VideoGen: A Reference-Guided Latent Diffusion Approach for High Definition Text-to-Video Generation

Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang,
Fu Li, Haocheng Feng, Errui Ding, Jingdong Wang
Department of Computer Vision Technology (VIS), Baidu Inc.
lixin41@baidu.com

Abstract

In this paper, we present VideoGen, a text-to-video generation approach, which can generate a high-definition video with high frame fidelity and strong temporal consistency using reference-guided latent diffusion. We leverage an off-the-shelf text-to-image generation model, e.g., Stable Diffusion, to generate an image with high content quality from the text prompt, as a reference image to guide video generation. Then, we introduce an efficient cascaded latent diffusion module conditioned on both the reference image and the text prompt, for generating latent video representations, followed by a flow-based temporal upsampling step to improve the temporal resolution. Finally, we map latent video representations into a high-definition video through an enhanced video decoder. During training, we use the first frame of a ground-truth video as the reference image for training the cascaded latent diffusion module. The main characterises of our approach include: the reference image generated by the text-to-image model improves the visual fidelity; using it as the condition makes the diffusion model focus more on learning the video dynamics; and the video decoder is trained over unlabeled video data, thus benefiting from high-quality easily-available videos. VideoGen sets a new state-of-the-art in text-to-video generation in terms of both qualitative and quantitative evaluation. See `https://videogen.github.io/VideoGen/` *for more samples.*

1. Introduction

There have been great progress in text-to-image (`T2I`) generation systems, such as DALL-E2 [12], Imagen [42], Cogview [10], Latent Diffusion [40], and so on. In contrast, text-to-video (`T2V`) generation, creating videos from text description, is still a challenging task as it requires not only high-quality visual content, but also temporally-smooth and realistic motion that matches the text. Moreover, it is hard to find large-scale datasets of text-video pairs.

In addition to extending the `T2I` network architecture, several recent `T2V` techniques explore the trained `T2I` model for improving the visual fidelity, e.g., utilizing the `T2I` model weights, or exploring image-text data. For example, CogVideo [23] and Make-A-Video [46] make use of the `T2I` model, by freezing or fine-tuning the `T2I` model weights. NÜWA [59] and Imagen Video [19] instead explore image-text pairs to improve `T2V` model training, through pre-training or joint-training.

In this paper, we propose VideoGen for generating a high-quality and temporally-smooth video from a text description. We leverage a `T2I` model to generate a high-quality image, which is used as a reference to guide `T2V` generation. Then, we adopt a cascaded latent video diffusion module, conditioned on the reference image and the text description, to generate a sequence of high-resolution smooth latent representations. We optionally use a flow-based scheme to temporally upsample the latent representation sequence. Finally, we learn a video decoder to map the latent representation sequence to a video.

The benefits of using a `T2I` model to generate a reference image lie in two-fold. On the one hand, the visual fidelity of the generated video is increased. This benefits from that our approach makes use of the large dataset of image-text pairs, which is richer and more diverse than the dataset of video-text pairs, through using the `T2I` model. This is more training-efficient compared to Imagen Video that needs to use the image-text pairs for joint training. On the other hand, using the reference image to guide the cascaded latent video diffusion model frees the diffusion model from learning visual content, and makes it focus more on learning the video dynamics. We believe that this is an extra advantage compared to the methods merely using the `T2I` model parameters [23, 46].

Furthermore, our video decoder only needs the latent representation sequence as input to generate a video, without requiring the text description. This enables us to train the video decoder over a larger set of easily-available un-

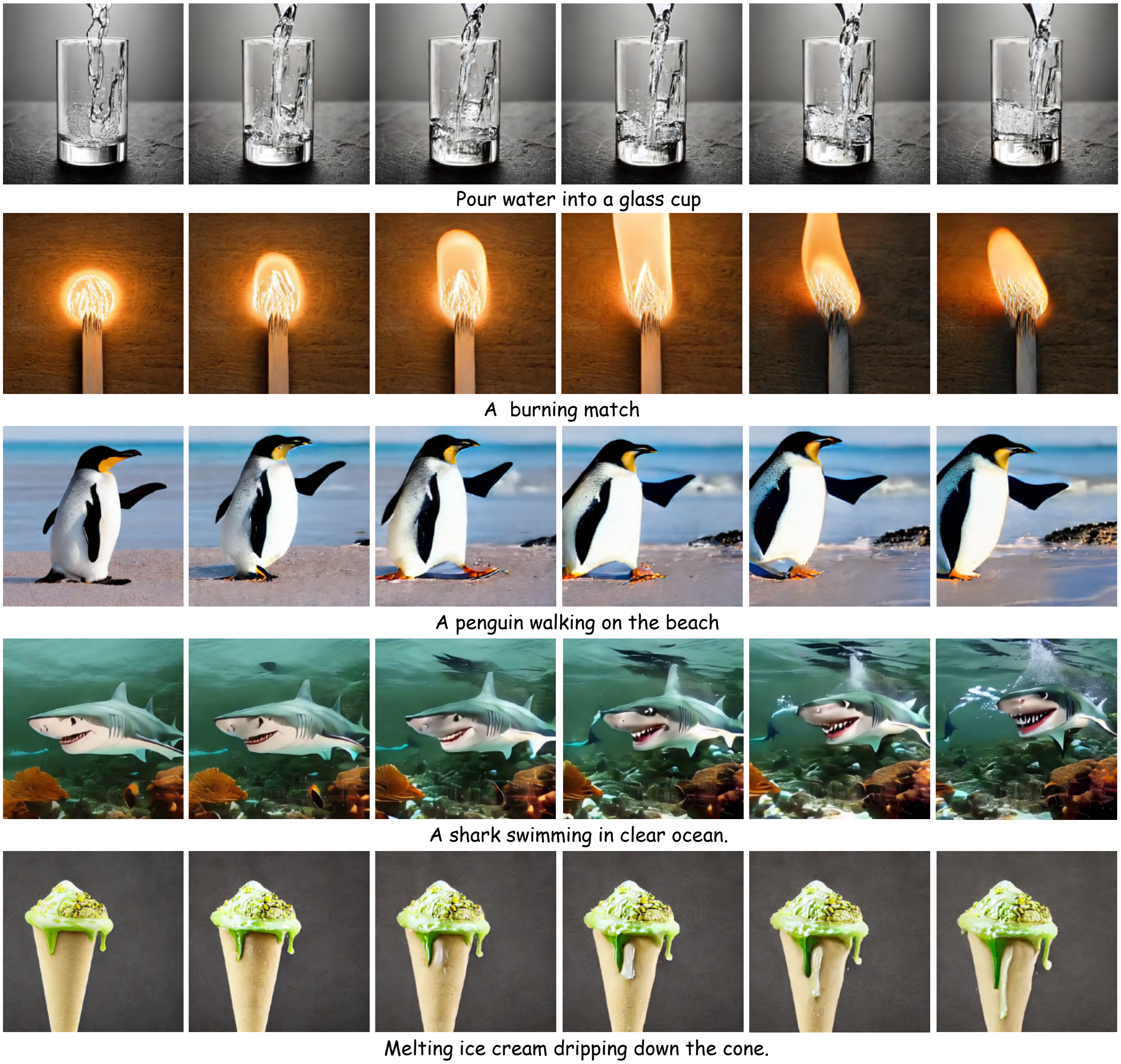

Figure 1. T2V generation examples of VideoGen. Our generated videos have rich texture details and stable temporal consistency. It is strongly recommended to zoom in to see more details.

labeled (unpaired) videos other than only video-text pairs. As a result, our approach benefits from high-quality video data, improving motion smoothness and motion realism of the generated video. Our key contributions are as follows:

- We leverage an off-the-shelf `T2I` model to generate an image from text description as a reference image, for improving frame content quality.

- We present an efficient and effective cascaded latent video diffusion model conditioned on the text description, as well as the reference image as the condition which makes the diffusion model focus more on learning the video motion.

- We are able to train the video decoder using easily-available unlabeled (unpaired) high-quality video data, which boosts visual fidelity and motion consistency of the generated video.

- We evaluate VideoGen against representative `T2V` methods and present state-of-the-art results in terms of quantitative and qualitative measures.

2. Related Work

Diffusion models. The generative technology has experienced rapid development, from the generative adversarial networks [17] in the past few years to the very popular diffusion models recently. Diffusion models [47, 20] have shown surprising potential and made great progress in generative tasks, such as text-to-speech [6, 7, 26], text-to-image [42, 37, 35, 40, 32, 2, 14, 5], text-to-3D [36, 57], text-to-video [22, 46, 18, 69, 19, 60, 23], image2image [43, 4, 56, 68, 41, 3] and vid2vid [12, 3]. Especially in the generation of images, such as Stable Diffusion [40], has reached the level of professional illustrators, which greatly improves the work efficiency of artists.

Text-to-image generation. The past years have witnessed tremendous progress in image-to-text generation. The early systems are mainly based on GAN [17], e.g., StyleCLIP [34], StyleGAN-NADA [15], VQGAN-CLIP [9], StyleT2I [29]. The most recent success is from the development of denoising diffusion model [20] and its efficient extension, latent diffusion model [40]. Examples include: DALL-E [38], DALL-E2 [37], Imagen [42], Stable Diffusion [40], CogView [10], Parti [64], GLIDE [32].

Our approach takes advantages of latent diffusion model [40] for text-to-video generation. This not only improves the diffusion sampling efficiency, but also allows to design the video decoder that only relies on videos, not on texts, allowing that the video decoder can be trained on high-quality unlabeled videos.

Text-to-video generation. Early text-to-video techniques include: leveraging a VAE with recurrent attention, e.g.,Sync-DRAW [30], and extending GAN from image generation to video generation [33, 28]. Other developments include GODIVA [58], NÜWA [59], CogVideo [23].

More recent approaches include: Tune-A-Video [60] and Dreamix [31] for applications with fine-tuning, Make-A-Video [46], MagicVideo [69], Video Diffusion Model [22] and Imagen Video [19], latent video diffusion models [18], which extend diffusion models from image generation to video generation,

Our approach differs from previous works in several aspects. First, our approach leverages the pretrained text-to-image generation model to generate a high-quality image for guiding video generation, leading to high visual fidelity of the generated video. This is clearly different from previous approaches. In Make-A-Video [46], an image is used to generate an embedding to replace the text embedding for image animation. In contrast, our approach uses an image as reference to guide video content generation. What's more, the image in Make-A-Video is mapped to an embedding through CLIP image encoder, that is mainly about semantic. In contrast, our approach uses the encoder trained with auto-encoder, and the output latent contains both semantics and details for reconstruction. This is why the results of Make-A-Video are more blurry. Second, we adopt latent video diffusion model, leading to more efficient diffusion sampling in comparison to Make-A-Video [46] and Imagen Video [19]. Reference-guidance for latent video diffusion model makes our approach differ from [18] that only conducts the study on a small dataset. Last, our design allows us to train the video decoder using high-quality unpaired videos.

3. Approach

Our approach VideoGen receives a text description, and generates a video. The inference pipeline is depicted in Figure 2. We generate a reference image from a pretrained and frozen Text-to-Image generation model. We then compute the embeddings of the input text and the reference image from pretrained and frozen text and image encoders. We send the two embeddings as the conditions for reference-guided latent video diffusion for generating latent video representation, followed by a flow-based temporal super-resolution module. Finally, we map the latent video representation to a video through a video decoder.

3.1. Reference Image Generation

We leverage an off-the-shelf text-to-image (`T2I`) generation model, which is trained over a large set of image-text pairs and can generate high-quality image. In our implementation, we adopt the SOTA model, Stable Diffusion[1] without any processing. We feed the text prompt into the `T2I` model. The resulting high-fidelity image is used as a reference image, and plays a critical role for effectively guiding subsequent latent representation sequence generation. During the training, we simply pick the first frame of the video as the reference, which empirically works well.

3.2. Reference-Guided Latent Video Diffusion

Cascaded latent video diffusion consists of three consecutive components: a latent video representation diffusion network, generating representations of spatial resolution 16×16 and temporal resolution 16, and two spatially super-resolution diffusion networks, raising the spatial resolutions to 32×32 and 64×64.

Architecture. We extend the 2D latent diffusion model [40] to the 3D latent diffusion model through taking into consideration the temporal dimension. We make two main modifications over the key building block that now supports both spatial and temporal dimensions.

Following Make-A-Video [46], we simply stack a 1D temporal convolution following each 2D spatial convolu-

[1] https://github.com/CompVis/stable-diffusion

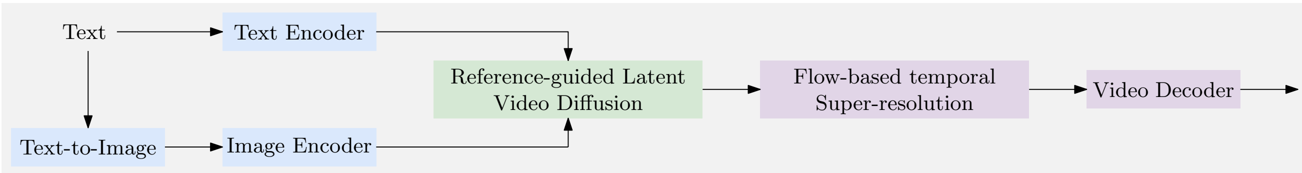

Figure 2. The VideoGen inference pipeline. The input text is fed into a pretrained Text-to-Image generation model, generating a reference image. The reference image and the input text are sent to a pretrained Image Encoder and a pretained Text Encoder. The output text and image embeddings are used as the conditions of Reference-guided Latent Video Diffusion, outputting the latent video representation. Then Flow-based temporal Super-resolution increases the temporal resolution, and is followed by Video Decoder, generating the final video. During the training process, the reference image is the first frame of the video.

tional layer in the network. The 2D spatial convolution is conducted for each frame separately, e.g., 16 frames in our implementation. Similarly, the 1D temporal convolution is conducted for each spatial position separately, e.g., 16×16, 32×32, and 64×64 for the three diffusion networks. Similar to Make-A-Video [46]. such a modification to the building block enables us to use the pretrained `T2I` model parameters to initialize the 2D convolutions. Similarly, we stack a temporal attention following each spatial attention.

Condition injection. We follow the scheme in LDM [40] to inject the text embedding into the network using cross-attention. We project the text description into an intermediate representation through a pretrained text encoder, CLIP text encoder in our implementation. The intermediate representation is then mapped into each diffusion network using a cross-attention layer.

The later diffusion network uses the bilinear $2\times$ upsampled representation output from the last diffusion network as an extra condition and concatenates it into the input. We follow Make-A-Video [46] to use FPS as a condition and inject its embedding into each diffusion model.

We project the reference image to a representation through a pretrained image encoder. In our implementation, we use the image encoder of the auto-encoder in Stable Diffusion, and process the image with three resolutions (16×16, 32×32, and 64×64), each corresponding to a diffusion network. We inject the representation of the reference image into the network by concatenating it with the first-frame representation of the input of the diffusion model, and concatenating zero representations with the representations corresponding to other frames.

3.3. Flow-based Temporal Super-resolution

We perform temporal super-resolution in the latent representation space. We estimate the motion flow according to the representations using a latent motion flow network. Then we warp the representations according to the estimated motion flow, and obtain a coarse longer video representations with $2\times$ upsampling. We next send each warped representation to a denoising diffusion network as a condition to get a refined representation. The final warp representation is a combination of the low-frequency component of the warped coarse representation and the high-frequency component of the refined representation. Consistent to the observation [8], our experiments find that the combined representations lead to more stable video generation. We perform this process three times and get $8\times$ upsampled video representations.

3.4. Video Decoder

The video decoder maps the video from the latent representation space to pixel space. We modify the Stable Diffusion $8\times$ upsampling image decoder for the video decoder. We stack a 1D temporal convolution following each 2D convolution and a temporal attention following each spatial attention. This modification also allows us to initialize the parameters of 2D convolutions and spatial attentions in the video decoder using the parameters of the pretrained image decoder.

3.5. Training

Our approach leverages existing models, e.g., CLIP text encoder for text description encoding, Stable Diffusion `T2I` generation model for reference image generation, Stable Diffusion image encoder for reference image encoding. In our implementation, we freeze the three models without retraining. The other three modules are independently trained from the video data with the help of pretrained image models. The details are as follows.

Reference-guided cascaded latent video diffusion. We compute the video representations by sending each frame into the image encoder as the denoising diffusion target. At each stage, the video spatial resolution is processed to match the spatial resolution of the latent representations. We simply pick the first frame in the video as the reference image for training.

The 2D convolution and spatial attention parameters of the first diffusion network are initialized from the pretrained Stable Diffusion `T2I` generation model. The temporal convolution and attention layers are initialized as the identity function. The second (third) diffusion network is initialized as the weights of the trained first (second) diffusion

Figure 3. For a text prompt, different reference images generate different videos.

network. The three diffusion networks are only the components receiving video-text pairs, WebVid-10M [1], for training.

Flow-based temporal super-resolution. We estimate the motion flow by extending IFRNet [25] from the pixel space to the latent representation space. We slightly modify the IFRNet architecture and simply change the first layer for processing latent representations. The ground-truth motion flow in the latent representation space is computed as: compute the motion flow in the pixel space using the pretrained IFRNet and resize the motion flow to the spatial size of the latent representation space.

The input representations of the flow-based temporal super-resolution part are directly computed from low temporal-resolution video. The ground-truth target representations of the denoising diffusion network for warped representation refinement are constructed by feeding the frames of high FPS video into the image encoder.

Video decoder. The 2D convolution and spatial attention weights are initialized from the pretrained Stable Diffusion image decoder, and the temporal convolution and attention are initialized as the identify function. During the training, we use the image encoder in StableDiffusion to extract video latent representations. We apply degradations (adding noise, blurring, and compression), which are introduced in BSRGAN [66], to the video, and extract the latent representations. The target video is still the original video, and without any processing. Video decoder and flow-based temporal super-resolution network are trained on unpaired videos with $40K$ clips of 100 frames that are collected from YouTube.

4. Experiments

4.1. Datasets and Metrics

We adopt the publicly available dataset of video-text pairs from WebVid-$10M$ [1] for training the reference-guided cascaded latent video diffusion network. We collected over $2,000$ $4K$-resolution videos of 60 FPS from YouTube and extracted 40000 clips for training the flow-based temporal super-resolution network, and the video decoder. Our other basic settings follow the open-sourced Stable Diffusion code [2] and remain unchanged. All our experiments are conducted on 64 A100-80G GPUs.

We evaluate our VideoGen on UCF-101 [49] and MSR-VTT [62]. For MSR-VTT, we use all $59,800$ captions from the test set to calculate CLIPSIM [58] (average CLIP similarity between video frames and text) following [46, 59]. UCF-101 contains 13,320 video clips from 101 categories that can be grouped into body movement, human-human interaction, human-object interaction, playing musical instruments, and sports. For UCF-101, we follow Make-A-Video [46] and construct the prompt text for each class.

Following previous methods [46, 22, 23], we report commonly-used Inception Score (IS) [44] and Frechet Video Distance (FVD) [54] [54] as the evaluation metrics on UCF-101. During the evaluation, we only generated

[2]https://github.com/CompVis/stable-diffusion

Figure 4. Qualitative comparison with Make-A-Video and Imagen Video. Compared with Make-A-Video, the lake ripples, boats and trees in our video are clearer. Similarly, although the video resolution of Imagen Video reaches 1280×768, the frames are very blurry compared with our result. The watermark in the last row is because the videos in the training set WebVid-10M contain the "shutterstock" watermark.

Table 1. `T2V` results on UCF-101. We report the performance for zero-shot and fine-tuning settings.

Method	Pretrain	Class	Resolution	IS ↑	FVD ↓
			Zero-Shot Setting		
CogVideo (Chinese)	Yes	Yes	480×480	23.55	751.34
CogVideo (English)	Yes	Yes	480×480	25.27	701.59
Make-A-Video	Yes	Yes	256×256	33.00	367.23
Ours	Yes	Yes	256×256	71.61 ± 0.24	554 ± 23
			Fine-tuning Setting		
TGANv2	No	No	128×128	26.60 ± 0.47	-
DIGAN	No	No	-	32.70 ± 0.35	577 ± 22
MoCoGAN-HD	No	No	256×256	33.95 ± 0.25	700 ± 24
CogVideo	Yes	Yes	160×160	50.46	626
VDM	No	No	64×64	57.80 ± 1.3	-
LVDM	No	No	256×256	-	372 ± 11
TATS-base	Yes	Yes	128×128	79.28 ± 0.38	278 ± 11
Make-A-Video	Yes	Yes	256×256	82.55	81.25
Ours	Yes	Yes	256×256	82.78 ± 0.34	345 ± 15

16×256×256 videos, because the C3D model [53] for IS and FVD, and the clip image encoder [3] for CLIPSIM do not expect higher resolution and frame rate.

[3] https://github.com/openai/CLIP

4.2. Results

Quantitative evaluation. We compare our VideoGen with some recent text-to-video generation methods, including Make-A-Video [46], CogVideo [23], VDM [22], LVDM

Table 2. T2V results on MSR-VTT. We report average CLIPSIM scores to evaluate the text-video alignment.

Method	Zero-Shot	Resolution	CLIPSIM ↑
GODIVA	No	128×128	0.2402
Nüwa	No	336×336	0.2439
CogVideo (Chinese)	Yes	480×480	0.2614
CogVideo (English)	Yes	480×480	0.2631
Make-A-Video	Yes	256×256	0.3049
Ours	Yes	256×256	0.3127

[18], TATS [16], MagicVideo [69], DIGAN [65] and Nüwa [59], etc. Because ImagenVideo [19] has neither open source nor public datasets results, we have only made a qualitative comparison with it. The results on MSR-VTT are given in Table 2. We can see that our VideoGen achieves the highest average CLIPSIM score without any fine-tuning on MSR-VTT, proving that the generated videos and texts have good content consistency.

The results on UCF-101 given in Table 1 show that in the cases of both the zero-shot and finetuning settings, the IS score of VideoGen performs the best. In the zero-shot setting, the IS score is greatly improved compared to the second best, from 33 to 71.6. The IS index measures the quality and category diversity of generated video and the high IS index indicates that the video quality and category diversity of our generated videos are excellent.

Figure 5. Visual comparison without and with the use of reference image. As we can see, the frames with reference-guided have more texture details in dark cloud and grass areas. Please zoom in to see more details.

The key reason for better results from our approach is that we generate a high-quality reference image using a well-trained T2I generation model, and accordingly the quality of generated video content is improved.

We also report the results in terms of FVD that measures the gap between the distribution of real videos and generated videos. Our approach performs the second best in the zero-shot setting. The most possible reason is that our training data distributes more differently from the UCF-101 dataset than the training data used by Make-A-Video. In the fine-tuning setting, we do not fine-tune the text-to-image generation model, the flow-based temporal super-resolution model, and the video decoder, and only fine-tunes the first latent video diffusion model. We guess that our FVD score would be better if we fine-tune the text-to-image model for generating a reference image whose content matches the distribution of UCF-101. The fine-tuning setting is not our current focus, and our current goal is general T2V generation.

Qualitative evaluation. In Figure 1, we show some examples generated from our VideoGen. Our results show rich and clear texture details, and excellent temporal stability and motion consistency. In Figure 4, we make a visual comparison with the two recent T2V methods, Imagen Video [19] and Make-A-Video [46]. It can be seen that although the video resolution of ImagenVideo reaches 1280×768, the frames are very blurry compared with our result. Compared with Make-A-Video, the lake ripples, boats and trees in our video are clearer.

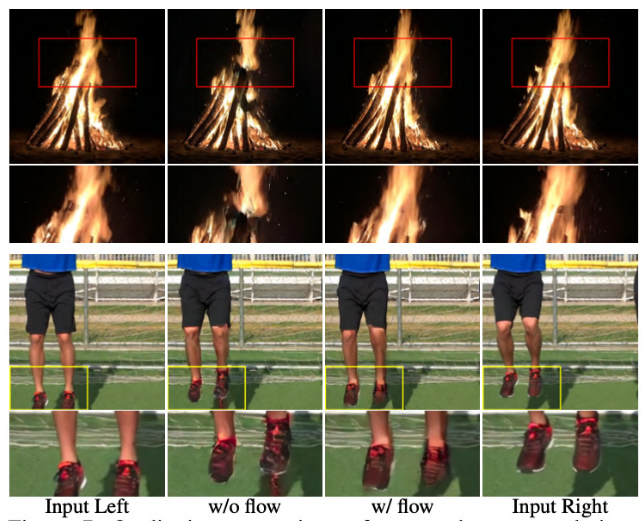

Figure 6. Qualitative comparison of temporal super-resolution without and with using motion flow. Using motion flow, the interpolated frame is more stable and more consistent to input left and right frames for the top example, and visually better for the bottom example. The first and third rows are two examples, and the second and four rows are zoomed-in of the patches in the red and yellow box.

4.3. Ablation Study

Reference image from text-to-image generation. In order to evaluate the effect of our T2V strategy guided by T2I reference, we conducted experiments by removing the reference condition for cascaded latent diffusion models. We randomly selected 1000 text prompts from the 59800 MSR-VTT test set and compared the CLIPSIM scores. We also

Table 3. Effect of reference guidance. We report average CLIPSIM score on 1000 texts randomly selected from the MSR-VTT testset. We also report the IS scores on the UCF101 dataset in the zero-shot setting.

	CLIPSIM ↑	IS ↑
without reference	0.2534	26.64 ± 0.47
with reference	0.3127	71.61 ± 0.24

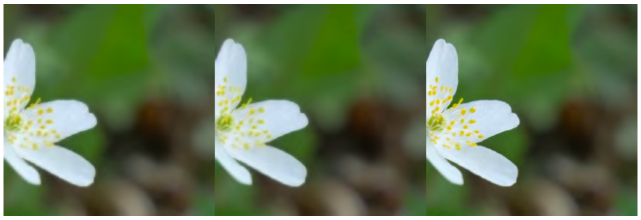

Figure 7. Visual comparison for the effectiveness of video decoder. The texture details of the the pistil and petals in our restored frame are clearer than those of original image decoder in the Stable Diffusion.

compared the IS index under zero-shot setting on the UCF-101 dataset. The comparison is given in Table 3. One can see that the `T2I` reference images greatly improve the IS and CLIPSIM scores. This empirically verifies the effectiveness of the reference image: improving the visual fidelity and helping the latent video diffusion model learn better motion. Figure 5 shows the visual comparison from the same text prompt. We can see that the visual quality and the content richness with reference image are much better. In Figure 3, we show three different reference images, with the same text prompt, our VideoGen can generate different videos.

Flow-based temporal super-resolution. We demonstrate the effectiveness of our flow-based temporal super-resolution by replacing flow-guided with spherical-interpolation guided. The comparison with two examples are given in Figure 6. We can observe that with motion flow the interpolated frames is more stable and continuous. Without flow-guided, as shown in Figure 6, the fire is broken and the right shoe has artifacts.

Video decoder. Figure 7 shows the visual comparison results between our video decoder and the original image decoder of the auto-encoder in Stable Diffusion. The frame from our video decoder has sharper textures. This is because we perform various degradations on the inputs during training, so that our video decoder has enhanced effect. Furthermore, the videos restored from the video decoder are temporally smoother.

4.4. User Study

Because Make-A-Video [46] and ImagenVideo [19], the two best performing methods at present, are not open

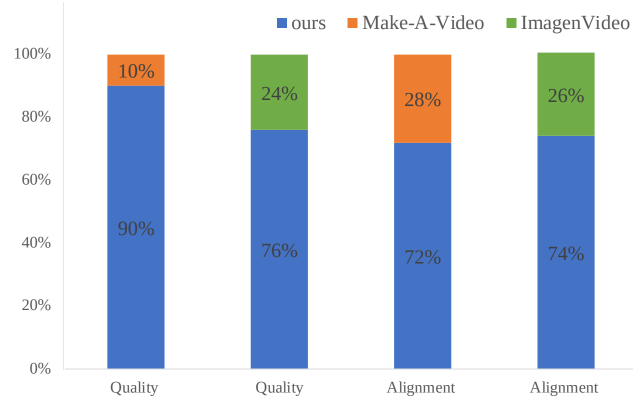

Figure 8. User Preferences. The first two bars are human evaluation results of our method compared to Make-A-Video and ImagenVideo for video quality (corresponding to the question: "Which video is of higher quality?"), respectively. Comparison with Make-A-Video, results from our approach are preferred 90%. Compared with ImagenVideo, 76% of our options are chosen. The latter two reveal the users' preference for text-video alignment ("Which video better represents the provided text prompt?"). Similarly, our VideoGen also outperforms baseline methods by a large margin.

sourced, we use the demos shown on their webpages for human evaluation. We conduct the user study on an evaluation set of 30 video prompts (randomly selected from the webpages of Make-A-Video and ImagenVideo). For each example, we ask 17 annotators to compare the video quality ("Which video is of higher quality?") and the text-video content alignment ("Which video better represents the provided text prompt?") between two videos from the baseline (ImagenVideo or Make-A-Video) and our method, presented in random order. As shown in Figure 8, in the video quality comparison with Make-A-Video, results from our VideoGen are preferred 90%. Compared with ImagenVideo, 76% of our options are chosen. Similarly, for the user study of the text-video alignment, our VideoGen also outperforms baseline methods by a large margin.

5. Conclusion

We present VideoGen, a text-to-video generation approach, and report the state-of-the-art video generation results. The success stems from: (1) Leverage the SOTA text-to-image generation system to generate a high-quality reference image, improving the visual fidelity of the generated video; (2) Use the reference image as a guidance of latent video diffusion, allowing the diffusion model to focus more on learning the motion; (3) Explore high-quality unlabeled (unpaired) video data to train a video decoder that does not depends on video-text pairs.

References

[1] 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. 5

[2] 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. 3

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 3

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

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3

[6] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 3

[7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*, 2021. 3

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356. IEEE, 2021. 4

[9] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 88–105. Springer, 2022. 3

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1, 3

[11] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1, 3

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

[14] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 3

[15] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3

[16] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 102–118. Springer, 2022. 7

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[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. 3, 7

[19] 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, 3, 7, 8

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

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

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3, 5, 6

[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. 1, 3, 5, 6

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[25] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 5

[26] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 3

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[28] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[29] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207, 2022. 3

[30] 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. 3

[31] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3

[32] 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. 3

[33] 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. 3

[34] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 3

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[37] 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. 3

[38] 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. 3

[39] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 250–266. Springer, 2022.

[40] 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, 3, 4

[41] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[43] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[44] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10-11):2586–2606, 2020. 5

[45] 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.

[46] 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, 3, 4, 5, 6, 7, 8

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012. 5

[50] 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.

[51] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[52] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image

generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021.

[53] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6

[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. 5

[55] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[56] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 3

[57] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[58] 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. 3, 5

[59] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 1, 3, 5, 7

[60] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022. 3

[61] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2018.

[62] 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

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

[64] 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. 3

[65] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 7

[66] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 5

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[68] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*, 2022. 3

[69] 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. 3, 7