos are sampled at a rate of 2 frames per second (FPS). Two scores are computed: the average optical flow magnitude (O_{avg}) and the mean deviation of optical flows (O_{md}). Videos shorter than 2 s lack sufficient frames for extraction at 2 FPS, so we exclude them when constructing the higher-quality subset.

Fig. 5 Video-caption pairs in various datasets. From top to bottom: HD-VILA-100M, HD-VG-130M (excluding HD-VG-40M), and HD-VG-40M. Compared to HD-VG-130M, HD-VILA-100M videos lack coherence in both visuals and accompanying text. In HD-VG-40M, static scenes and meaningless text are filtered out, enhancing the dataset's quality for text-to-video generation

Generally, the distribution of real-world optical flow magnitude O_{avg}, should be similar to the Gaussian distribution. However, as shown in Fig. 3, the area within the red box does not conform to the tail shape of a Gaussian distribution. This is because internet videos may contain scenes that depict an image, and corresponding video clips may remain completely still. These cases of insufficient motion could mislead the video generative model. To eliminate these instances, we apply a filtering rule of $O_{avg} > 0.2$, resulting in the removal of 3.71% of videos.

Some scenes may not be static, but rather consist solely of an image with translation or scaling transformations. An example is shown in Fig. 4b, where an image of a car is slowly zoomed in. These movements are overly simplistic and fail to accurately reflect real-world object motion, thereby diminishing the effectiveness of video generative models. Such global transformations can be readily identified using the ratio of O_{avg}/O_{md}. O_{md} signifies the diversity across frames in the optical flow. When the value of O_{md} is lower than O_{avg}, it indicates that the motion across frames is largely uniform, typically indicative of global transformations such as translation and scaling.

One issue with this filtering strategy is that video clips involving camera zooming, scaling, and translation also demonstrate high O_{avg}/O_{md} values. We observe that in such real-life scenarios, changes in the viewing angle can cause shifts in object occlusion relationships. This is illustrated in Fig. 4c, where the background surrounding the yellow ball undergoes a change. These variations in content lead to inconsistent optical flow across frames, resulting in relatively high O_{md} values. Therefore, we finally employ a filtering strategy in which we keep videos satisfying either $O_{avg}/O_{md} < 2$ or $O_{md} > 6$, which is able to remove image transformation animations while retaining real-world camera transformations. It removes 9.58% of videos from the dataset.

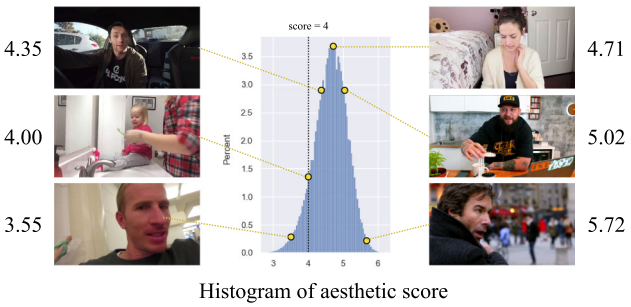


Fig. 6 The distribution of aesthetic scores alongside samples depicting the same theme (human) across various aesthetic scores. Videos with scores above 4 exhibit good aesthetic quality

6.09 6.41 6.83

Fig. 7 Samples from our dataset with aesthetic scores exceeding 6. Art films typically score around 6, while videos with scores exceeding 6.5 mostly feature people drawing

3.2.3 Aesthetics Evaluation

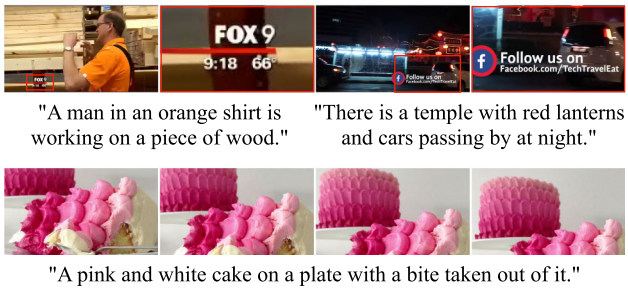
We apply the widely-used LAION-Aesthetics Predictor V2[3] to evaluate the aesthetics of video frames. The distribution of aesthetic scores is shown in Fig. 6. We also provide samples of humans to compare aesthetic effects within the same theme. Videos with aesthetic scores of 4 and below are usually uploaded by ordinary users. Although the content is clear, the composition and lighting are relatively random, and the contrast is low. Videos with an aesthetic score around 4.7, i.e., the majority of the dataset, have standard composition and aesthetic effects in line with mainstream aesthetics. Videos with an aesthetic score closer to 6 have more artistic effects, such as asymmetrical composition or exaggerated background blurring. To enhance the beauty of the data, we filtered out the samples with an aesthetic score below 4 and removed 9.37% of the videos.

Very high-quality videos, such as art film slices, typically have an aesthetic score of around 6. Few samples have aesthetic scores of 6.5 and above. This is because the LAION-Aesthetics Predictor V2 tends to give higher scores to artistic paintings rather than realistic scenes. For videos with aesthetic scores higher than 6.5, many of them are about static painting images, which are removed by our motion filtering in Sect. 3.2.2. The remaining videos mostly depict people painting, as shown in the right two samples in Fig. 7.

3.2.4 Summary

After implementing the aforementioned data processing steps, the HD-VG-130M dataset is refined into a higher-quality subset of 40 million samples, addressing issues such as meaningless texts, lack of movement, and low aesthetics. This subset is named HD-VG-40M. A visualization comparing data samples from HD-VILA-100M, HD-VG-130M, and HD-VG-40M is presented in Fig. 5. In comparison to HD-VG-130M, the videos in HD-VILA-100M lack semantic coherence, and the accompanying text fails to describe the video contents. In HD-VG-40M, videos containing static scenes and meaningless text are further filtered out, resulting in higher-quality data for text-to-video generation. Despite removing more than half of the samples, our higher-quality subset remains larger than most of the existing open-source text-to-video generation datasets, as shown in Table 1. We will demonstrate later that fine-tuning with our higher-quality subset can further enhance the performance of video generation.

[3] https://github.com/christophschuhmann/improved-aesthetic-predictor.



"A man in an orange shirt is working on a piece of wood." "There is a temple with red lanterns and cars passing by at night."

"A pink and white cake on a plate with a bite taken out of it."

Fig. 8 The Panda-70M dataset contains video samples where the text is independent of the picture content, as well as videos consisting solely of the translation of static images

Panda-70M (Chen et al., 2024) is a recently released large-scale text-video dataset, making a significant contribution to the AIGC community. Panda-70M focuses more on meticulous video captioning but somehow neglects the importance of visual content filtering. Both our dataset and Panda-70M collect data from YouTube. However, as discussed above, not all internet videos are suitable for training video generation models, leading to improper samples in Panda-70M, as illustrated in Fig. 8. In comparison, our work conducts detailed processing on the visual contents, filling the gaps left by open-source data in this area. Moreover, we introduce a novel spatiotemporal interaction strategy to enhance model design. Consequently, our model exhibits enhanced visual quality and text-video alignment compared to the model trained on Panda-70M in Tables 6 and 7.

4 High-Quality Text-to-Video Generation

In this section, we introduce how we build the text-to-video generation framework. We first describe how we reinforce both spatial and temporal interactions. Then, we introduce the detailed architecture of our model and the super-resolution processing for generating high-definition videos.

4.1 Spatiotemporal Connection

To reduce computational costs and leverage pretrained image generation models, space-time separable architectures have gained popularity in text-to-video generation (Ho et al., 2022b; Hong et al., 2022). These architectures handle spatial operations independently on each frame, while temporal operations consider multiple frames for each spatial posi-

tion. In the following, we refer to the features predicted by 2D/spatial modules in space-time separable networks as "spatial features", and "temporal features" vice versa.

The quality of spatiotemporal features is important for video generation, as it can affect temporal consistency and text-content alignment performance (Hong et al., 2022; Ho et al., 2022a). The interaction between spatial and temporal features is also essentially, as it determines how the spatial and temporal features are combined. This interaction has been highlighted in previous video-related studies (Bertasius et al., 2021; Zeng et al., 2020) and verified in cross-modality learning (Gu et al., 2023; Ruan et al., 2023). However, as discussed in Sect. 1, prior works have neglected the crucial interaction between spatial and temporal features. The methodologies of existing spatiotemporal strategies are illustrated in Fig. 9a–c. None of them capture the interaction between spatial and temporal features. To address this limitation, we propose the mutual reinforcement of these features through a series of cross-attention operations. As shown in Fig. 9d, our swap attention mechanism enhances bidirectional guidance between spatial and temporal features by treating one feature as the query and the other as the key/value. To ensure the reciprocity of information flow, we also interchange the role of the "query" in adjacent layers. In the following, we introduce the details of this design.

First, denote a basic operation

$$\text{CrossAttention}(x, y) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \tag{1}$$

with

$$Q = W_Q^{(i)} \cdot x, \quad K = W_K^{(i)} \cdot y, \quad V = W_V^{(i)} \cdot y, \tag{2}$$

where $W_Q^{(i)}$, $W_K^{(i)}$, and $W_V^{(i)}$ are learnable projection matrices in the i-th layer. The direction of cross-attention, specifically whether Q originates from spatial or temporal features, plays a decisive role in determining the impact of cross-attention. In general, spatial features tend to encompass a greater amount of contextual information, which can improve the alignment of temporal features with the input text. On the other hand, temporal features have a complete receptive field of the time series, which may enable spatial features to generate visual content more effectively. To leverage both aspects effectively, we propose a strategy of swapping the roles of Q and K, V in adjacent two blocks. This approach ensures that both temporal and spatial features receive sufficient information from the other modality, enabling a comprehensive and mutually beneficial interaction.

Global attention greatly increases the computational costs in terms of memory and running time. To improve efficiency, we conduct 3D window attention. Given a video feature in the shape of $F \times H \times W$ and a 3D window size of $F_w \times H_w \times W_w$, we organize the windows to process the feature in a non-overlapping manner, leading to $\lceil \frac{F}{F_w} \rceil \times \lceil \frac{H}{H_w} \rceil \times \lceil \frac{W}{W_w} \rceil$ distinct 3D windows. Within each window, we perform spatiotemporal cross-attention. By adopting the 3D window scheme, we effectively reduce computational costs without compromising performance.

Following prior text-to-image arts (Blattmann et al., 2023b; Rombach et al., 2022), we incorporate $2\times$ down/upsampling along the spatial dimension to establish a hierarchical structure. Furthermore, research (Habibian et al., 2019; Pessoa et al., 2020) has pointed out that the temporal dimension is sensitive to compression. In light of these considerations, we do not compress the temporal dimension and conduct shift windows (Liu et al., 2022), which advocates an inductive bias of locality. On the spatial dimension, we do not shift since the down/upsampling already introduces connections between neighboring non-overlapping 3D windows.

To this end, we propose a Swapped spatiotemporal Cross-Attention (Swap-CA) in 3D windows. Let t^l and s^l represent the predictions of 2D and 1D modules. We utilize Multi-head Cross Attention (MCA) to compute their interactions by Swap-CA as

$$\begin{aligned}
\tilde{s}^l &= \text{Proj}_{in}^l \odot \text{GN}(s^l);\\
\tilde{t}^l &= \text{Proj}_{in}^l \odot \text{GN}(t^l);\\
h^l &= \text{3DW-MCA}(\text{LN}(\tilde{s}^l),\ \text{LN}(\tilde{t}^l)) + \tilde{s}^l;\\
\bar{h}^l &= \text{FFN} \odot \text{LN}(h^l) + h^l;\\
z^l &= t^l + s^l + \text{Swap-CA}(s^l, t^l)\\
&= t^l + s^l + \text{Proj}_{out}^l(\bar{h}^l),
\end{aligned} \tag{3}$$

where Group Norm (GN), Projection (Proj), Layer Norm (LN), and 3D Window-based Multi-head Cross-Attention (3DW-MCA) are learnable modules. By initializing the output projection Proj_{out}^{l-1} by zero, we have $z^l = t^{l-1} + s^{l-1}$, i.e., Swap-CA is skipped so that it is reduced to a basic addition operation. This allows us to initially train the diffusion model using addition operations, significantly speeding up the training process. Subsequently, we can switch to Swap-CA to enhance the model's performance.

Then for the next spatial-temporal separable block, we apply 3D Shifted Window Multi-head Cross-Attention (3DSW-MCA) and interchange the roles of s and t, as

$$h^{l+1} = \text{3DSW-MCA}(\text{LN}(\tilde{t}^{l+1}), \text{LN}(\tilde{s}^{l+1})) + \tilde{t}^{l+1}. \tag{4}$$

In all 3DSW-MCA, we shift the window along the temporal dimension by $\lceil \frac{F_w}{2} \rceil$ elements. We implement window shifting following Liu et al. (2022).

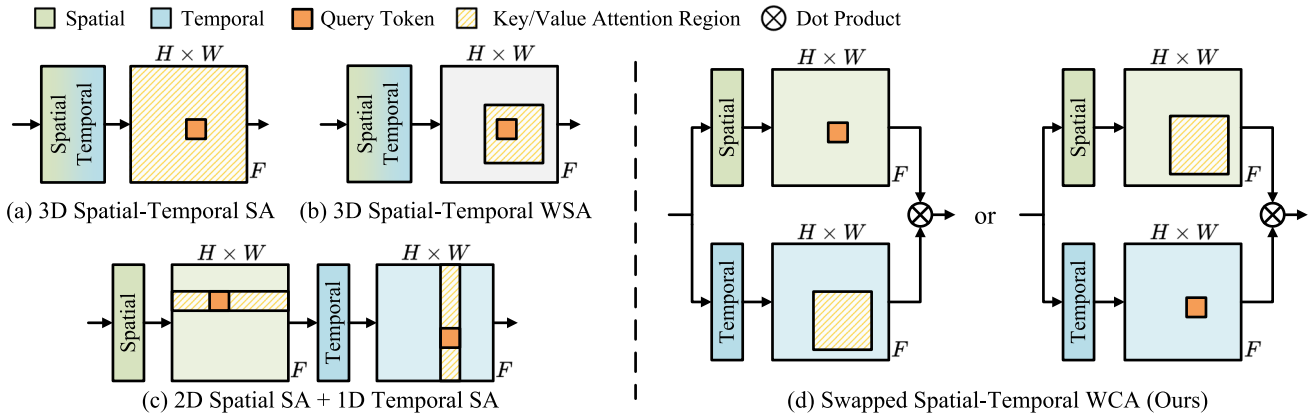


Fig. 9 The paradigm of the proposed Swapped spatiotemporal Cross-Attention (Swap-CA) in comparison with existing video attention schemes. Instead of only conducting self-attention in (**a**)–(**c**), we perform cross-attention between spatial and temporal modules in a U-Net, which encourages more spatiotemporal mutual reinforcement

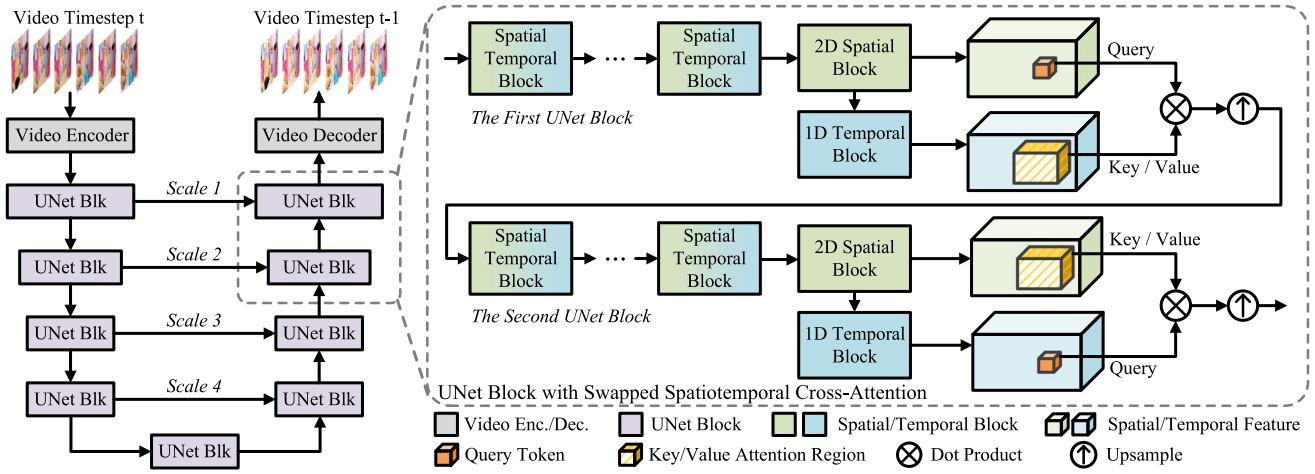


Fig. 10 An illustration of our video diffusion model incorporating Swapped spatiotemporal Cross-Attention (Swap-CA). At the end of each U-Net block, we employ a swapped cross-attention scheme on 3D windows to facilitate a comprehensive integration of spatial and temporal features. In the case of two consecutive blocks, the first block employs temporal features to guide spatial features, while in the second block, their roles are reversed. This reciprocal arrangement ensures a balanced and mutually beneficial interaction between the spatiotemporal modalities throughout the model

4.2 Overall Architecture

We adopt the LDM (Rombach et al., 2022) model as the text-to-image backbone. We employ an auto-encoder to compress the video into a down-sampled 3D latent space. Within this latent space, we perform diffusion optimization using an hourglass spatial-temporal separable U-Net model. Text features are extracted with a pretrained CLIP (Radford et al., 2021) model and inserted into the U-Net model through cross-attention on the spatial dimension.

The detailed architecture of our framework is illustrated in Fig. 10. To balance performance and efficiency, we use Swap-CA only at the end of each U-Net encoder and decoder block. In other positions, we employ a straightforward fusion technique using a $1 \times 1 \times 1$ convolution to merge spatial and temporal features. To enhance the connectivity among temporal modules, we introduce skip connections that connect temporal modules separated by spatial down/upsampling modules. This strategy promotes stronger integration and information flow within the temporal dimension of the network architecture.

4.3 Super-Resolution Towards Higher Quality

To obtain visually satisfying results, we further perform Super-Resolution (SR) on the generated video. One key to improving SR performance is designing a degradation model that closely resembles the actual degradation pro-

Table 2 Ablation study on spatiotemporal interaction strategies. We report the FVD (Unterthiner et al., 2018) and CLIPSIM (Radford et al., 2021) on 1K samples from the WebVid-10M (Bain et al., 2021) validation set

Attention Type	Q	K, V	Param. (G)	Mem. (GB)	Time (ms)	FVD ↓	CLIPSIM ↑
–	–	–	1.480	9.37	135.35	566.16	0.3070
Global	T	S	1.601	22.96	202.12	555.35	0.3091
	S	T	1.601	22.96	205.00	496.25	0.3073
	Swapped		1.601	22.96	201.51	485.86	0.3092
3D Window	T	S	1.601	9.83	150.49	563.12	0.3086
	S	T	1.601	9.83	149.93	490.60	0.3076
	Swapped		1.601	9.83	148.24	**475.09**	**0.3107**

Computational cost is evaluated on inputs of shape $4 \times 16 \times 32 \times 32$. Details can be found in the appendix. T and S represent spatial and temporal features, respectively
Bold indicates the best scores

cess (Wang et al., 2021). In our scenario, the generated video quality suffers from both the diffusion and auto-encoder processes. Therefore, we adopt the hybrid degradation model in Real-ESRGAN (Wang et al., 2021) to simulate possible quality degradation caused by the generated process. During training, an original video frame is downsampled and degraded using our model, and the SR network attempts to perform SR on the resulting low-resolution image. We adopt RCAN (Zhang et al., 2018) with 8 residual blocks as our SR network. It is trained with a vanilla GAN (Goodfellow et al., 2014) to improve visual satisfaction. With a suitable degradation design, our SR network can further reduce possible artifacts and distortion in the frames, increase their resolution, and improve their visual quality.

5 Experiments

In this section, we present the experimental results on text-to-video generation. We first introduce the implementation details, then provide an analysis of method design, and finally compare the performance with existing methods.

5.1 Implementation Details

Our model predicts 16 frames at a resolution of 344×192 (with a latent space of 16×43×24). Then a 4×upscaling is produced in our SR model, resulting in a final output resolution of 1376 × 768. Our model is trained with 32 NVIDIA V100 GPUs. We utilize our HD-VG-130M as training data to promote the generation visual qualities. Furthermore, considering that the textual captions in HD-VG-130M are annotated by BLIP-2 (Li et al., 2023), which may have some discrepancies with human expressions, we adopt a joint training strategy with WebVid-10M (Bain et al., 2021) to ensure the model could generalize well to diverse humanity textual inputs. This approach allows us to benefit from the large-scale text-video pairs and the superior visual qualities of HD-VG-130M while maintaining the generalization ability to diverse textual inputs in real scenarios, enhancing the overall training process. Our model is finally fine-tuned on the HD-VG-40M subset to further promote the performance. More details can be found in the appendix.

5.2 Ablation Studies

In this section, we conduct in-depth analyses of the designs of our text-to-video generation model and the construction of our dataset.

5.2.1 Spatiotemporal Inter-Connection

We first evaluate the design of our swapped cross-attention mechanism. As shown in Table 2, using temporal features as Q generally leads to better CLIP similarity (CLIPSIM) (Radford et al., 2021), revealing a better text-video alignment. The reason might be that language cross-attention only exists in spatial modules. Thus, using spatial features to guide temporal ones implicitly enhance semantic guidance. Reversely, using spatial as Q leads to significantly better FVD, revealing better video quality. The reason might be that the spatial features can better perceive the overall video by using temporal features as guidance. This experiment demonstrates the benefits of introducing cross-attention, as well as the different acts of spatial and temporal features. Combining these two aspects, we propose to swap the roles of x and y every two blocks. In this way, both the temporal and spatial features can get sufficient information from the other modality, leading to improved FVD and CLIPSIM scores. 3D window attention not only significantly lowers computational costs but also leads to a slight performance improvement. Previous studies (Li et al., 2021; Wang et al., 2022) have observed similar performance improvements by integrating a module to enhance local information within transformer-like structures. We show comparative examples in Fig. 11. These results demonstrate how our cross-attention design distinctly enhances scene quality and video dynamics.

We conduct comparisons with other attention strategies in Table 3. We re-implement these designs within our framework, i.e., replacing our proposed swapped cross-attention

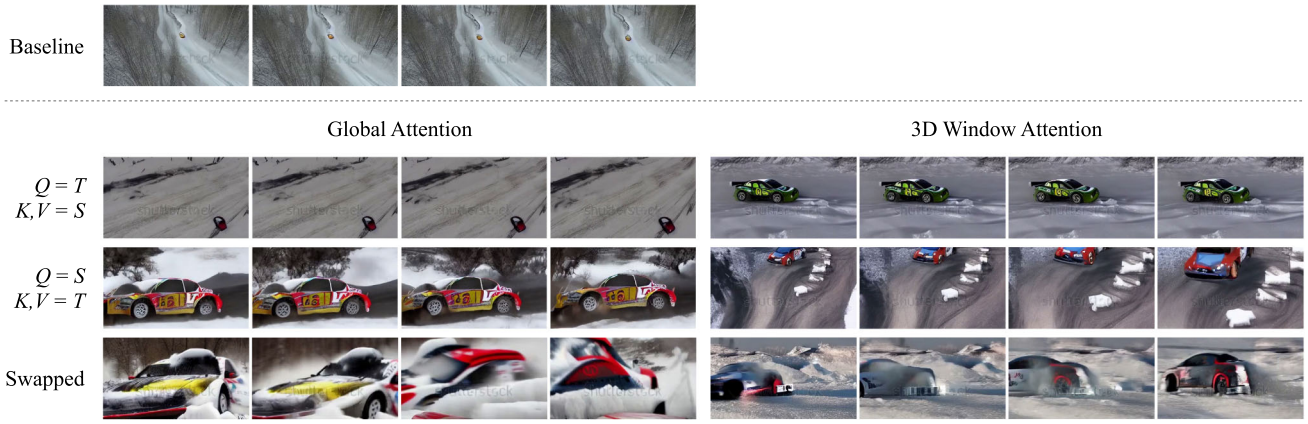


Fig. 11 Subjective ablation study results. The input prompt is "Rally racing car ice racing, realistic"

Table 3 Ablation study on attention strategies. The baseline is the same as the one in Table 2

Methods	FVD ↓	CLIPSIM ↑
Baseline	566.16	0.3070
Tune-A-Video (2023)	717.34	0.3084
CogVideo (2022)	534.48	0.3010
3D Spatiotemporal WSA	500.49	0.3072
Swap-CA (ours)	**475.09**	**0.3107**

Bold indicates the best scores

with them. Specifically, 3D spatial-temporal WSA is realized by first adding spatial and temporal features together and then applying 3D window self-attention. For Tune-A-Video (Wu et al., 2023), we use its ST-Attn, which models temporal consistency by querying relevant positions in previous frames. For CogVideo (Hong et al., 2022), we use its dual-channel attention, where the spatial and temporal features are parallel and then mixed with weighting factors that sum to 1. All other settings remain consistent with the setting in Table 2. The custom attention mechanism utilized in the one-shot model, Tune-A-Video (Wu et al., 2023), appears to be less effective in the open-domain setting. This might be because manually defining the modeling of temporal consistency limits the capacity of generation. While CogVideo (Hong et al., 2022) and 3D spatial-temporal WSA surpass the baseline, they bring less performance improvement compared with our Swap-CA. This shows that the capacity of our swapped cross-attention approach is superior to 3D window self-attention or feature fusion for spatiotemporal interaction.

We further evaluate the effect of different window sizes. The final window size is set to $8 \times 3 \times 6$, i.e., $F_w = 8$, $H_w = 3$, and $W_w = 6$. The rationale behind choosing $H_w = 3$ and $W_w = 6$ is to match the spatial resolution of the core feature in U-net, ensuring that the window attention in the core block can fully perceive the video contents. As for F_w, we set it to 8 to achieve a broader temporal attention view while reducing computation complexity. Table 4 shows the ablation study we performed on window sizes, following the experimental setup in Table 2 of our main paper. Due to NVIDIA software differences, the memory values are not the same in Table 2 and Table 4. Our final configuration, $8 \times 3 \times 6$, achieves the best FVD, CLIPSIM scores and comparable efficiency. Performance can be adversely affected by window sizes that are too small, as they may limit the model's perception. Conversely, for larger window sizes, the model might already have a comprehensive understanding of the global features through the 2D and 1D self-attention modules implemented prior to the swapped cross-attention. Note that, unlike previous works (Liu et al., 2021) that use window attention for almost all layers, our window attention is only used in the proposed swapped cross-attention, accounting for just 7% of the total parameters. As such, large window sizes at this stage could be redundant. It's also important to note that local features and their interactions are crucial in many contexts, and smaller window sizes can help prevent important local features from being overlooked. Ultimately, a window size of $8 \times 3 \times 6$ strikes a good balance between these considerations, effectively capturing both global and local features without redundancy or oversight.

5.2.2 Video Generation Dataset

Visual Contents The advantages of HD-VG-130M extend beyond watermark removal. As shown in Table 5, we evaluate the effect of our HD-VG-130M. After adding HD-VG-130M in training, the result on the validation set of WebVid-10M (Bain et al., 2021) has been improved by 45.34 in FVD. The visual comparison can also be found in Fig. 13. Training with HD-VG-130M not only eliminates watermarks but also elevates the scenic beauty and enriches the level of detail, leading to a comprehensive improvement in the visual quality of the generated videos. These results indicate the comprehensive superiority of our HD-VG-130M in terms of video

Table 4 Ablation study on attention window size

Window Size ($F_w \times H_w \times W_w$)	Param. (G)	Mem. (GB)	Time (ms)	FVD ↓	CLIPSIM ↑
$8 \times 1 \times 3$	1.601	10.07	149.42	525.91	0.3056
$4 \times 3 \times 6$	1.601	10.07	152.14	485.43	0.3064
$8 \times 3 \times 6$ (final setting)	1.601	10.07	153.16	475.09	0.3107
$16 \times 3 \times 6$	1.601	10.07	153.23	487.08	0.3072
Global attention	1.601	23.51	205.58	485.86	0.3092

Table 5 Video generation effect of training on different datasets

Training Data	FVD ↓
w/o HD-VG-130M	475.09
w/ HD-VG-130M	429.75
w/ HD-VG-130M + fine-tuning with higher-quality subset	418.40

Fig. 12 Illustration comparing the impact of training with HD-VG-130M (top row in each group) and subsequent fine-tuning on the HD-VG-40M higher-quality subset (bottom row in each group). Yellow arrows indicate meaningless texts generated by training without HD-VG-40M

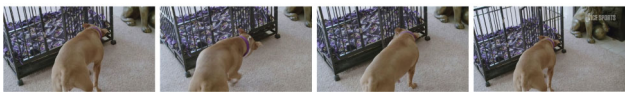
Fig. 13 Generation results without and with using HD-VG-130M for training the model

Fig. 14 Results of the text-to-video model trained on de-watermarked WebVid-10M

quality, diversity, and volume for training text-conditioned video generation models.

Here, we discuss whether watermark removal, if allowed, would solve the problem. For scientific research, we attempted to use E2FGVI (Li et al., 2022) to remove watermarks from WebVid-10M. As shown in Fig. 14, the generated videos have blurry textures. The locations of these blurry areas are in line with the locations of the original watermarks, indicating that the de-watermarking method causes blurriness, and this blurriness damages the training of the video generation model. Furthermore, it's important to consider that removing watermarks without permission can often lead to copyright infringement and legal repercussions. As a result,

mPLUG-2: "A man and woman are talking."
BLIP-2: "A woman **in a black coat** talking to a man."



mPLUG-2: "A dog is in a cage."
BLIP-2: "A dog **standing next to a cage on the floor**."

Fig. 15 Comparison between using mPLUG-2 and BLIP-2 for annotating the contents of video

the best solution currently available is to directly collect non-watermarked videos. This underscores the significance of our HD-VG-130M dataset.

Finally, we assess the effect of additional data processing. As shown in Table 5, fine-tuning with the higher-quality subset enhances the FVD score to 418.40. Furthermore, visual comparisons are presented in Fig. 12. For the top-left sample, training without HD-VG-40M has a risk of generating static scenes, while for the bottom-left sample, the absence of HD-VG-40M in training leads to the space shuttle remaining stationary in each frame, essentially appearing as a translation transformation of a static image. In the case of the right two samples, training without HD-VG-40M may generate meaningless text, as indicated by the yellow arrows. After fine-tuning on the higher-quality subset, these issues are resolved, and the aesthetics of the generated results improve, with better contrast, clearer edges, and more vivid colors.

Video Captions We further evaluated different captioning models. We experimented with a state-of-the-art video captioning model, mPLUG-2 (Xu et al., 2023), but observed that it provides less detailed descriptions (e.g., BLIP-2 predicts "black coat" while mPLUG-2 does not in the first row of Fig. 15) or misinterprets the scene (e.g., mistakes the dog to be inside the cage in the second row of Fig. 15). As a result, using videos captioned with mPLUG-2, the CLIPSIM is decreased to 0.3046.

In addition, we assessed the impact of training with HD-VILA-100M (Xue et al., 2022) instead of HD-VG-130M. As HD-VILA-100M only provides subtitles and lacks scene detection (with potential multiple transitions), significant performance degradation is observed in FVD (429.75 → 692.99) and CLIPSIM (0.3082 → 0.2671), despite joint training with WebVid. This experiment highlights the crucial role of our scene detection and video captioning procedures.

5.2.3 Super Resolution

The results before and after super resolution is presented in Fig. 16. As we can see, before super resolution, the visual content is already pleasing and of high contrast. Our super resolution further makes the edges clearer, such as for the ball of yarn and the man's hair. This result also indicates that high quality mainly comes from the video generation model, showing the importance of training a good text-to-video generation base model.

Table 6 Comparison of video visual quality on the UCF101 dataset

Method	Zero-shot	FVD ↓
VideoGPT (2021)	✗	2880.6
MoCoGAN (2018)	✗	2886.8
MoCoGAN-SG2 (2022)	✗	1821.4
MoCoGAN-HD (2021)	✗	1729.6
DIGAN (2022b)	✗	1630.2
StyleGAN-V (2022)	✗	1431.0
PVDM (2023)	✗	343.6
CogVideo (2022)	✓	701.6
MagicVideo (2022)	✓	699.0
LVDM (2022a)	✓	641.8
ModelScope (2023b)	✓	639.9
Video LDM (2023b)	✓	550.6
LaVie (2023)	✓	526.3
AnimateDiff (2024)	✓	499.3
AnimateDiff+Panda (2024)	✓	421.9
Ours w/o FT	✓	410.0
Ours w/ FT	✓	**398.1**

Bold indicates the best score for zero-shot video generation

5.3 Quantitative and Qualitative Comparison

To thoroughly evaluate the performance of our VideoFactory, we benchmark it on three distinct datasets: the WebVid-10M (Bain et al., 2021) (Val) dataset, which shares the same domain as part of our training data, as well as the UCF101 (Soomro et al., 2012) and the MSR-VTT (Xu et al., 2016) datasets in a zero-shot setting. Note that class-conditioned video generation methods cannot be applied to the MSR-VTT open-domain setting. For methods without released code, we report the scores provided in their papers. We also demonstrated results with and without fine-tuning on the HD-VG-40M higher-quality subset, denoted as "Ours w/o FT" and "Ours w/ FT" respectively. For fair comparison, we did not use super-resolution when evaluating objective performance.

Evaluation on UCF101 As mentioned in Sect. 3, the textual annotations in UCF101 are class labels. We first follow (Ho et al., 2022b; Singer et al., 2022) and rewrite the labels of 101 classes to descriptive captions, and then generate 100 samples for each class. As shown in Table 6, the FVD of our methods reaches 398.1, which achieves the best compared with other methods both in zero-shot setting and beats most

Fig. 16 Generation results before and after super resolution

Table 7 Comparison of text-video alignment on the MSR-VTT dataset

Method	Zero-shot	CLIPSIM ↑
GODIVA (2021)	✗	0.2402
NUWA (2022)	✗	0.2439
LVDM (2022a)	✓	0.2381
CogVideo (2022)	✓	0.2631
ModelScope (2023b)	✓	0.2795
AnimateDiff (2024)	✓	0.2869
AnimateDiff +Panda (2024)	✓	0.2880
Video LDM (2023b)	✓	0.2929
LaVie (2023)	✓	0.2949
Ours w/o FT	✓	0.3005
Ours w/ FT	✓	**0.3021**

Bold indicates the best score for zero-shot video generation

Table 8 Comparison of text-to-video generation performance on the WebVid dataset

Methods	FVD ↓	CLIPSIM ↑
LVDM (2022a)	455.53	0.2751
ModelScope (2023b)	414.11	0.3000
Ours w/ FT	**322.13**	**0.3104**

Bold indicates the best scores

of the methods which have tuned on UCF101. The results verify that our proposed VideoFactory could generate more coherent and realistic videos.

Evaluation on MSR-VTT As shown in Table 7, we also evaluate the CLIPSIM on the widely used video generation benchmark MSR-VTT. We randomly choose one prompt per example from MSR-VTT to generate 2990 videos in total. Although in a zero-shot setting, our method achieves the best compared to other methods with an average CLIPSIM score of 0.3021, which suggests the semantic alignment between the generated videos and the input text. Moreover, note that the state-of-the-art AnimateDiff (2024) training on Panda (2024) performs inferior to ours for both FVD on UCF101 and CLIPSIM on MSR-VTT, demonstrating the effectiveness of both our dataset and model designs.

Evaluation on WebVid-10M (Val) Referring to Table 8, we randomly extract 5K text-video pairs from WebVid-10M which are exclusive from the training data to form a validation set and conduct evaluations on it. Our approach achieves an FVD of 292.35 and a CLIPSIM of 0.3070, outperforming existing methods and showcasing the superiority of our approach.

Subjective Results In Fig. 17, we show comparison results against Make-A-Video, Imagen Video, and Video LDM. The prompts and generated results are collected from their official project website. We also evaluate Gen-2 ,[4] a popular platform in the AIGC field. Make-A-Video only generates 1:1 videos, which limits the user experience. When compared with Imagen Video and Video LDM, our model generates the panda and golden retriever with more vivid details. Despite setting the motion intensity parameter to the maximum, Gen-2 cannot simulate the splashing motion of water. We showcase additional samples of our model in Fig. 18 and more in the supplementary.

Failure Case Study The typical failure case of our text-to-video generation model is that our text encoder, CLIP (Radford et al., 2021), can sometimes misinterpret concepts, leading to unintended results. For instance, with the input prompt "A cat singing in a barbershop quartet," the term "barbershop quartet" signifies musical performance in a specific style. However, our text encoder might inadvertently emphasize "barbershop", introducing a corresponding background to the video. To address this, we can use GPT−3.5 for prompt refinement, after which our model can generate a vivid cat singing on the stage. A visual demonstration (we use the w/o FT version for convenience) can be found in Fig. 19.

6 Conclusion

In this paper, we introduce a high-quality open-domain video generation framework that produces watermark-free, high-definition, widescreen videos. We enhance spatial and

[4] https://research.runwayml.com/gen2.

Fig. 17 Text-to-video generation results compared with Make-A-Video, Imagen Video, Video-LDM, and Gen-2 (Cases of the first three methods are collected from their public project websites)

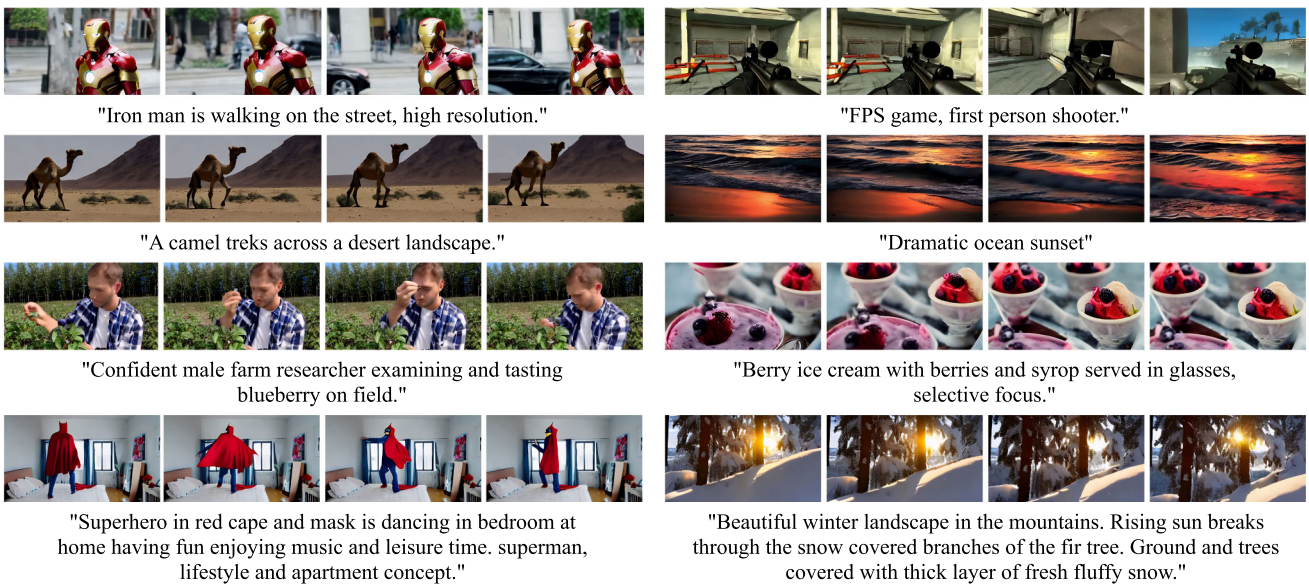
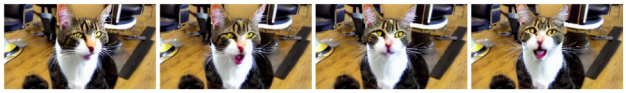


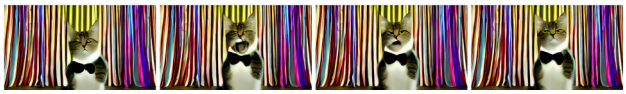
Fig. 18 Samples generated by our VideoFactory (w/ FT) exhibit high quality, featuring clear motion, intricate details, and precise semantic alignment

temporal modeling using a novel swapped cross-attention mechanism, allowing spatial and temporal information to complement each other effectively. Additionally, we provide the HD-VG-130M dataset, featuring 130 million open-domain text-video pairs in widescreen, watermark-free, high-definition format, maximizing the potential of our model. A higher-quality subset is constructed to further promote the performance. Experimental results demonstrate that our method generates videos with superior spatial quality, temporal consistency, and alignment with text. Analysis also demonstrates the effectiveness of our dataset and processing designs.

Future directions for our work may involve refining BLIP-2 captions using large language models and changing the backbone to more powerful text-to-image generation baselines. Another direction is improving data captioning, such as



Original prompt: "A cat singing in a barbershop quartet. 4k HD, vivid"



Revised prompt: "A cat singing in stage, wearing a bow tie. The background is vertical striped curtain. 4k HD, vivid."

Fig. 19 Failure case study of our text-to-video generation model and a quick solution

combining image captions and video captions, or introducing multi-frame captioning like Cap3D (Luo et al., 2024, 2023a) using multi-view images for 3D captioning. The field of video generation has experienced significant growth recently. Due to limited resources, we cannot match the capabilities of some closed-source industrial products. However, we believe that our contributions, particularly the open-source dataset and comprehensive experimental analysis, will benefit the advancement of this field.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-025-02349-y.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

An, J., Zhang, S., Yang, H., Gupta, S., Huang, J. B., Luo, J., & Yin, X. (2023). Latent-Shift: Latent diffusion with temporal shift for efficient text-to-video generation. arXiv.

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In *Conference on Computer Vision and Pattern Recognition*.

Bain, M., Nagrani, A., Varol, G., & Zisserman, A. (2021). Frozen in time: A joint video and image encoder for end-to-end retrieval. In *International Conference on Computer Vision*.

Betker, J., Goh. G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal. P., Chu, C., Jiao, Y., & Ramesh, A. (2022). eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv.

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding?

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., & Ramesh, A. (2023). Improving image generation with better captions.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., & Rombach, R. (2023a). Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023b). Align your latents: High-resolution video synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*.

Chen, D. L., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics*.

Chen, T. S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H. W., Jeon, B. E., Fang, Y., Lee, H. Y., Ren, J., Yang, M. H., & Tulyakov, S. (2024). Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Conference on Computer Vision and Pattern Recognition*.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Conference and Workshop on Neural Information Processing Systems*.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., & Tang, J. (2021). CogView: Mastering text-to-image generation via transformers. In *Conference and Workshop on Neural Information Processing Systems*.

Ding, M., Zheng, W., Hong, W., & Tang, J. (2022). CogView2: Faster and better text-to-image generation via hierarchical transformers. In *Conference and Workshop on Neural Information Processing Systems*.

Esser, P., Chiu, J., Atighehchian, P., et al. (2023). Structure and content-guided video synthesis with diffusion models. arXiv.

Esser, P., Kulal, S., Blattmann, A., et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. arXiv.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2014). Generative adversarial nets. In *Conference and Workshop on Neural Information Processing Systems*.

Gu, B., Fan, H., & Zhang, L. (2023). Two birds, one stone: A unified framework for joint learning of image and video style transfers. In *International Conference on Computer Vision*.

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., & Dai, B. (2024). Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. International Conference on Learning Representations.

Habibian, A., van Rozendaal, T., Tomczak, J. M., & Cohen, T. (2019). Video compression with rate-distortion autoencoders.

He, Y., Yang, T., Zhang, Y., Shan, Y., & Chen, Q. (2022a). Latent video diffusion models for high-fidelity long video generation. arXiv.

He, Y., Yang, T., Zhang, Y., Shan, Y., & Chen, Q. (2022b). Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv.

Heilbron, F. C., Escorcia, V., Ghanem, B., & Niebles, J. C. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Conference on Computer Vision and Pattern Recognition*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Conference and Workshop on Neural Information Processing Systems*.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A. A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022a). Imagen Video: High definition video generation with diffusion models. arXiv.

Ho, J., Salimans, T., Gritsenko, A. A., Chan, W., Norouzi, M., & Fleet, D. J. (2022b). Video diffusion models. In *Conference and Workshop on Neural Information Processing Systems*.

Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). CogVideo: Large-scale pretraining for text-to-video generation via transformers. arXiv.

Joshi, B. J., Stewart, K., & Shapiro, D. (2017). Bringing impressionism to life with neural style transfer in *Come Swim*. In *ACM SIGGRAPH Digital Production Symposium*.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi. H. (2023a). Text2video-zero: Text-

to-image diffusion models are zero-shot video generators. In *International Conference on Computer Vision*.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi. H. (2023b). Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. arXiv.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., Cheng, Y., Chiu, M., Dillon, J., Essa, I., Gupta, A., Hahn, M., Hauth, A., Hendon, D., Martinez, A., Minnen, D., Ross, D. A., Schindler, G., Sirotenko, M., Sohn, K., Somandepalli, K., Wang, H., Yan. J., Yang, M., Yang, X., Seybold, B., & Jiang, L. (2023). Videopoet: A large language model for zero-shot video generation. arXiv.

Lee, S., Chung, J., Yu, Y., Kim, G., Breuel, T. M., Chechik, G., & Song. Y. (2021). ACAV100M: automatic curation of large-scale datasets for audio-visual video representation learning.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv.

Li,Y., Min, M. R., Shen, D., Carlson, D. E., & Carin, L. (2018). Video generation from text. In *AAAI Conference on Artificial Intelligence*.

Li, Y., Zhang, K., Cao, J., Timofte, R., & Gool, L. V. (2021). Localvit: Bringing locality to vision transformers. arXiv.

Li, Z., Lu, C., Qin, J., Guo, C., & Cheng, M. (2022). Towards an end-to-end framework for flow-guided video inpainting. In *Conference on Computer Vision and Pattern Recognition*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In *Conference on Computer Vision and Pattern Recognition*.

Luo, T., Rockwell, C., Lee, H., & Johnson, J. (2023a). Scalable 3d captioning with pretrained models. In *Conference and Workshop on Neural Information Processing Systems*.

Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., & Tan. T. (2023b). VideoFusion: Decomposed diffusion models for high-quality video generation. In *Conference on Computer Vision and Pattern Recognition*.

Luo, T., Johnson, J., & Lee, H. (2024). View selection for 3d captioning via diffusion ranking. In *European Conference on Computer Vision*.

Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., & Chen, Q. (2023). Follow your pose: Pose-guided text-to-video generation using pose-free videos. arXiv.

Mansimov, E., Parisotto, E., Ba, L. J., & Salakhutdinov, R. (2016). Generating images from captions with attention. In *International Conference on Learning Representations*.

Mathieu, M., Couprie, C., & LeCun, Y. (2016). Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations*.

Menapace, W., Lathuiliere, S., Tulyakov, S., Siarohin, A., & Ricci, E. (2021). Playable video generation. In *Conference on Computer Vision and Pattern Recognition*.

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen. M. (2021). GLIDE: Towards photo-realistic image generation and editing with text-guided diffusion models. arXiv.

Pan, Y., Qiu, Z., Yao, T., Li, H., & Mei, T. (2017). To create what you tell: Generating videos from captions. In *ACM International Conference on Multimedia*.

Pessoa, J., Aidos, H., Tomas, P., & Figueiredo, M. A. T. (2020). End-to-end learning of video compression using spatio-temporal autoencoders. In *IEEE Workshop on Signal Processing Systems*.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Muller, J., Penna, J., & Rombach, R. (2024). SDXL: improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss. C., Radford, A., Chen, M., & Sutskever. I. (2021). Zero-shot text-to-image generation.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen. M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv.

Reed, S. E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In *Conference on Computer Vision and Pattern Recognition*, pp. 3202–3212.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*.

Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N. J., Jin, Q., & Guo, B. (2023). MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Conference on Computer Vision and Pattern Recognition*.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K. et al. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Conference on Computer Vision and Pattern Recognition*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, S. K. S., Lopes, R. G., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. S., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Conference and Workshop on Neural Information Processing Systems*.

Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal generative adversarial nets with singular value clipping.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y. Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). Make-A-Video: Text-to-video generation without text-video data. arXiv.

Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). StyleGAN-V: A continuous video generator with the price, image quality and perks of stylegan2. In *Conference on Computer Vision and Pattern Recognition*, pp 3626–3636.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv.

Sun, D., Yang, X., Liu, M. Y., & Kautz, J. (2018). PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Conference on Computer Vision and Pattern Recognition*.

Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D. N., & Tulyakov, S. (2021). A good image generator is what you need for high-resolution video synthesis. arXiv.

Tulyakov, S., Liu, M., Yang, X., & Kautz,. J. (2018). MoCoGAN: Decomposing motion and content for video generation. In *Conference on Computer Vision and Pattern Recognition*.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2018). Towards accurate generative models of video: A new metric and challenges. arXiv.

Villegas, R., Babaeizadeh, M., Kindermans, P., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., & Erhan. D. (2022). Phenaki: Variable length video generation from open domain textual description. arXiv.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *Conference and Workshop on Neural Information Processing Systems*.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie. S. (2011). The Caltech-UCSD birds-200-2011 dataset.

Wang, P., Wang, X., Wang, F., Lin, M., Chang, S., Li, H., & Jin, R. (2022). KVT: k-nn attention for boosting vision transformers. In *European Conference on Computer Vision*.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). Video-to-video synthesis. arXiv.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., & Wang, W. Y. (2019). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research.

Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops, pp. 1905–1914*.

Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., Guo, Y., Wu, T., Si, C., Jiang, Y., Chen, C., Loy, C. C., Dai, B., Lin, D., Qiao, Y., & Liu, Z. (2023). LAVIE: high-quality video generation with cascaded latent diffusion models. arXiv.

Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., & Duan, N. (2021). GODIVA: Generating open-domain videos from natural descriptions. arXiv.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., & Duan, N. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision*.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Hsu, W., Shan, Y., Qie, X., & Shou. M. Z. (2023). Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *International Conference on Computer Vision*.

Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., & Zhou J. mplug-2: A modularized multi-modal foundation model across text, image and video.

Xu, J., Mei, T., Yao, T., & Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In *Conference on Computer Vision and Pattern Recognition*.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*.

Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., & Guo, B. (2022). Advancing high-resolution video-language representation with large-scale video transcriptions. In *Conference on Computer Vision and Pattern Recognition*.

Yan, W., Zhang, Y., Abbeel, P., & Srinivas, A. (2021). Videogpt: Video generation using vq-vae and transformers. arXiv.

Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, & D., Wen, F. (2023). Paint by example: Exemplar-based image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition*.

Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A. G., Yang, M., Hao, Y., Essa, I., & Jiang, L. (2022a) MAGVIT: masked generative video transformer. arXiv.

Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J. W., & Shin. J. (2022b). Generating videos with dynamics-aware implicit generative adversarial networks. arXiv.

Yu, S., Sohn, K., Kim, S., & Shin, J. (2023). Video probabilistic diffusion models in projected latent space. In *Conference on Computer Vision and Pattern Recognition*.

Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., Farhadi, A., & Choi, Y. (2021). MERLOT: multimodal neural script knowledge models. In *Conference and Workshop on Neural Information Processing Systems*.

Zeng, Y., Fu, J., & Chao, H. (2020). Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision*.

Zhang, D. J., Wu, J. Z., Liu, J., Zhao, R., Ran, L., Gu, Y., Gao, D., & Shou, M. Z. (2023). Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv.

Zhang, H., Xu, T., & Li, H. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *International Conference on Computer Vision*.

Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. arXiv.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pp. 286–301.

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., & Feng, J. (2022). MagicVideo: Efficient video generation with latent diffusion models. arXiv.

Zhou, L., Xu, C., & Corso, J. J. (2018). Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*