ART•V: Auto-Regressive Text-to-Video Generation with Diffusion Models

Wenming Weng[1] *, Ruoyu Feng[1], Yanhui Wang[1], Qi Dai[2], Chunyu Wang[2], Dacheng Yin[1],
Zhiyuan Zhao[2], Kai Qiu[2], Jianmin Bao[2], Yuhui Yuan[2], Chong Luo[2] †, Yueyi Zhang[1], Zhiwei Xiong[1]
[1]University of Science and Technology of China [2]Microsoft Research Asia
https://warranweng.github.io/art.v

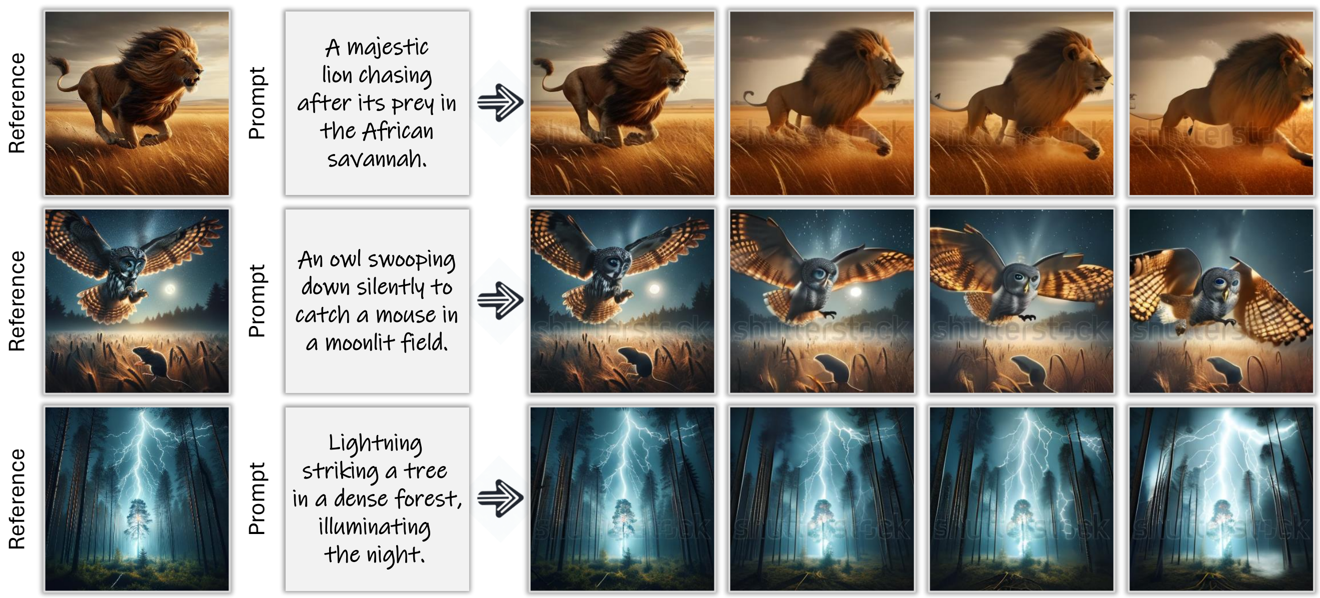

Figure 1. Exemplary results of text-image-to-video generation using our proposed approach, ART•V. Our method skillfully captures object motion while preserving the overall scene, showcasing rich details and maintaining a high level of aesthetic quality. Reference images are generated by DALL-E 3 [1].

Abstract

We present ART•V, an efficient framework for auto-regressive video generation with diffusion models. Unlike existing methods that generate entire videos in one-shot, ART•V generates a single frame at a time, conditioned on the previous ones. The framework offers three distinct advantages. First, it only learns simple continual motions between adjacent frames, therefore avoiding modeling complex long-range motions that require huge training data. Second, it preserves the high-fidelity generation ability of the pre-trained image diffusion models by making only minimal network modifications. Third, it can generate arbitrarily long videos conditioned on a variety of prompts such as text, image or their combinations, making it highly versatile and flexible. To combat the common drifting issue in AR models, we propose masked diffusion model which implicitly learns which information can be drawn from reference images rather than network predictions, in order to reduce the risk of generating inconsistent appearances that cause drifting. Moreover, we further enhance generation coherence by conditioning it on the initial frame, which typically contains minimal noise. This is particularly useful for long video generation. When trained for only two weeks on four GPUs, ART•V already can generate videos with natural motions, rich details and a high level of aesthetic quality. Besides, it enables various appealing applications, e.g. *composing a long video from multiple text prompts.*

* This work is done when the author is an intern with MSRA.
† Corresponding author.

1. Introduction

Recently, text-to-image (T2I) generation [1, 3, 44] has been significantly advanced by generative diffusion models [20, 34, 50–52] and large scale text-image datasets such as Laion5B [47]. The success has also catalyzed a remarkable proliferation of research in text-to-video (T2V) generation [6, 9–16, 18, 21, 22, 26, 30, 39, 49, 58, 59, 61, 62, 66–68, 72, 74–76], driven by the intrinsic allure of the potential breakthroughs.

Existing T2V methods [18, 21] usually adopt a straightforward framework in which they generate entire videos at once using a spatial-temporal U-Net. However, they often produce videos with unrealistic motions. This is because learning the long-range motions is a highly ambiguous and complex task, which requires a significantly larger training dataset than that used in T2I, such as Laion5B [47], which unfortunately is prohibitively expensive to collect and train on. Even the largest video dataset available [7] represents only a fraction of Laion5B. Therefore, we argue that achieving the "stable diffusion" moment in T2V using this framework is difficult.

In this work, we present ART•V, a framework that generates video frames auto-regressively. As shown in Fig. 2, it first obtains a key frame as initialization. Then, with the key frame, or multiple copies of it, depending on the length of the conditioning sequence, ART•V generates subsequent frames auto-regressively, one frame at a time. The conditioning frames, typically one or two previous frames, are concatenated and injected into a pre-trained image diffusion model [44] using T2I-Adapter [33] (similar as ControlNet [73] but smaller), for conditional generation. The resulting model is more efficient compared to previous methods, as it only needs to learn simple continuous motions between adjacent frames. Besides, it minimizes alternations to the pre-trained image diffusion model, eliminating the necessity for additional temporal layers, and preserving its high-fidelity generation capability. Contrary to conventional wisdom, our auto-regressive model matches the inference speed of one-shot video models, while facilitating larger batch sizes during training.

To combat drifting in AR models, we propose masked diffusion, which learns a mask that determines which information can be directly drawn from reference images, rather than from network predictions, to reduce the chance of generating inconsistent appearance. The static noise, obtained by subtracting the reference image from the input noised image, is a short-cut to propagate reference images to the diffusion model. Therefore, the network only needs to predict the remaining part of the noise, which we call as dynamic noise. Fig. 3 shows an overview of the proposed masked diffusion. Moreover, we further enhance the generation process by conditioning it on the initial frame, which sets the tone for the overall scene and appearance details, further promoting global coherence. We call the above scheme anchored conditioning, benefiting long video generation as well. Finally, we perform noise augmentation to the reference frames to bridge the gap between training and testing. We combine the above techniques to arrive at ART•V, which effectively mitigates the drifting issue.

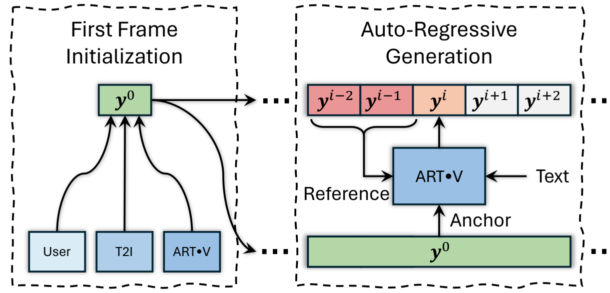

Figure 2. Overview of our video generation system ART•V, consisting of first frame initialization process and auto-regressive generation process. The first frame can be initialized by users, T2I models [1, 3, 44] or our ART•V itself.

We train our model on five million text-video pairs filtered from the WebVid-10M dataset [7]. Due to limited GPU resouces, we only train the model for two weeks on four A100 GPUs. However, we find that ART•V can already generate videos with natural motions, rich details and a high level of aesthetic quality. Though trained on low-resolution data, ART•V can directly generate impressive high-resolution videos, as shown in Fig. 6. It also achieves better quantitative results than the previous methods (they only represent proof-of-concept results since the methods are not fairly comparable due to differences in model size, training data and GPU resources). Fig. 1 shows some examples. Most importantly, the simplicity of our model makes it highly scalable to larger training data and longer training time, which we believe can further improve the results. Besides, ART•V enables various appealing applications. For example, it can generate long videos from multiple text prompts for story telling. It can also animate single images based on descriptive texts.

2. Related Work

Text-to-Video Generation. The problem has seen remarkable progress recently. Early T2V models demonstrated the possibility of generating videos in simple close-set domain [17, 28, 29, 31, 32, 35] and further exploited Transformer-based model [55] to achieve open-domain generation [23, 56, 64, 65]. Recently, diffusion-based T2V systems [6, 9–16, 18, 21, 22, 26, 30, 39, 42, 49, 58, 59, 61, 62, 66–68, 72, 74–76] have shown groundbreaking progress. Models like ModelScope [58] and Imagen Video [21] trained T2V models from scratch, demanding a huge text-video dataset and numerous GPU resources which is prohibitive for most cases. In contrast, most works [9, 26, 30, 49, 59, 62, 66, 68, 72] leveraged T2I model priors such as Stable Diffusion [44] for T2V by freezing or finetuning the pre-trained weights, showcasing compelling results. However, these methods, usually generating entire videos in one-shot, suffer from generating unrealistic large motions or very limited motions. In this work, we propose ART•V, a generation

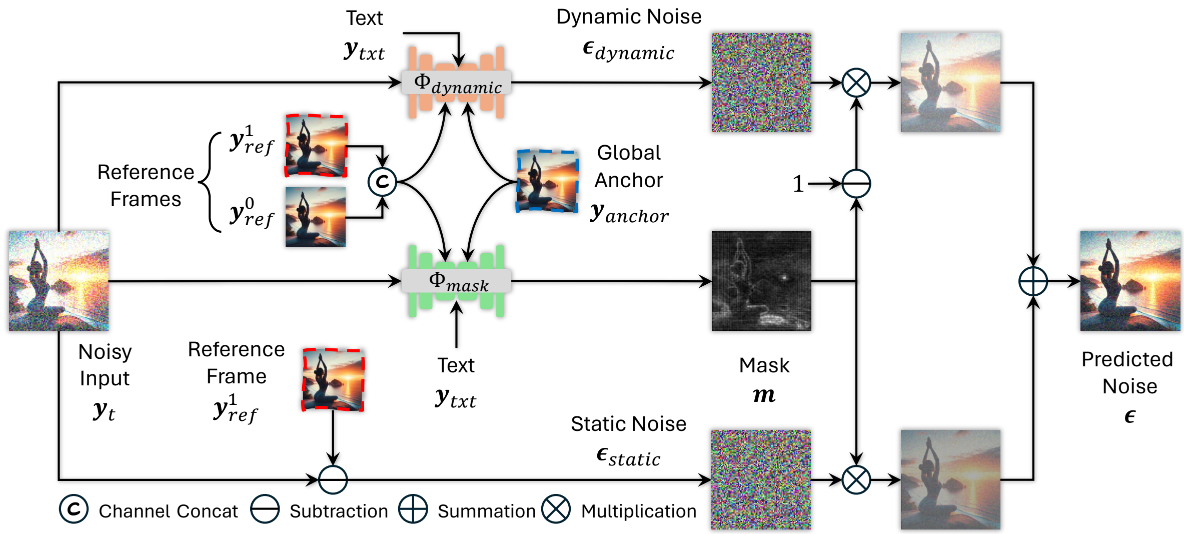

Figure 3. Illustration of the proposed masked diffusion model (MDM), conditioned on text, two reference frames and a global anchor frame. The predicted noise of MDM is composed of dynamic noise and static noise, which are scaled by a predicted mask. We employ two sub-networks $\Phi_{dynamic}$ and Φ_{mask} to predict dynamic noise and mask, respectively. Static noise is directly derived by subtraction of noisy input and reference frame. We initialize $\Phi_{dynamic}$ with Stable Diffusion 2.1 [44], while Φ_{mask} is randomly initialized. Reference frames and global anchor frame are injected into two sub-networks by using T2I-Adapter [33] and cross attention [44], respectively. Notably, the diffusion process is conducted in the latent space as in [44]. The autoencoder is omitted here for brevity.

system that avoids the challenge of learning complex long-range motion via auto-regressive first-order motion prediction, facilitating efficient training.

Auto-Regressive Video Generation. This is a burgeoning research area that aims to generate realistic and coherent videos by predicting each frame based on previously generated frames. Generally, three strategies have been employed. The first is pixel-level auto-regression. Some representative methods attempt to estimate the joint distribution of pixel value auto-regressively [25], speed up the processing by realizing a parallelized PixelCNN [43], and scale the techniques of auto-regressive Transformer architectures [55] to accommodate modern hardware accelerators [63]. The second is frame-level auto-regression. Huang *et al.* [24] proposed auto-regressive GAN to predict frames based on a single still frame. By overcoming error accumulation problem of AR, the complementary masking is introduced to promote the generation quality. The third is latent-level auto-regression, which significantly saves processing time due to reduced data redundancy and achieves a good time-quality trade-off [41, 48, 57, 71]. Our ART•V generation system, belonging to latent-level auto-regression, is the first attempt exploiting auto-regressive framework in the context of T2V with diffusion models.

3. Method

3.1. System Overview

Fig. 2 shows an overview of ART•V. Given a text prompt $\boldsymbol{y}_{txt}$ and an optional reference frame $\boldsymbol{y}^0$, it generates a video $\mathbf{V} = \{\boldsymbol{y}^0, \boldsymbol{y}^1, ..., \boldsymbol{y}^i, ...\boldsymbol{y}^N\}$. If $\boldsymbol{y}^0$ is not available, the system can use existing T2I models to generate one, or uses ART•V itself to generate one conditioned on blank images.

It trains a conditional diffusion model $\Phi(\cdot;\theta)$ parameterized by θ to perform auto-regressive generation, which is formulated as

$$\boldsymbol{y}^i = \Phi\left(\boldsymbol{y}_{txt}, \mathcal{R}^i; \theta\right), \tag{1}$$

where $\mathcal{R}^i$ denotes the set of conditional frames for generating $\boldsymbol{y}^i$. In implementation, $\mathcal{R}^i$ includes the previous two frames and an global anchor frame, denoted as $\boldsymbol{y}^{i-1}_{ref}$, $\boldsymbol{y}^{i-2}_{ref}$ and $\boldsymbol{y}_{anchor}$, to encode first-order motions.

Our model is built on Stable Diffusion 2.1 [44] (SD2.1). To support image conditional generation, the two reference frames are concatenated along the channel dimension and injected into SD2.1 in a T2I-Adapter [33] style, while the global anchor frame adopts cross attention for injection. We do not introduce additional temporal modeling modules such as 3D convolutions and attention layers, which are required by previous T2V models. This is because we only need to model short motions between adjacent frames. In the following, we will elaborate our proposed techniques for alleviating the drifting issue in AR models.

3.2. Masked Diffusion Model (MDM)

In standard diffusion process, all pixels are predicted from random noises by networks which have large chance of generating appearances inconsistent with the previous frames. As prediction proceeds auto-regressively, the accumulated errors will eventually lead to drifting. The core idea of

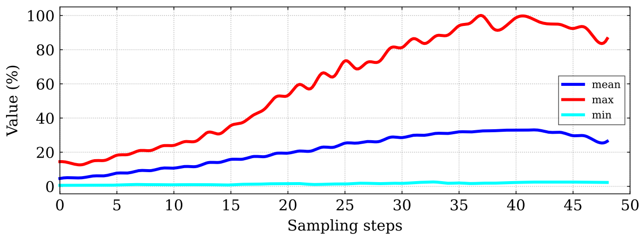

Figure 4. Value distribution of the estimated mask by mask diffusion model during different sampling steps. The maximum sampling step is 50.

MDM is to implicitly learn a mask determining which information can be drawn directly from closely related conditional images rather than network predictions to reduce inconsistency. Fig. 3 shows an overview of MDM.

As shown in Fig. 3, MDM has two U-Nets for predicting noise and mask, respectively. The static noise, directly obtained by subtracting the reference image from the input noised image, is a short-cut to propagate information in the reference image to the diffusion process. We find that the model tends to copy more from reference images at later denoising steps, which effectively reduces the risk of generating inconsistent high-frequency appearances that cause drifting. This is illustrated in Fig. 4. The U-Net hence only needs to predict the remaining part of the noise, which we call as dynamic noise. In the following, we will formally introduce the method.

Diffusion Model Preliminaries. Diffusion model has a forward and a backward process, respectively. The forward process gradually adds noises to the clean data $\boldsymbol{y}_0 \sim q(\boldsymbol{y}_0)$, which can be formulated as:

$$q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}) = \mathcal{N}\left(\boldsymbol{y}_t; \sqrt{1-\beta_t}\boldsymbol{y}_{t-1}, \beta_t \mathbf{I}\right), \tag{2}$$

where $t \in \{1, ..., T\}$ and $\beta_t \in (0,1)$ is a fixed variance schedule. Denote that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, we can directly sample $\boldsymbol{y}_t$ in a closed form from the distribution $q(\boldsymbol{y}_t|\boldsymbol{y}_0)$ at an arbitrary timestep t:

$$\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The backward process reverses the forward process, which eventually maps Gaussian noises $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the target data. Specifically, the backward denoising process solves the posterior $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t)$, which can be approximated by training a deep neural network $\Phi(\cdot;\theta)$ to predict the noise $\boldsymbol{\epsilon}$ added to the data. The training objective is formulated as:

$$\mathbb{E}_{\boldsymbol{y}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t}\left[\left\|\boldsymbol{\epsilon} - \Phi\left(\boldsymbol{y}_t, \boldsymbol{c}, t; \theta\right)\right\|_2^2\right], \tag{4}$$

where $\mathbf{c}$ denotes the conditions that represent the reference and global anchor frames, and texts in our ART•V system.

Mask Prediction and Dynamic Noise. In MDM, noise prediction in Eq. (4) is realized by two networks: dynamic noise prediction network $\Phi_{dynamic}(\cdot;\theta_0)$ and mask prediction network $\Phi_{mask}(\cdot;\theta_1)$. Without loss of generality, we define $\sigma = \sqrt{\bar{\alpha}_t}$ and $\lambda = \sqrt{1-\bar{\alpha}_t}$. We omit t for brevity. We reformulate Eq. (3) as:

$$\begin{aligned} \boldsymbol{y}_t &= \sigma\boldsymbol{y}_0 + \lambda\boldsymbol{\epsilon} \\ &= (\boldsymbol{y}_{ref} + \boldsymbol{y}_{res}) + \lambda\boldsymbol{\epsilon} \\ &= \boldsymbol{y}_{ref} + (\boldsymbol{y}_{res} + \lambda\boldsymbol{\epsilon}) \\ &= \boldsymbol{y}_{ref} + \boldsymbol{\epsilon}', \end{aligned} \tag{5}$$

where $\boldsymbol{y}_{ref}$ is the reference frame, and $\boldsymbol{y}_{res}$ denotes the residual component between $\sigma\boldsymbol{y}_0$ and $\boldsymbol{y}_{ref}$. Therefore, the $\boldsymbol{\epsilon}$, which needs to be predicted by the diffusion model $\Phi(\cdot;\theta)$ in Eq. (4), can be derived from Eq. (5):

$$\begin{aligned} \boldsymbol{\epsilon} &= \frac{\boldsymbol{y}_{ref} + \boldsymbol{\epsilon}' - \sigma\boldsymbol{y}_0}{\lambda} \\ &= \frac{\boldsymbol{y}_{ref} - \sigma\boldsymbol{y}_0}{\lambda} + \frac{\boldsymbol{\epsilon}'}{\lambda} \\ &= \frac{\boldsymbol{\epsilon}''}{\lambda} + \frac{\boldsymbol{\epsilon}'}{\lambda} \\ &= \boldsymbol{\epsilon}_{static} + \boldsymbol{\epsilon}_{dynamic}, \end{aligned} \tag{6}$$

where $\boldsymbol{\epsilon}_{static}$ and $\boldsymbol{\epsilon}_{dynamic}$ represents the static noise and dynamic noise, respectively.

We can see from Eq. (5) and Eq. (6) that the static noise $\boldsymbol{\epsilon}_{static}$ is from the reference image $\boldsymbol{y}_{ref}$, which can be directly propagated to the output and is expected to mitigate error accumulation. In our implementation, we make approximation $\boldsymbol{\epsilon}_{static} \simeq \boldsymbol{y}_{ref} - \boldsymbol{y}_t$. In such a way, $\boldsymbol{\epsilon}_{static}$ can be directly derived from reference images and noised input input, which do not need to be predicted. The dynamic noise $\boldsymbol{\epsilon}_{dynamic}$ contains the residual component $\boldsymbol{y}_{res}$ that changes dynamically, which needs to be predicted by our noise prediction network $\Phi_{dynamic}(\cdot;\theta_0)$. In order to determine the contributions of static and dynamic noises, we employ the mask prediction network $\Phi_{mask}(\cdot;\theta_1)$ to predict a mask $\boldsymbol{m}$. Eventually, the final predicted noise of our mask diffusion model is obtained by:

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{m} \cdot \boldsymbol{\epsilon}_{static} + (1-\boldsymbol{m}) \cdot \boldsymbol{\epsilon}_{dynamic}. \tag{7}$$

The two networks can be optimized by Eq. (4).

3.3. Noise Augmentation

Drifting issue in our ART•V generation system arises not only from prediction error but also from train-test discrepancy. During training, the model utilizes ground truth frames as references and the global anchor. However, during testing, it conditions on generated frames prone to noises. Inspired by [44], we slightly corrupt reference and

Table 1. Quantitative comparisons with SoTA for zero-shot video generation on UCF-101 [53] and MSR-VTT [69].

Methods	Training Data	UCF-101[53]			MSR-VTT[69]		
		Zero-shot	FVD ↓	IS ↑	Zero-shot	FVD ↓	CLIPSIM ↑
GODIVA [64]	MSR-VTT [69]	Yes	-	-	No	-	0.2402
NUWA [65]	MSR-VTT [69]	Yes	-	-	No	-	0.2439
Make-A-Video [49]	WebVid-10M [7] + HD-VILA-100M [70]	Yes	367.23	33.00	Yes	-	0.3049
VideoFactory [59]	WebVid-10M [7] + HD-VG-130M [59]	Yes	410.00	-	Yes	-	0.3005
ModelScope [58]	WebVid-10M [7] + LAION-5B [46]	Yes	410.00	-	Yes	550.00	0.2930
VideoGen [26]	WebVid-10M [7] + Private-HQ-2K [26]	Yes	554.00	71.61	Yes	-	0.3127
Lavie [62]	WebVid-10M [7] + LAION-5B [46]	Yes	526.30	-	Yes	-	0.2949
VidRD [16]	WebVid-2M [7] + TGIF [27] + VATEX [60] + Pexels [5]	Yes	363.19	39.37	Yes	-	-
PYoCo [15]	Private-data [15]	Yes	355.19	47.76	Yes	-	0.3204
LVDM [18]	WebVid-2M [7]	Yes	641.80	-	Yes	742.00	0.2381
CogVideo [23]	WebVid-5.4M [7]	Yes	702.00	25.27	Yes	1294.00	0.2631
MagicVideo [76]	WebVid-10M [7]	Yes	699.00	-	Yes	998.00	-
Video-ldm [9]	WebVid-10M [7]	Yes	550.61	33.45	Yes	-	0.2929
VideoComposer [61]	WebVid-10M [7]	Yes	-	-	Yes	580.00	0.2932
VideoFusion [30]	WebVid-10M [7]	Yes	639.90	17.49	Yes	581.00	0.2795
SimDA [68]	WebVid-10M [7]	Yes	-	-	Yes	456.00	0.2945
ART•V + W/O Image (Ours)	WebVid-5M [7]	Yes	567.20	26.89	Yes	356.50	0.2897
ART•V + SDXL [38] (Ours)	WebVid-5M [7]	Yes	539.57	36.21	Yes	413.01	**0.3022**
ART•V + GT Image (Ours)	WebVid-5M [7]	Yes	**315.69**	**50.34**	Yes	**291.08**	0.2859

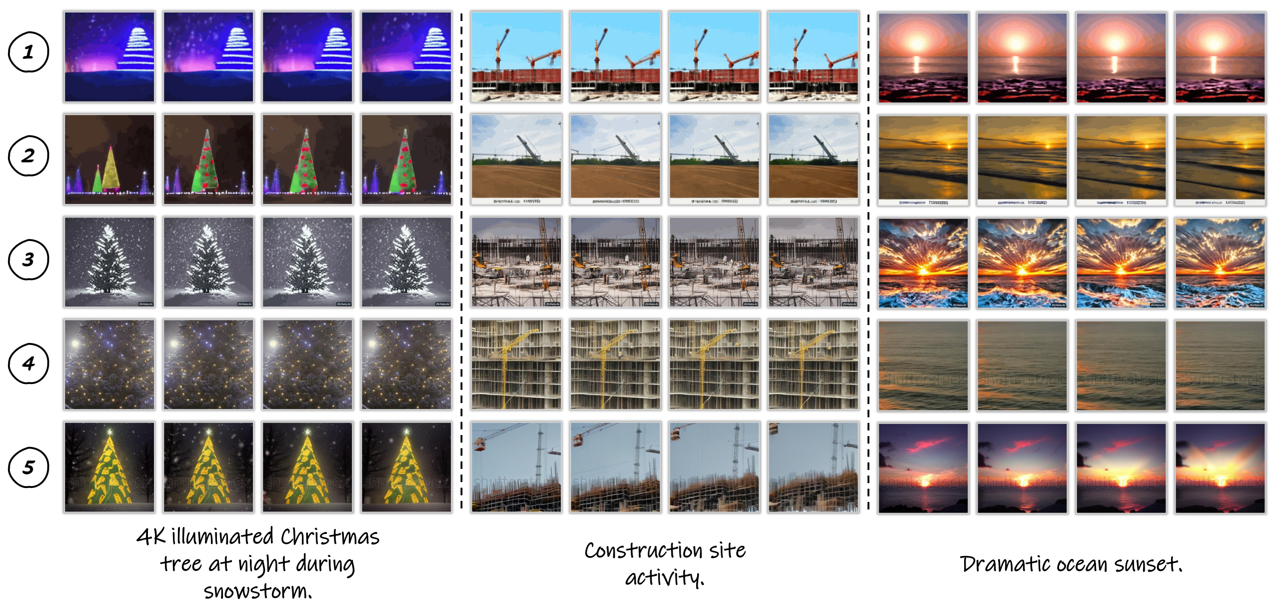

Figure 5. Visual comparisons of text-to-video generation. The results of row 1 to row 5 are sampled from VDM [22], CogVideo [23], Make-A-Video [49], ModelScope [58] and Our ART•V.

global anchor frames using the forward diffusion process in Eq. (3). In particular, for each training step, we randomly sample a noise level $t \in [0, T_{max}]$. In such a way, the model has the chance to see clean reference frames and corrupted ones, respecting the case of inference and expected to address the error accumulation problem during inference. Following [44], we also use the noise level t_{test} as an additional condition by adding it to the time step embedding of diffusion model. In inference, we use a fixed noise level of $t_{\text{test}} = 200$, validated by the ablation study in Sec. 4.2.

3.4. Anchored Conditioning

In addition to using masked diffusion model and noise augmentation to address drifting issue in our ART•V generation system, we introduce a novel design, anchored conditioning, expected to promote model capacity for long video generation. One key challenge in generating long videos is to maintain consistency in terms of scenes and objects throughout videos, solved by a global anchor in ART•V.

In detail, we use the first frame, which is free from noises, as a stable anchor frame $\boldsymbol{y}_{anchor}$ to preserve the content, in whole videos. In training, we randomly select one frame within a fixed time window range preceding the current one to serve as the global anchor frame. We expirically choose time window range as 10, to create relatively large motion variations. We use cross attention [55] to inject the global anchor frame to the diffusion model. The strategy addresses the inherent challenges in long text-to-video generation, providing a robust mechanism for faithfully retaining the scenes and objects.

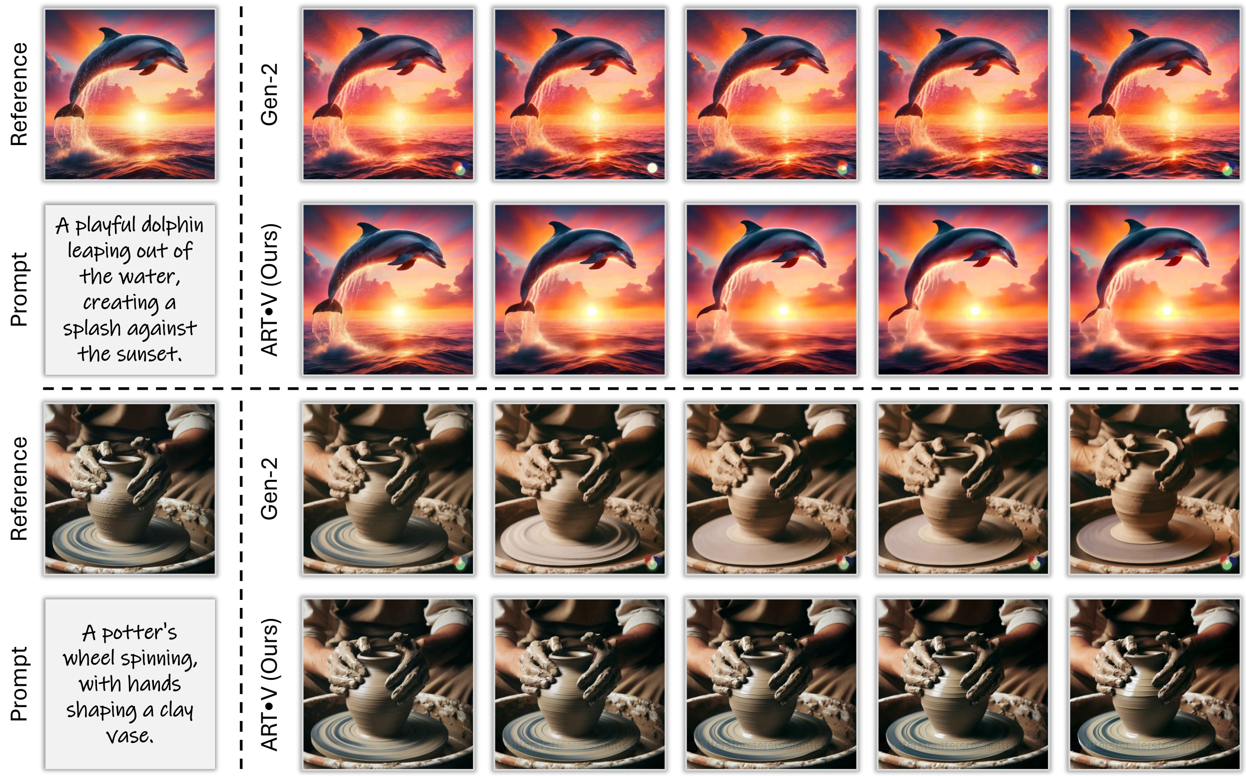

Figure 6. Visual comparisons of text-image-to-video generation. Reference image generated by DALL-E 3 [1]. Notebly, ART•V is trained on 320×320 video data, while the inference is performed on 768×768 in these cases.

4. Experiment

Datasets and Evaluation Metrics. To make quantitative and qualitative comparisons, we choose the publicly available datasets: WebVid-10M [7], MSR-VTT [69] and UCF-101 [53]. We split WebVid-10M to training subset and testing subset. We make data cleaning on the training subset. In specific, we use the public code [4] to compute the motion score of each video and then only retain the videos whose motion scores are between $[1, 20]$. Subsequently, we compute a CLIP score for each video and retain the top 5 million data that have largest CLIP scores [40]. We train our model on this cleaned 5M dataset. MSR-VTT [69] and UCF-101 [53] are utilized for evaluation. We report the Frechet Video Distance (FVD) [54], Frechet Inception Distance (FID) [36], Inception Score (IS) [45] and CLIPSIM (average CLIP similarity between video frames and text) [40] for quantitative comparison.

Implementation Details. We implement our method using Pytorch [37] and use AdamW solver for optimization. We train our diffusion model with 1000 noising steps and a linear noise schedule. The exponential moving average (EMA) of model weights with 0.9999 decay is adopted during training. We set the learning rate as $1e^{-5}$ and keep it constant during the training process. We use a batch size of 640. For noise augmentation, we set the maximum noise level T_{max} as 550. For inference, we employ classifier-free guidance [19] to amplify the effect of the conditional signals of reference frames $\boldsymbol{y}_{ref}$, global anchor frame $\boldsymbol{y}_{anchor}$ and text prompts $\boldsymbol{y}_{text}$. The guidance scales of $\boldsymbol{y}_{ref}$, $\boldsymbol{y}_{anchor}$ and $\boldsymbol{y}_{text}$ are set as 0.25, 0.25 and 6.5, respectively. During training, we randomly drop these conditions with a drop rate of 10%.

4.1. Application

We now demonstrate a wide range of applications of our ART•V system. Our ART•V, only trained once without task-specific finetuning, can skillfully support multiple generation tasks. In contrast, existing models like VideoCrafter1 [10] needs to train two individual models for T2V and TI2V, causing large training cost.

Text-to-Video Generation. We first exploit our ART•V to perform text-to-video generation, without the image provided by T2I models [1, 3, 44] or users. It is worth noting that, our model is trained using joint conditions of text and images. Notably, we randomly drop the image condition with a drop rate of 10% during training. It suggests that the training cases of text-to-video take a small proportion. However, we observe that, ART•V, trained for text-image-to-video generation, is able to skillfully generate video by using text condition only. Specifically, when there is no

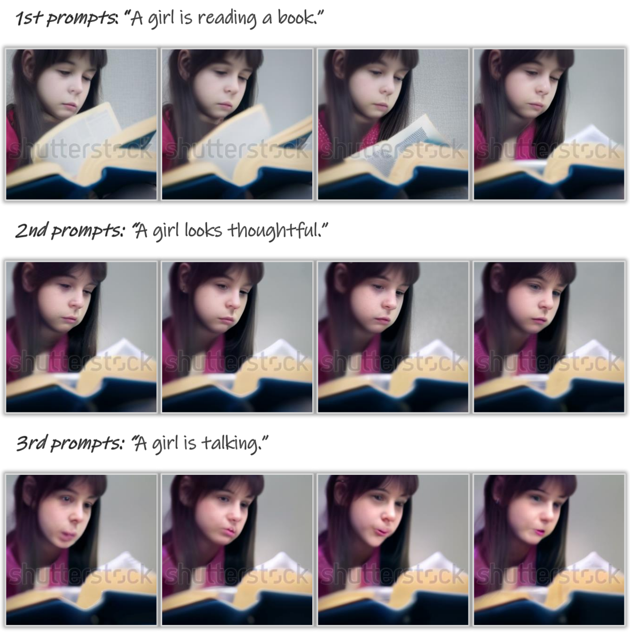

Figure 7. Visual result of multi-prompt long text-to-video generation. 16 frames are generated for each prompt.

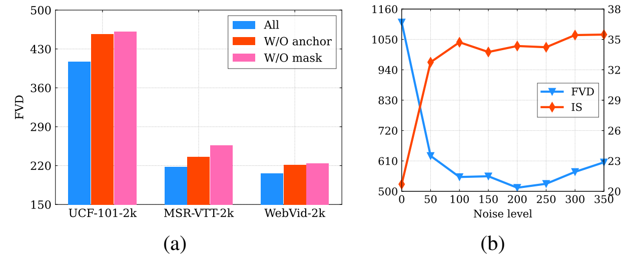

Figure 8. (a) Ablation results of mask diffusion model and anchor conditioning on UCF-101-2k [53], MSR-VTT-2k [69] and WebVid-2k [7]. (b) Investigation results of noise augmentation on UCF-101-2k [53].

provided reference frames, we directly use our model to generate one from the text prompt, leaving the conditioned reference frames blank. Then, we generate the subsequent frames conditioned on the generated reference frames. We demonstrate the quantitative results in Tab. 1. We compare our method with the existing state-of-the-art methods on UCF-101 [53] and MSR-VTT [69] in a zero-shot setting. It can be clearly observed that, our method ART•V, achieving FVD score of 567.20 and IS score of 26.89 in UCF-101, consistently outperforms existing methods such as VideoFusion [30], MagicVideo [76], LVDM [18] and CogVideo [23]. In MSR-VTT, we keep the top performance in terms of FVD, and even outperform ModelScope [58] that utilizes additional high-quality datasets for training. In Fig. 5, we also demonstrate some exemplary results of different methods using the same text prompts. The visual results also support the conclusions above, demonstrating the visually-satisfying results compared to the existing methods. In addition, we believe if ART•V is finetuned for T2V task, we will achieve better results.

Text-Image Conditioned Video Generation. ART•V also offers the ability to animate a still image based on text prompts. We either employ the existing T2I models such as Stable Diffusion [44], Midjourney [3], and DALL-E 3 [1] to generate reference images or directly use the images provided by users. The numbers are reported in Tab. 1. We make two variants of our method, termed "ART•V+SDXL" and "ART•V+GT Image", which utilize the image generated by SDXL [38] and GT image as the first frame, respectively. As can be clearly observed, when conditioned on an additional image, our ART•V achieves better results in terms of FVD and IS in UCF-101. Especially, we achieve the SoTA results, FVD of 315.69 and IS of 50.34 in UCF-101, FVD of 291.08 in MSR-VTT, when GT image is taken as reference image. It demonstrates the superior performance of text-image conditioned generation of our ART•V.

We demonstrate some visual exemplar videos generated by our method in Fig. 6. In these cases, ART•V is exploited to generate high-resolution videos of 768 × 768, though the model is trained on 320 × 320. We compare to a well-known video generation system Gen-2 [2] provided by a commercial company. We generate the reference frame using DALLE 3 [1], which is then fed to ART•V and Gen-2 to generate videos, respectively.

We observe that both ART•V and Gen-2 are able to animate the given image using the text description, demonstrating good visual fidelity. Notably, our ART•V exhibits a superior ability to preserve appearance compared to Gen-2. As can be seen from the second case of Fig. 6, Gen-2 shows the severe color shifting problem, while our ART•V preserves the content in the reference images, thanks to the proposed masked diffusion model and anchored conditioning. In addition, the exceptional visual quality of ART•V demonstrates that our method can achieve tuning-free, high-resolution video generation, thereby significantly reducing the training costs. Nevertheless, Gen-2 show superior results in terms of visual detail and temporal consistency due to additional high-quality training data and temporal interpolation models, which is beyond the scope of this paper. In contrast, we train ART•V only using WebVid-5M, which has low resolution and quality.

Multi-Prompt Long Video Generation. ART•V is suitable for long video generation due to its auto-regressive nature. We can repeat the auto-regressive process to generate an arbitrarily long video, and require different segments of the video to be conditioned on different prompts. The leading frame of a video segment should be conditioned on the ending frames of last video segment to promote coherence and continuity. Fig. 7 shows an example. We can see that our system can generate videos with coherent scenes and objects, and meanwhile the motions in each segment are faithful to the corresponding prompts.

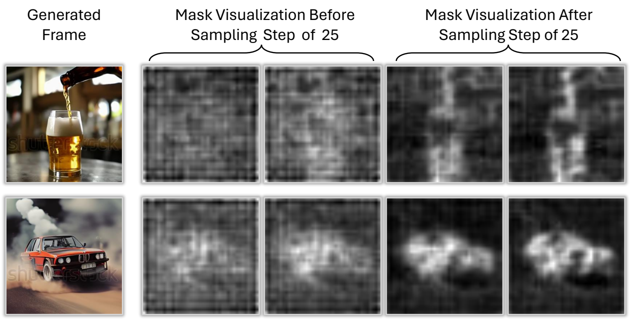

Figure 9. Visualization of estimated mask by mask diffusion model. Reference image is generated by SDXL [38].

4.2. Ablation Study

Masked Diffusion Model. We propose masked diffusion model to alleviate the error accumulation in our ART•V system. To validate its effectiveness, we introduce a baseline which drops the mask prediction network. It trains a single network to predict the noise in Eq. (7) as in standard diffusion models. As shown in Fig. 8 (a), when we drop the masked diffusion, the performance drops significantly on all evaluation datasets. We also visually compare the videos generated by different methods in Fig. 10. We can see that the model suffers from severe drifting without masked diffusion. In addition, the image quality is also degraded, losing many sharp details. These results demonstrate the importance of masked diffusion.

We show the normalized strengths of the predicted masks at different time steps in Fig. 4. The average strength increases as the denoising step, suggesting that the diffusion model will use copy more from the reference images in later denoising steps, where diffusion models focus on high-frequency appearance generation [8]. So, our model will generate images that have similar appearance as the reference images, thus can effectively alleviate the drifting issue. Fig. 9 shows the masks at different denoising steps, which validates our conjecture.

Noise Augmentation. In addition to masked diffusion model, we propose noise augmentation to further reduce error accumulation. We investigate the effect of applying different levels of noises, *i.e.* t_{test}, during inference. The numeric results are in Fig. 8 (b). As can be observed, adopting noise augmentation brings significant performance boosts in terms of FVD and IS metrics. When we increase the noise level, IS achieves consistently better results but the gains become marginal after exceeding 100. In contrast, the FVD gets the best result when the noise level is 200 and shows performance drop when noise level exceeds 200. It is worth noting the value of 200 is approximately the average of the noise levels we applied during training. Fig. 10 shows the visual results of ablating noise augmentation, which is adversely affected by the noise artifacts and reveals the necessity of noise augmentation.

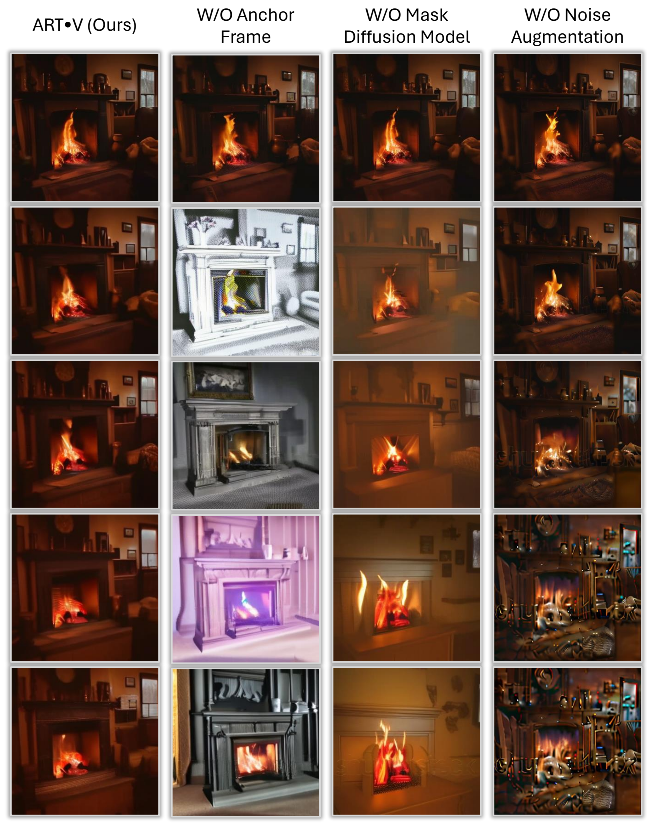

Figure 10. Visual results of ablation study. Reference image is generated by SDXL [38] by using prompt "*interior, fireplace.*"

Anchored Conditioning. Here we validate the effectiveness of anchor frame. We manually set the anchor frame to be zero and keep the model structure unchanged. As can be seen from Fig. 8 (a), without the anchor frame as an additional condition, the model shows a clear performance drop in terms of FVD for all evaluation datasets. The visual results of Fig. 10 showcase the obvious domain shifting problem with loss of high frequency details when removing anchor frame. These results indicate the anchored conditioning is an essential to retain the overall appearance.

5. Conclusion

This paper realizes a novel text-to-video generation system, termed ART•V, to generate videos conditioned on texts or images in an auto-regressive frame generation manner. To address the error accumulation problem and support long video generation, we implement our ART•V generation system by proposing mask diffusion model that carefully utilizes the priors of reference images, noise augmentation that closes the train-test discrepancy and anchored conditioning that assures scene consistency. As validated by comprehensive experiments, we demonstrates superior performance over various comparison methods.

References

[1] https://openai.com/dall-e-3. 1, 2, 6, 7

[2] https://research.runwayml.com/gen2. 7

[3] https://www.midjourney.com/home. 1, 2, 6, 7

[4] https://github.com/Breakthrough/PySceneDetect. 6

[5] https://huggingface.co/datasets/Corran/pexelvideos. 5

[6] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 5, 6, 7

[8] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 8

[9] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. 2, 5

[10] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 6

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

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

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

[14] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023.

[15] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 5

[16] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 2, 5

[17] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, pages 598–613, 2018. 2

[18] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 5, 7

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeuIPS*, 33:6840–6851, 2020. 1

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2, 5

[23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 5, 7

[24] Jiahui Huang, Yew Ken Chia, Samson Yu, Kevin Yee, Dennis Küster, Eva G Krumhuber, Dorien Herremans, and Gemma Roig. Single image video prediction with auto-regressive gans. *Sensors*, 22(9):3533, 2022. 3

[25] Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *ICML*, pages 1771–1779. PMLR, 2017. 3

[26] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2, 5

[27] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5

[28] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018. 2

[29] Yue Liu, Xin Wang, Yitian Yuan, and Wenwu Zhu. Cross-modal dual learning for sentence-to-video generation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1239–1247, 2019. 2

[30] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, pages 10209–10218, 2023. 2, 5, 7

[31] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *ICCV*, pages 1426–1434, 2017. 2

[32] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1096–1104, 2017. 2

[33] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3

[34] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021. 1

[35] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 2

[36] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 6

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5, 7, 8

[39] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[41] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020. 3

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

[43] Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando Freitas. Parallel multiscale autoregressive density estimation. In *ICML*, pages 2912–2921. PMLR, 2017. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 6, 7

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[46] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 2

[48] Younggyo Seo, Kimin Lee, Fangchen Liu, Stephen James, and Pieter Abbeel. Harp: Autoregressive latent video prediction with high-fidelity image generator. In *ICIP*, pages 3943–3947. IEEE, 2022. 3

[49] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 5

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 1

[51] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeuIPS*, 32, 2019.

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2020. 1

[53] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 6, 7

[54] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeuIPS*, 30, 2017. 2, 3, 5

[56] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[57] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021. 3

[58] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 5, 7

[59] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 2, 5

[60] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 5

[61] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 5

[62] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 5

[63] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 3

[64] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2, 5

[65] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, pages 720–736. Springer, 2022. 2, 5

[66] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7623–7633, 2023. 2

[67] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023.

[68] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 2, 5

[69] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5, 6, 7

[70] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 5

[71] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[72] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[74] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 2

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

[76] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 5, 7