Grid Diffusion Models for Text-to-Video Generation

Taegyeong Lee* Soyeong Kwon* Taehwan Kim
Artificial Intelligence Graduate School, UNIST
{taegyeonglee, soyoung17, taehwankim}@unist.ac.kr

Abstract

Recent advances in the diffusion models have significantly improved text-to-image generation. However, generating videos from text is a more challenging task than generating images from text, due to the much larger dataset and higher computational cost required. Most existing video generation methods use either a 3D U-Net architecture that considers the temporal dimension or autoregressive generation. These methods require large datasets and are limited in terms of computational costs compared to text-to-image generation. To tackle these challenges, we propose a simple but effective novel grid diffusion for text-to-video generation without temporal dimension in architecture and a large text-video paired dataset. We can generate a high-quality video using a fixed amount of GPU memory regardless of the number of frames by representing the video as a grid image. Additionally, since our method reduces the dimensions of the video to the dimensions of the image, various image-based methods can be applied to videos, such as text-guided video manipulation from image manipulation. Our proposed method outperforms the existing methods in both quantitative and qualitative evaluations, demonstrating the suitability of our model for real-world video generation.

1. Introduction

The advancement of diffusion models has resulted in significant improvements in the performance of text-to-image models [9, 18, 23, 24, 26]. Unlike GAN-based models, the diffusion model is easier to train, offering desirable properties such as distribution coverage, a stationary training objective, and easy scalability [8]. Based on these strengths, various studies [5, 16, 45, 47] are being conducted to manipulate or generate images from text using diffusion, and research on generating videos [4, 6, 12, 15, 17, 29] from text is also actively being pursued. However, video generation is more challenging than image generation because videos have higher dimensions [17], there is a scarcity of text-video datasets [11, 15], and it incurs higher costs [15, 17, 34] than generating an image from text. Previous studies [4, 15, 17, 29] generate a video by using additional temporal dimensions and super-resolution models to maintain the temporal consistency and resolution of videos. This characteristic of videos makes efficiency an important issue in video generation, which is one reason why many video generation studies [12, 17, 46] focus on efficiency. Unlike the existing video generation paradigm, we propose novel grid diffusion models that reduce the high dimensionality of videos to that of images, allowing for high-quality video generation without substantial GPU memory costs and a large paired dataset. We leverage the strengths of diffusion actively to generate videos from text.

Our model consists of two stages: 1) key grid image generation and 2) autoregressive grid image interpolation. To reduce video generation to image generation, we select four frames from the video in chronological order and generate an image as in Figure 2, called a *key grid image*. The image consists of four inside frames that represent the video generated from the text. We fine-tune a pre-trained text-to-image model [26] using the prompt as the condition to generate the key grid image. According to Stable Diffusion [26], due to VAE latent in representing global spatial image structure, we can prevent the naive generation of the four similar inside frames and generate individual inside frames in the key grid image with temporal consistency.

However, unlike prior text-to-video generation models [13–15, 17, 29, 34] that generate only 16 frames, our key grid image consists of four inside frames. Therefore, we need to interpolate inside frames of the key grid image while maintaining temporal consistency and order. Since we reduce the video to an image dimension, we can use an image manipulation method [5]. Inspired by [5], we propose an autoregressive grid image interpolation method. Our interpolation model takes the masked grid image as the input and the previously generated key grid image as the condition. Our model concatenates the embedding spaces of the two images in the latent dimension. This enables us to generate coherent video frames that are consistent within the current grid image and with the previously generated grid image.

[*] These authors contributed equally to this work.

Also, to generate more frames, we use the next key grid image generation model by autoregressively using the previous key grid image as a condition. This approach allows our model to maintain temporal consistency and generate videos with more than 28 frames.

Additionally, since we represent a video as a grid image, our model can be applied to various applications with image-based models, such as video manipulation from using image manipulation.

In our experiments, we achieve better performance than existing text-to-video models without large paired training datasets and can generate more frames with a fixed amount of GPU memory costs. These results indicate that our grid diffusion for text-to-video generation can be applied in the real world.

In summary, our contributions are as follows:

- We propose simple but effective novel grid diffusion models for efficient text-to-video generation by reducing the temporal dimension of video.
- We generate high-quality videos using a fixed amount of GPU memory regardless of the number of frames and less training data.
- Since our model represents video as grid image, one may easily apply image based models for corresponding video tasks such as video manipulation and video style editing.
- In experimental results, our model is able to generate faithful and high quality videos from text and outperforms baselines in both quantitative and qualitative evaluations.

2. Related Work

Text-to-Image Generation. Research on generating high-quality images from text has been long studied [25, 38, 42, 44], and recent advancements in diffusion models have enabled the generation of high-quality images from general text, leading to significant societal impact. Recent studies have utilized architectures such as Transformers, variational autoencoder (VAE), and diffusion models to generate higher-resolution and more general images from text descriptions. For instance, DALLE [23] and Parti [41] train Transformer models on large-scale text-image paired datasets to enable the generation of images from general text inputs. On the other hand, models like GLIDE [18], DALLE2 [24], and Stable Diffusion [26] utilize diffusion models to generate images. These diffusion-based models have shown promising results in image generation tasks. We propose an approach that leverages the strengths of diffusion models and uses Stable Diffusion [26] which has been pre-trained on large-scale text-image pair datasets, to generate high quality videos from text without the temporal dimension.

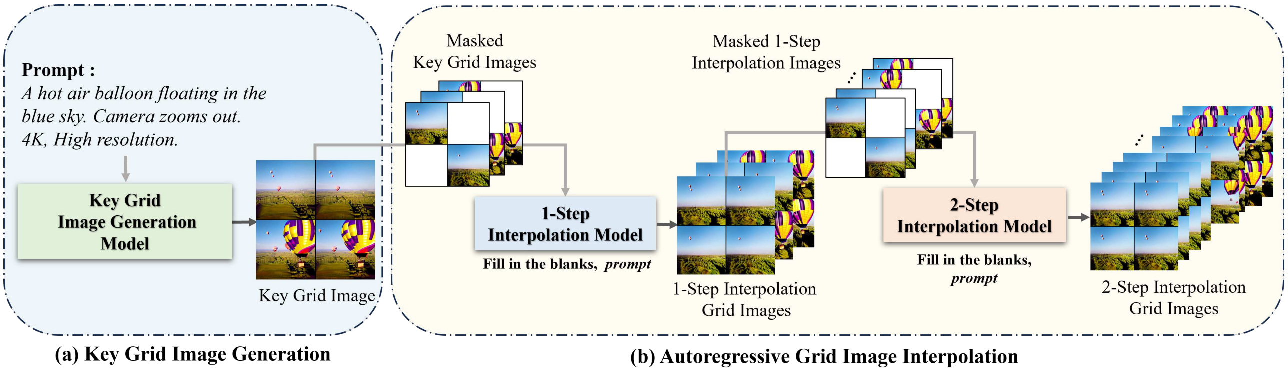
Text-to-Video Generation. Text-to-video generation is confronted with two major challenges: the lack of a large-scale high-quality text-video dataset and the complexity of modeling the temporal dimension [1, 29]. Make-A-Video [29] extends a diffusion-based text-to-image model, DALLE2 [24], to text-to-video by leveraging joint text-image priors and introducing super-resolution strategies for high-definition and high frame-rate video generation. Video diffusion models [14] trains image and video jointly with the 3D U-Net diffusion model architecture. Latent-Shift [12] generates video by shifting the spatial U-Net feature map forward and backward in the temporal dimension which enables to ensure temporal coherence in the video and efficiency. Also, the PYoCo [11] extends text-to-image diffusion models into a 3D dimension and fine-tune a pre-trained diffusion model. Additionally, it utilizes a noise prior and a pre-trained eDiff-I [3] model for generating videos. Despite active research being conducted in the field of text-to-video generation, there are still challenging issues due to the complex model structure and large training data required. We address these problems through a simple architecture with an effective approach, presenting a new paradigm of text-to-video generation without large text-video paired training datasets.

3. Method

We propose a simple but effective novel approach for text-to-video generation using a grid diffusion model. As shown in Figure 1, our model consists of two main stages: (i) *key grid image* generation and (ii) autoregressive grid image interpolation. In the first stage, we generate a key grid image that represents video from the given text. In the second stage, we interpolate the generated key grid image to generate the video. This approach enables us to generate high-quality videos with a fixed amount of GPU memory costs and less training data than existing text-to-video generation models [13–15, 17, 29, 34], and also allows for video manipulation in the image dimension.

3.1. Key Grid Image Generation

To generate videos by reducing the temporal dimension, we generate a key grid image. The key grid image consists of four inside frames representing the primary motions or events of the video. Figure 2 shows key grid image generation process and model training overview. In training, we first select four frames from the video in chronological order. Second, we arrange the frames in the selected order. The generated key grid image has a resolution of 512×512 and is composed of four inside frames, each with a resolution of 254×254. To train key grid image generation model, we fine-tune a Stable Diffusion [26] model pre-trained with LAION-5B [28] on 0.1 million samples of Webvid-10M [2]. As described in [26], VAE latent may capture global spatial structure in image, therefore we can encode temporal dynamics order with the same interval. Also we empirically


Prompt :
A hot air balloon floating in the blue sky. Camera zooms out. 4K, High resolution.
Key Grid Image Generation Model
Key Grid Image
Masked Key Grid Images
1-Step Interpolation Model
Fill in the blanks, prompt
1-Step Interpolation Grid Images
Masked 1-Step Interpolation Images
2-Step Interpolation Model
Fill in the blanks, prompt
2-Step Interpolation Grid Images

(a) Key Grid Image Generation

(b) Autoregressive Grid Image Interpolation

Figure 1. **Overview of our approach**. Our approach consists of two stages. In the first stage (a), our key grid image generation model generates a key grid image following the input prompt. In the second stage (b), our model generates masked grid images by applying masking between each of the four frames and performs a 1-step interpolation using 'Fill in the blanks,' as a prefix with the prompt. Then, our model conducts a 2-step interpolation with the 2-step interpolation model, using the masked grid image from the 1-step interpolation images as input.

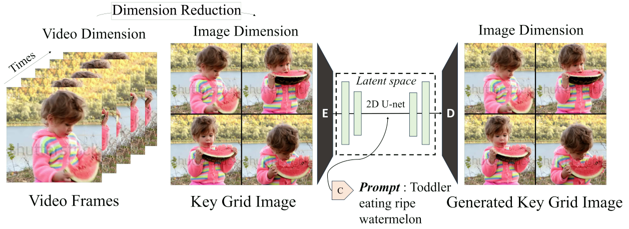


Figure 2. **Visualization of key grid image generation model training**. We train the key grid image generation model with 2D U-Net after representing the video as a key grid image, conditioned on the prompt.

find that the key grid image dataset adequately reflects motion, and therefore, our key grid image generation model effectively represents scene changes and dynamic motions.

3.2. Autoregressive Grid Image Interpolation

Since our key grid image is composed of a total of four inside frames, more frames are required to generate a video. In addition, it is desirable to have each frame connected to each other and keep temporal consistency between frames. Therefore, we propose and train interpolation models to generate the output grid image conditioned on the previously generated grid image and the masked input grid image in an autoregressive manner. As shown in Figure 1, our interpolation models are composed of two models (1-Step Interpolation and 2-Step Interpolation). We train interpolation models using grid images with different intervals for each model. Specifically, as shown in Figure 3, we use a grid image which is masked on the second and third frames as an input so that our model can interpolate and generate frames between the first and fourth frames in the grid image. Building upon Instructpix2pix [5] approach, we concatenate the embeddings of the input grid image and the conditioning previous grid image. We initialize the weights of the U-Net as pre-trained on the LAION-5B [28] and then we expand the input channels of the U-Net architecture. Since it corresponds to the image editing task that fills the masked image, we generate the filled grid image by using "Fill in the blanks" as a prefixed instruction and text prompt as condition for interpolation.

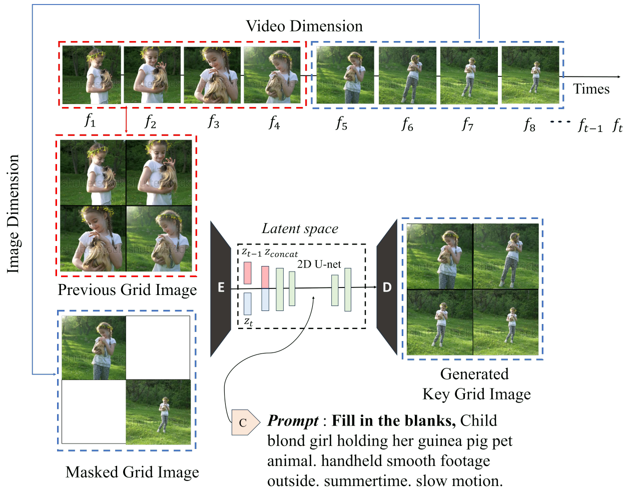

Video Dimension
Times
f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_{t-1} f_t
Image Dimension
Previous Grid Image
Latent space
z_{t-1} z_{concat}
2D U-net
z_t
E
D
Generated Key Grid Image
Masked Grid Image
C
Prompt : Fill in the blanks, Child blond girl holding her guinea pig pet animal. handheld smooth footage outside. summertime. slow motion.

Figure 3. **Visualization of the interpolation model training**. In the training process, we select 8 frames f_1 to f_8 from the frames of the original video in chronological order. Among them, we use f_1 to f_4 as the previous grid image, which serves as a condition and f_5 to f_8 as the input image which has masked on f_6 and f_7.

3.3. Inference for Video Generation

Our model conducts text-to-video generation following the process described in Figure 4. The inference process of our model consists of three steps: key grid image generation, 1-step interpolation, and 2-step interpolation. Firstly, we utilize our key grid image generation model to generate a key grid image. The generated key grid image comprises four

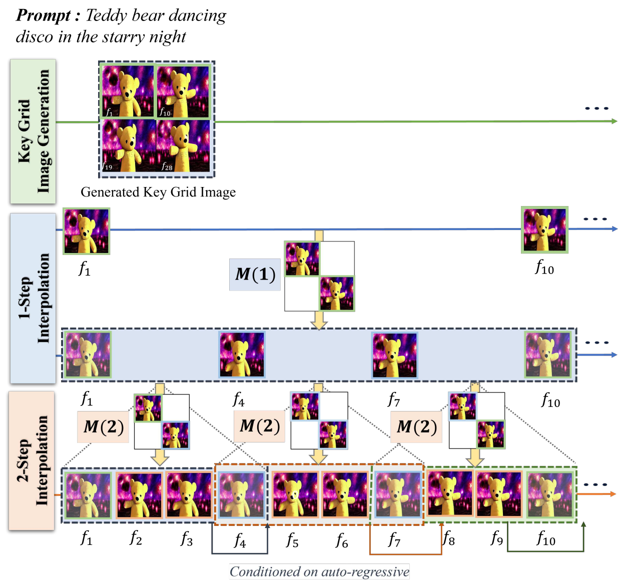


Figure 4. **Our inference procedure.** We generate a key grid image following a text prompt with our key grid image generation model. Our interpolation model generates frames between them given the masked grid image (denoted as $M(1)$ and $M(2)$ in the figure), while also ensuring temporal consistency by generating frames autoregressively conditioned on previous frames.

inside frames, with each frame being spaced at an interval of nine frames. Secondly, as depicted in Figure 4 ($M1$), we interpolate between the first and second frames by applying masks, namely between f_1 and f_{10} in the figure, to fill the gap with two frames. We repeat this process to generate more fine-grained frames in Figure 4 ($M2$). To enforce temporal consistency, when generating frames between the first and fourth frames in a grid image, we use previously generated grid image as conditioning image and repeat this process in an autoregressive manner. Consequently, by integrating the four inside frames produced via the key grid image generation process with the eight frames derived from interpolation steps, we successfully generate a video composed of 28 frames, derived from 2×2 grid images.

3.4. Video Generation with More Frames

To expand more frames from a text prompt, we train a next key grid generation model. It generates the next key grid image autoregressively, conditioned on the previous key grid image. Then we interpolate the newly generated key grid image, as illustrated in Figure 4. Consequently, our model is capable of generating more frames with both context and temporal consistency while adhering to a fixed GPU memory constraint as described in Section 4.4.

3.5. Extensions of Our Method

Text-guided video manipulation. As discussed above, by reducing video generation to image generation, we can apply various image-based methods to video domain. We explore text-guided video manipulation among them. First, we select four frames from the original video to create a key grid image. Then with the prompt as a condition, we manipulate the key grid image by using Instructpix2pix [5]. Subsequently, we generate the video by interpolating the key grid image using our interpolation model. Our interpolation model can generate videos with temporal consistency by autoregressively conditioning on previously generated frames. Also we empirically find that the Instructpix2pix [5] model does not change the temporal order and according to the [5], this model edits style while preserving contents approximately.

Video generation with higher resolution. Since we generate videos using a text-to-image model, we can apply text-to-image model with high resolution such as SD-XL [20] on our grid diffusion method. SD-XL can generate images with a resolution of 1024×1024. Therefore, by applying a 2×2 grid, as shown in Figure 9, we can generate a video with a resolution of 510×510. This shows that our approach can be flexibly extended to high resolution video generation by using text-to-image models.

4. Experiment

4.1. Experimental Setup

Datasets. To train our model, we use randomly selected 0.1 million samples from Webvid-10M [2]. Webvid-10M consists of a total of 10.7 million short videos, each of which has a paired textual description. To evaluate video generation performance, we use three datasets in a zero-shot manner: MSR-VTT [37], UCF-101 [30], and CGCaption. The CGCaption dataset, which we created, comprises a total of 500 prompts from GPT-4 [19] to assess video generation performance for real-world prompts. We provide these captions in the supplementary material.

Baselines. For evaluation, we compare our model to existing text-to-video generation models such as [1, 10, 15, 17, 29, 34, 35, 46]. VideoFusion [17] trained on Webvid-10M [2] and other public datasets, based on diffusion model, utilizes a 3D U-Net and is designed for efficiency. Additionally, it provides a pre-trained model, which serves as the main baseline for our experiments. For other models, we use the respective reported scores. As a baseline for video generation with more frames, along with VideoFusion [17], we additionally use FreeNoise [21] that focuses on text to long video generation and also provide a pre-trained model.

Implementation detail. We fine-tune the key grid image generation and interpolation models using Stable Diffusion 1.5 [26]. In the key grid image generation model, the inference step is set to 80, the batch size is 28, and the training step is 82K. For both 1-step and 2-step interpolation models, the inference step is set to 20, the batch size is 20, and the training step is 54K. We use two NVIDIA A100 80GB GPUs for training and 50K training steps. Please refer to the supplementary material for more detail.

4.2. Quantitative Results

4.2.1 Text to Video Generation

We compare our model with baselines using CLIPSIM [22] (average CLIP similarity between video frames and text), Frechet Video Distance (FVD) [31] and Inception score (IS) [27] as evaluation metrics.

MSR-VTT experiment. To evaluate the MSR-VTT [37] test set in a zero-shot manner, following prior work [11], we generate 2,990 videos with 16 frames and a resolution of 254×254 for each frame. As shown in Table 1, our model outperforms other methods [1, 4, 10, 15, 17, 29, 33, 36, 43, 46] trained on large datasets on CLIPSIM [22] and FVD [31], and achieved state-of-the-art performance.

UCF-101 experiment. To evaluate text-to-video generation on UCF-101 [30] in a zero-shot manner, we use the text prompts for each class, as provided by previous work [11]. For the IS score [27], we generate 20 videos for each prompt and to calculate FVD [31], we sample 2,048 videos for evaluation, following prior work [11]. As shown in Table 2, our model outperforms other models [4, 11, 17, 29, 32, 43, 46] trained on much larger datasets.

CGcaption experiment. Since the CGcaption dataset only consists of prompts, we evaluate the model using CLIPSIM [22]. As shown in Table 2, in CGcaption, with a variety of real-world prompts, our model obtains higher CLIPSIM compared to VideoFusion [17].

Previous studies [1, 14, 17, 29] were trained on large datasets such as Webvid-10M [2] or 10M subset from HD-VILA-100M [39], but our model is trained using only 0.1 million samples from Webvid-10M. According to the experimental results, our model outperforms other methods on MSR-VTT, UCF-101, and CGcaption. These results demonstrate that our model is remarkably effective, generating high-quality videos with significantly less training dataset than prior methods [1, 14, 17, 29].

4.2.2 Video Generation with More Frames

For the evaluation of video generation capabilities beyond 16 frames, we generate videos of 64 and 128 frames on the MSR-VTT [37] dataset in a zero-shot manner. All settings are consistent with Section 4.2.1. To evaluate the quality of the generated frames from video generation with more frames, we use the Block-FVD [40] which divides a video into several clips, to calculate the average FVD [31] of these clips. We also use CLIPSIM [22] to compare the text faithfulness of our model between the generated frames and text.

		MSR-VTT [37]	
Method	Data (M)	CLIPSIM (↑)	FVD (↓)
CogVideo [15]	5.4	0.2631	1294
Video LDM [4]	10	0.2929	-
Make-A-Video [29]	20	0.3049	-
Latent-Shift [1]	10	0.2773	-
MMVG [10]	10	0.2644	-
MagicVideo [46]	27	-	998
VideoFactory [33]	10	0.3005	-
VideoComposer [33]	10	0.2932	580
SimDA [36]	10	0.2945	456
Show-1 [43]	10	0.3072	538
VideoFusion [17]	10	0.2930	550
Ours	**0.1**	**0.3096**	**375**

Table 1. **Text-to-video generation on MSR-VTT [37].** Our method gives significant performance gains compared to the prior work both in CLIPSIM [22] and FVD [31] metrics. Data is training dataset size (million).

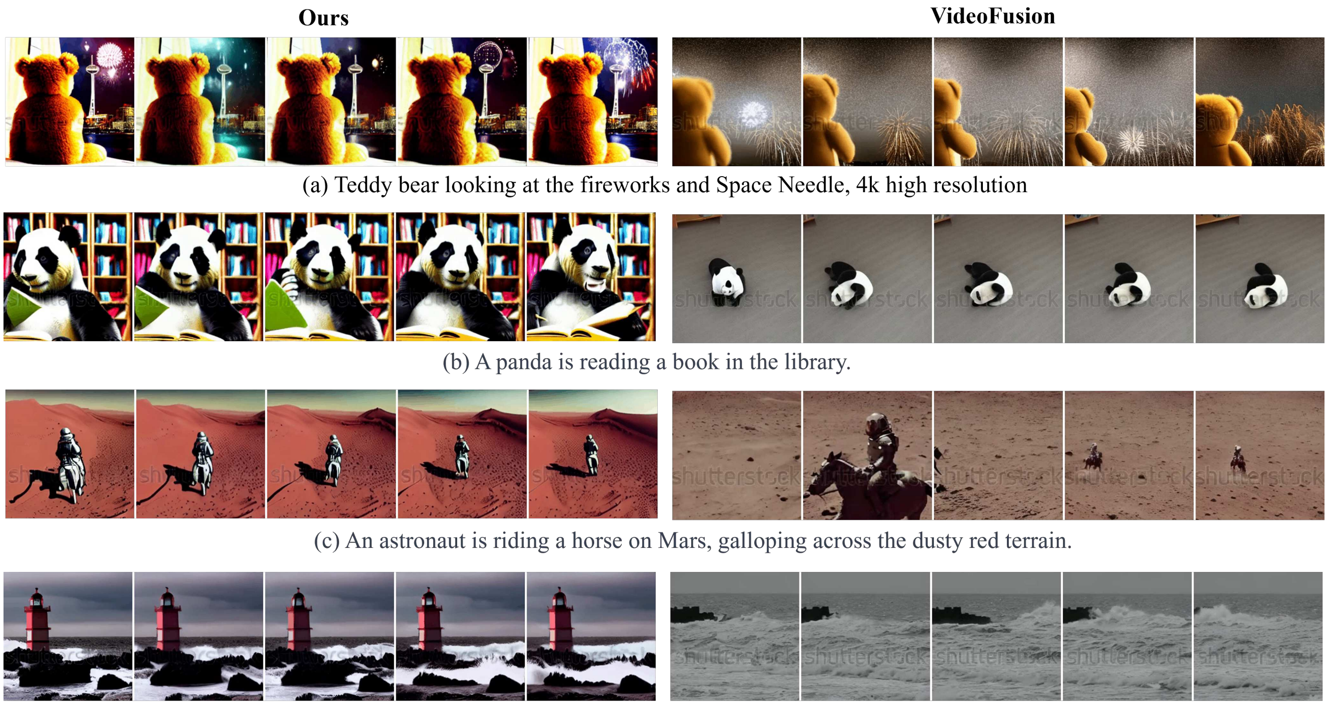
		UCF-101 [30]		CGcaption
Method	Data (M)	IS(↑)	FVD(↓)	CLIPSIM(↑)
CogVideo [15]	5.4	25.27	701	-
Make-A-Video [29]	10	33.00	367	-
Video LDM [4]	10	33.45	550	-
MagicVideo [46]	10	-	655	-
VideoFactory [32]	10	-	410	-
Show-1 [43]	10	35.42	394	-
PYoCo [11]	10	47.76	355	-
VideoFusion [17]	10	-	639	0.3025
Ours	**0.1**	**62.88**	**340**	**0.3282**

Table 2. **Text-to-video generation on UCF-101 [30] and CGcaption.** Our method gives significant performance gains compared to the prior work both in IS [27], FVD [31] and CLIPSIM [22] metrics.

		MSR-VTT [37]	
	Frames	CLIPSIM (↑)	B-FVD-16 (↓)
VideoFusion [17]	64	0.2626	1106
FreeNoise [21]	64	0.2996	517
Ours	64	**0.3044**	**370**
VideoFusion [17]	128	0.2532	1239
FreeNoise [21]	128	**0.3034**	726
Ours	128	0.3000	**364**

Table 3. **Text-to-video generation on MSR-VTT [37] with more frames.** To simplify, we name BlockFVD [40] as B-FVD-X where X denotes the length of the short clips.

We chose VideoFusion [17] and FreeeNoise [21] capable of generating more frames from a prompt as our baselines, and they also provide pre-trained models to use. Table 3 shows the results of the performance evaluation for text-to-video generation with more frames. As can be seen in Table 3, our model shows better performance than VideoFusion [17]. In comparison with FreeNoise [21], which focuses on long video generation, our model shows competitive results in CLIPSIM [22] and better results in Block-



(a) Teddy bear looking at the fireworks and Space Needle, 4k high resolution

(b) A panda is reading a book in the library.

(c) An astronaut is riding a horse on Mars, galloping across the dusty red terrain.

(d) Waves crashing against a lone lighthouse.

Figure 5. **Text-to-video generation comparison with VideoFusion [17].** Our model can generate high-quality videos that align better with the given text. **Please refer to the supplementary material for more video samples.**

Ours vs. VideoFusion [17]	TM		VQ		TC		MQ	
	WIN	LOSS	WIN	LOSS	WIN	LOSS	WIN	LOSS
MSR-VTT	48.40	17.08	46.32	29.50	48.45	28.57	47.53	25.00
UCF-101	58.54	17.00	55.10	30.46	54.12	28.08	56.54	29.16
CGcaption	51.25	28.34	60.00	24.28	52.10	23.16	55.32	25.08
Ours vs. VideoCrafter [7]	TM		VQ		TC		MQ	
	WIN	LOSS	WIN	LOSS	WIN	LOSS	WIN	LOSS
MSR-VTT	47.00	18.50	48.32	27.36	52.49	26.31	46.21	24.79
UCF-101	56.00	26.36	57.40	25.47	56.32	19.50	55.50	22.82
CGcaption	50.47	29.20	55.75	26.28	58.51	22.34	56.80	29.76

Table 4. **Comparison with VideoFusion [17] and VideoCrafter [7] in human evaluation on three datasets.** TM is text matching, VQ is video quality, TC is temporal consistency, and MQ is motion quality. We report winning and loss percentages of ours and omit TIE due to space.

FVD [40]. These results demonstrate that our model can generate videos with more frames effectively with a fixed GPU memory consumption.

4.3. Qualitative Results

4.3.1 Text to Video Generation

Qualitative analysis. Figure 5 shows the videos generated by our model and VideoFusion [17]. As shown in Figure 5, our model is capable of generating videos that are more aligned with the given text compared to VideoFusion. As observed in Figure 5, our model generates videos with varied motions, effectively representing the content of the text.

Human evaluation. We conduct human evaluation on Amazon Mechanical Turk (AMT) with 30 participants to evaluate text matching, video quality, temporal consistency and motion quality by our method in comparison to VideoFusion [17] and VideoCrafter [7] which are publicly available. For human evaluation, we randomly sample 100 generated videos from each of MSR-VTT [37], UCF-101 [30], and CGcaption datasets, in total 300 samples. Please refer to the supplementary material for more details. Table 4 shows the results of human evaluation: in all aspects, participants preferred our model significantly more than the baselines. These results demonstrate that our model is more suitable for text matching and capable of generating high-quality videos while maintaining temporal consistency and motion quality, despite using a smaller training dataset compared to VideoFusion [17] and VideoCrafter [7].

4.3.2 Video Generation with More Frames

Figure 6 shows the results of our model and baselines [17, 21] when generating more frames. As seen in Figure 6, our model generates a 128 frame video maintaining temporal consistency and text alignment. In contrast, when

Input Prompt: *there is a man skating through the iceland filled with trees*

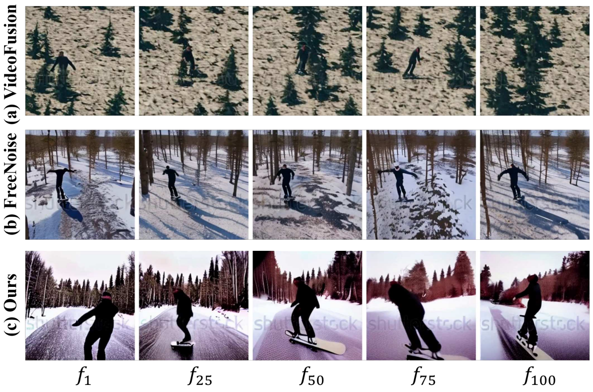


Figure 6. **Text-to-video generation comparison with VideoFusion [17] and FreeNoise [21] on MSR-VTT for 128 frames.**

generating videos longer than 16 frames with VideoFusion [17], it generates videos with background noise or a more monotonous outcome. This may be due to the difference in the number of frames in the training set videos and the number of inference frames [21]. As illustrated in Figure 6, when considering 25 frame intervals, our model generates videos with more dynamic motion, while FreeNoise [21] tends to generate videos with relatively static motions. These results indicate that our model maintains competitive quality in more frame generation, even when compared with models focused on long video generation such as FreeNoise [21].

4.3.3 Text-Guided Video Manipulation

As mentioned in Section 3.5, by exploring a new method from image manipulation to video manipulation, we can manipulate videos easily and simply, without the need for additional training for video manipulation. Figure 7 shows the result of video manipulation derived from image manipulation. This provides the opportunity for diverse extensions in the video manipulation task. More samples are provided in the supplementary material.

4.4. Efficiency Comparison

To evaluate the efficiency of our model, we compare the inference GPU memory usage of VideoFusion [17] and FreeNoise [21] with our model based on the number of frames in the video. As shown in Figure 8, the GPU memory usage of our model remains almost the same as the memory usage when generating a single image from Stable Diffusion [26], regardless of the number of frames in the video. As observed, our model demonstrates a decrease of 74.08% in consumed memory when generating 128 frames, compared to FreeNoise [21]. But it is observed that the GPU memory usage of VideoFusion [17] and FreeNoise [21] increases significantly as the number of frames in the video increased, which may prevent to run the model in limited GPU memory unlike our model. In terms of inference time cost, FreeNoise [21] takes 1.68 seconds per frame for a 64 frames video and 1.62 seconds per frame for a 128 frame video, while our process requires 1.71 seconds per frame for 64 frames and 1.76 seconds per frame for 128 frames. These results show that our model is efficient in terms of GPU memory consumption and maintains competitive inference time cost when generating videos with more frames. Although generating video with more frames may take longer, we can still generate videos on a fixed GPU memory.

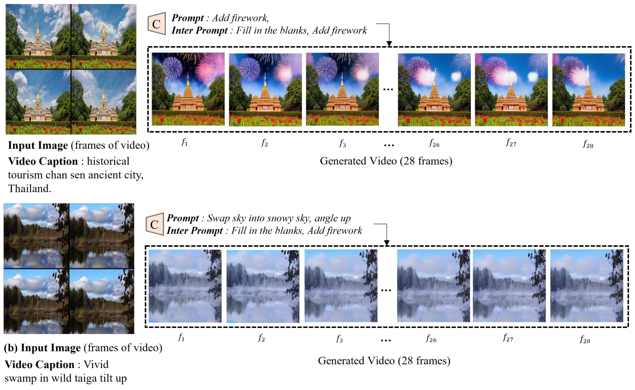

Prompt : Add firework,
Inter Prompt : Fill in the blanks, Add firework
Input Image (frames of video)
Video Caption : historical tourism chan sen ancient city, Thailand.
f_1
f_2
f_3
...
f_{26}
f_{27}
f_{28}
Generated Video (28 frames)
Prompt : Swap sky into snowy sky, angle up
Inter Prompt : Fill in the blanks, Add firework
(b) Input Image (frames of video)
Video Caption : Vivid swamp in wild taiga tilt up
Generated Video (28 frames)

Figure 7. **The result of video manipulation**. We select input images from Webvid-10M [2] videos.

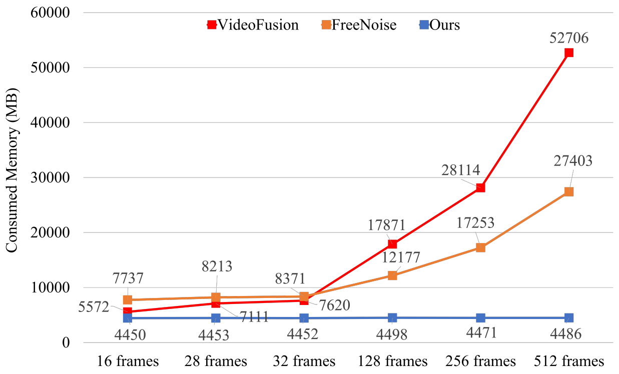


Figure 8. **The efficiency comparison of GPU memory usage.**

4.5. Ablation Study

To explore the impact of the proposed components in key grid image generation, interpolation models, convolution layers and attention layers, we conduct an ablation study on MSR-VTT [37].

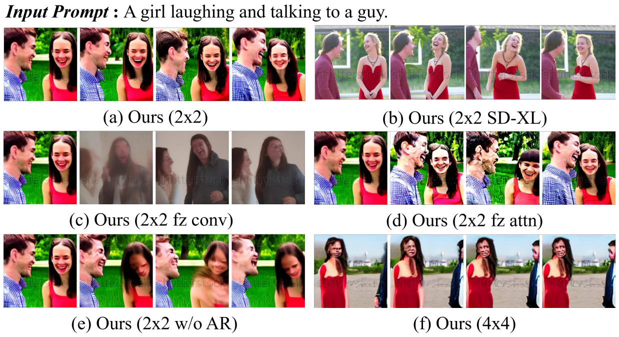
Autoregressive Frame Interpolation. Table 5 shows the results of the ablation study for both the autoregressive interpolation model and the non-autoregressive model. The non-autoregressive interpolation model is trained using a prompt "Fill in the blanks", which simply fills in the masked grid image without any conditions from previous frames. Our model with autoregressive interpolation outperforms the non-autoregressive one in CLIPSIM [22] and FVD [31] on MSR-VTT [37]. As shown in Figure 9, the autoregressive model exhibits better temporal consistency and produce smoother frame generation compared to the non-autoregressive model. Also we conduct human evaluation under the same settings as Section 4.3.1, and as shown in Table 5, our model outperformed ablated non-autoregressive model by a significant margin. These results indicate that autoregressive interpolation model interpolates each frame in a dependent manner, which helps keep temporal consistency across the entire video.

4×4 Key Grid Image Generation. We generate a key grid image consisting of four inside frames in a 2×2 grid that represents the video from text. In ablation study, as shown in Figure 9, instead of using key grid image in a 2×2 grid, we generate key grid image in a 4×4 grid. A 4×4 grid image is composed of 16 frames with a resolution of 126×126. As shown in Table 5, the 2×2 model shows higher CLIPSIM [22] and IS score [27] than the 4×4 model. This is because the pre-trained Stable Diffusion (SD 1.5) [26] model appears to be more capable of generating higher quality images for the 254×254 resolution of a 2×2 grid, compared to the 126×126 resolution of a 4×4 grid. However, the 4×4 model performed slightly better than the 2×2 model in FVD score [31]. This may be due to the tendency for the FVD to worsen as the resolution of the video increases [17]. In the human evaluation results, as shown in Table 5, our model outperformed 4×4 grid ablated model by a significant margin. However, even though 4×4 has a smaller resolution and lower quality than 2×2, generating 16 images of 126×126 is more efficient and requires no interpolation.

Convolution Layers vs. Attention Layers. Since we utilize pre-trained SD 1.5 [26] models for U-Net and VAE, our model can generate grid images by representing spatial structure in the latent space. To analyze the impact of the convolution and attention layers on U-Net of our model, we fine-tuned SD 1.5 model with frozen attention layers and another model with frozen convolution layers. In Table 5, our model with frozen attention layers performed slightly worse compared to our model in FVD [31] and CLIPSIM [22]. But our model with frozen convolution layers performed significantly worse than our model in FVD and CLIPSIM. Also in Figure 9 (c), the model with frozen convolution layers generates a grid image that fails to maintain consistency with inside frames. These results show that the convolution layers in U-Net may have the ability to cover long range correlation in the latent of the grid image.

	MSR-VTT		Human evaluation							
			TM		VQ		TC		MQ	
	CLIPSIM	FVD	WIN	LOSS	WIN	LOSS	WIN	LOSS	WIN	LOSS
Ours (4×4)	0.2902	**343**	50.50	17.70	60.28	17.34	58.50	19.26	56.07	22.46
Ours (w/o AR)	0.2982	504	56.50	22.53	60.83	21.77	55.13	19.13	59.13	21.86
Ours (fz conv)	0.2872	724	75.43	10.16	76.20	11.16	75.53	12.26	71.16	13.83
Ours (fz attn)	0.2956	512	51.26	26.26	49.63	26.86	52.96	22.93	52.53	23.23
Ours	**0.3096**	375	-							

Table 5. **Video generation evaluation on MSR-VTT [37] in ablation study.** The video generated by the 4×4 grid model has a resolution of 126×126. For human evaluation, we report winning and loss percentages of ours vs. ablated models and omit TIE due to space.



(a) Ours (2x2) (b) Ours (2x2 SD-XL)

(c) Ours (2x2 fz conv) (d) Ours (2x2 fz attn)

(e) Ours (2x2 w/o AR) (f) Ours (4x4)

Figure 9. **Text-to-video generation comparison with ablated models and Ours (SD-XL) on MSR-VTT [37].**

5. Conclusion

In this paper, we propose novel grid diffusion models for text-to-video generation, addressing the challenges posed by the lack of large text-video paired datasets and the high GPU memory costs on video generation. Unlike previous studies, by representing the video as a grid image, we can generate high-quality videos using a fixed amount of GPU memory, regardless of the number of frames. Furthermore, various methods in the image dimension can be easily applied to our model such as video manipulation. Our model has a limitation as it relies on a pre-trained text-to-image model, but the generated videos contain rich visual content. In the experimental results, our model outperforms the baselines in both quantitative and qualitative evaluations. As future work, we will explore applying our model to other generative tasks with different modalities such as sound.

Acknowledgments. We thank Dong Gyu Lee for the help with human evaluation. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00608, Artificial intelligence research about multi-modal interactions for empathetic conversations with humans, No.2021-0-02068, Artificial Intelligence Innovation Hub & No.2020-0-01336, Artificial Intelligence Graduate School Program (UNIST)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00219959).

References

[1] 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, 4, 5

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 4, 5, 7

[3] 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. 2

[4] 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. *arXiv preprint arXiv:2304.08818*, 2023. 1, 5

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1, 3, 4

[6] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023. 1

[7] 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

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1

[10] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. *arXiv preprint arXiv:2211.12824*, 2022. 4, 5

[11] 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. 1, 2, 5

[12] 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. 1, 2

[13] 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. 1, 2

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

[15] 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. 1, 2, 4, 5

[16] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 1

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 1, 2, 4, 5, 6, 7, 8

[18] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[19] OpenAI. Gpt-4 technical report, 2023. 4

[20] 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. 4

[21] 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. 4, 5, 6, 7

[22] 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. 5, 8

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2

[24] 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. 1, 2

[25] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022. 1, 2, 4, 7, 8

[27] 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. 5, 8

[28] 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. *arXiv preprint arXiv:2210.08402*, 2022. 2, 3

[29] 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. 1, 2, 4, 5

[30] 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. 4, 5, 6

[31] 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. 5, 8

[32] 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. 5

[33] 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. 5

[34] 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. 1, 2, 4

[35] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pretraining for neural visual world creation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 4

[36] 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. 5

[37] 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. 4, 5, 6, 7, 8

[38] T Xu, P Zhang, Q Huang, H Zhang, Z Gan, X Huang, and X AttnGAN He. Fine-grained text to image generation with attentional generative adversarial networks. arxiv 2017. *arXiv preprint arXiv:1711.10485*. 2

[39] 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 *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[40] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 5, 6

[41] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[42] Mingkuan Yuan and Yuxin Peng. Text-to-image synthesis via symmetrical distillation networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1407–1415, 2018. 2

[43] 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. 5

[44] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[45] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1

[46] 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. 1, 4, 5

[47] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. *arXiv preprint arXiv:2211.15388*, 2022. 1