BIVDiff: A Training-Free Framework for General-Purpose Video Synthesis via Bridging Image and Video Diffusion Models

Fengyuan Shi[1] Jiaxi Gu[2] Hang Xu[2] Songcen Xu[2] Wei Zhang[2] Limin Wang[1,3*]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2] Huawei Noah's Ark Lab [3] Shanghai AI Laboratory
https://bivdiff.github.io

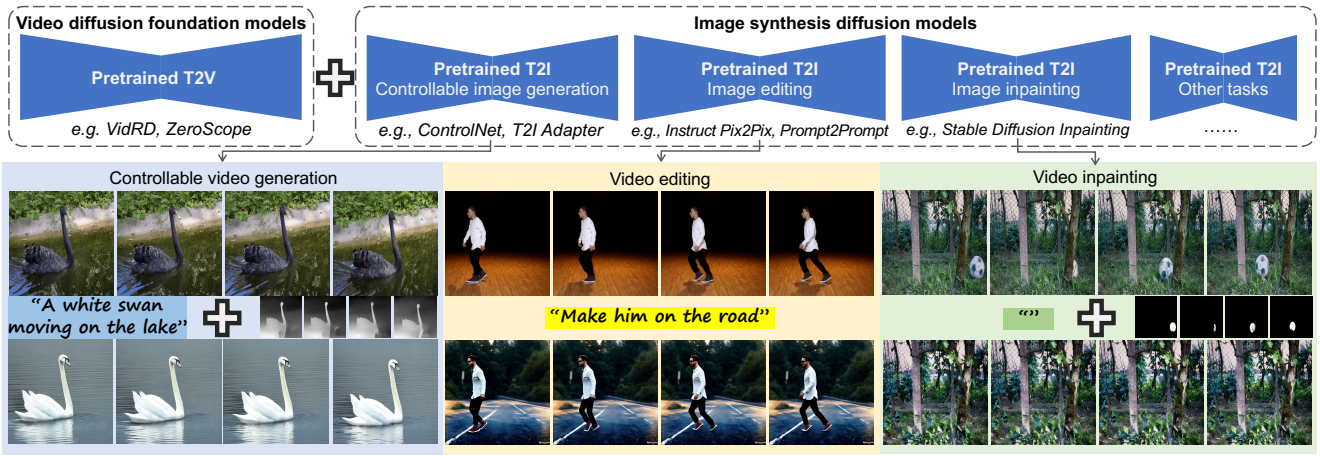


Figure 1. Given an image diffusion model (IDM) for a specific image synthesis task, and a text-to-video diffusion foundation model (VDM), our model can perform training-free video synthesis, by bridging IDM and VDM.

Abstract

Diffusion models have made tremendous progress in text-driven image and video generation. Now text-to-image foundation models are widely applied to various downstream image synthesis tasks, such as controllable image generation and image editing, while downstream video synthesis tasks are less explored for several reasons. First, it requires huge memory and computation overhead to train a video generation foundation model. Even with video foundation models, additional costly training is still required for downstream video synthesis tasks. Second, although some works extend image diffusion models into videos in a training-free manner, temporal consistency cannot be well preserved. Finally, these adaption methods are specifically designed for one task and fail to generalize to different tasks. To mitigate these issues, we propose a training-free general-purpose video synthesis framework, coined as **BIVDiff**, *via bridging specific image diffusion models and general text-to-video foundation diffusion models. Specifically, we first use a specific image diffusion model (e.g., ControlNet and Instruct Pix2Pix) for frame-wise video generation, then perform Mixed Inversion on the generated video, and finally input the inverted latents into the video diffusion models (e.g., VidRD and ZeroScope) for temporal smoothing. This decoupled framework enables flexible image model selection for different purposes with strong task generalization and high efficiency. To validate the effectiveness and general use of BIVDiff, we perform a wide range of video synthesis tasks, including controllable video generation, video editing, video inpainting, and outpainting.*

1. Introduction

Diffusion models [11, 29, 31] have shown impressive capabilities in generating diverse and photorealistic images. By scaling up dataset and model size, large-scale text-to-image diffusion models [4, 10, 20, 24, 25, 27] gain strong generalization ability and make tremendous breakthroughs in text-to-image generation. By fine-tuning these powerful image generation foundation models on high-quality data in specific areas, various downstream image synthesis tasks also come a long way, such as controllable image genera-

*Corresponding author (lmwang@nju.edu.cn).

tion [19, 37], image editing [2, 9, 17], personalized image generation [6, 26], and image inpainting [25]. However, video diffusion models are less explored for different video synthesis tasks due to several critical issues.

First, training video generation foundation models requires substantial training on a massive amount of labeled video data, heavily depending on a large scale of computing resources [7, 8, 12, 13, 28]. Even with video foundation models available, additional training on high-quality data in specific areas is still required for downstream video synthesis tasks such as controllable video generation [5, 32] and video editing [15, 18]. To improve training efficiency, Tune-A-Video [33] fine-tunes a pre-trained text-to-image model on the input video. Although Tune-A-Video can learn temporal consistency, this kind of per-input fine-tuning is still time-consuming. And it may overfit the small number of input videos and its generalization ability is limited (e.g., poor motion editability). Second, while some works extend image diffusion models into videos in a training-free manner, their temporal consistency cannot be well kept and flickers can still be observed (e.g., Fig. 6), due to the weak temporal modeling. Finally, previous works are usually proposed for one specific task and it requires different methods to extend from images to videos for different downstream video synthesis tasks with limited cross-task generality.

Image generation models can exhibit strong generalization and diversity, and yield many powerful downstream image synthesis models through fine-tuning. But frame-wise video generation with image models would lead to temporal inconsistency. Video generation foundation models can generate temporally coherent videos but require additional costly training for downstream video synthesis tasks. A question arises naturally: *Is it possible to build a training-free framework for general-purpose video synthesis by jointly leveraging the strengths of both pre-trained image and video diffusion models?* The key challenge is how to design a simple and general interface to bridge these two types of diffusion models to efficiently achieve temporal consistency in video synthesis.

To this end, we propose a general training-free video synthesis framework (BIVDiff), via bridging the *specific* image diffusion models and a *general* text-to-video diffusion model. Specifically, we first use a task-specific image diffusion model (like ControlNet [37], Instruct Pix2Pix [2]) to generate the target video in a frame-by-frame manner, then perform DDIM Inversion [30] on the generated video, and finally input the inverted latents into the video diffusion model (VDM) for temporal smoothing. Decoupling image and video models enables flexible model selection for different synthesis purposes, which endows the framework with strong task generalization and high efficiency (Fig. 1).

Despite using inverted latents by image DDIM Inversion, VDM tends to generate contents inconsistent with IDM in some cases, due to the distribution shifts. Moreover, for the case with a large gap between the latent distributions of image and video diffusion models, VDMs will fail to generate videos. For example, in the case of inputting source videos, the initial noisy frame latents obtained by frame-wise DDIM Inversion of image diffusion models are highly correlated, making some VDMs (e.g., VidRD [8]) with i.i.d. random latent requirement collapse to meaningless noises (Fig. 10). Accordingly, we introduce an improved version called Mixed Inversion. Specifically, we perform DDIM Inversion with both image and video diffusion models. Both latents by Image and Video DDIM Inversion encode the content of videos. The former could be further temporally smoothed by VDM but its distribution may be different from the one required by VDM. The latter cannot be further temporally smoothed by VDM but the distribution is consistent with VDM. We use a weighted sum of these two latents to adjust the distribution of initial latents fed into VDM. With this Mixed Inversion, we can flexibly adjust the latent distribution to make VDMs produce more consistent and better results, and trade off between temporal smoothing and open generation capability of VDMs. To validate the effectiveness of BIVDiff, we perform experiments on various representative video synthesis tasks, including 1) Controllable Video Generation; 2) Video Editing; and 3) Video Inpainting and Outpainting. Our contributions are summarized as follows:

- We propose a general training-free video synthesis framework, via bridging downstream task-specific image diffusion models and text-to-video diffusion models. Our BIVDiff is simple, efficient, and generalizable for different video synthesis tasks.
- We introduce Mixed Inversion, i.e., mixing the DDIM inverted latents of image and video diffusion models, to adjust the latent distribution to make VDMs produce more consistent and better results, and trade off between temporal smoothing and open generation capability of VDMs.
- We perform extensive experiments on various video synthesis tasks, including controllable video generation video editing, video inpainting, and outpainting, demonstrating the effectiveness and general use of BIVDiff.

2. Related Work

2.1. Diffusion Models for Image Synthesis

The emergence of diffusion models [11, 29, 31] has significantly advanced the progress of text-to-image generation. ADM [4] proposes classifier guidance for text-driven image generation. GLIDE [20] introduces classifier-free guidance [10] to improve image quality further. DALLE-2 [24] trains a prior model on CLIP text latents for better text-image alignment. Imagen [27] shows that text encoding with large language models (e.g., T5 [23]) is effective at

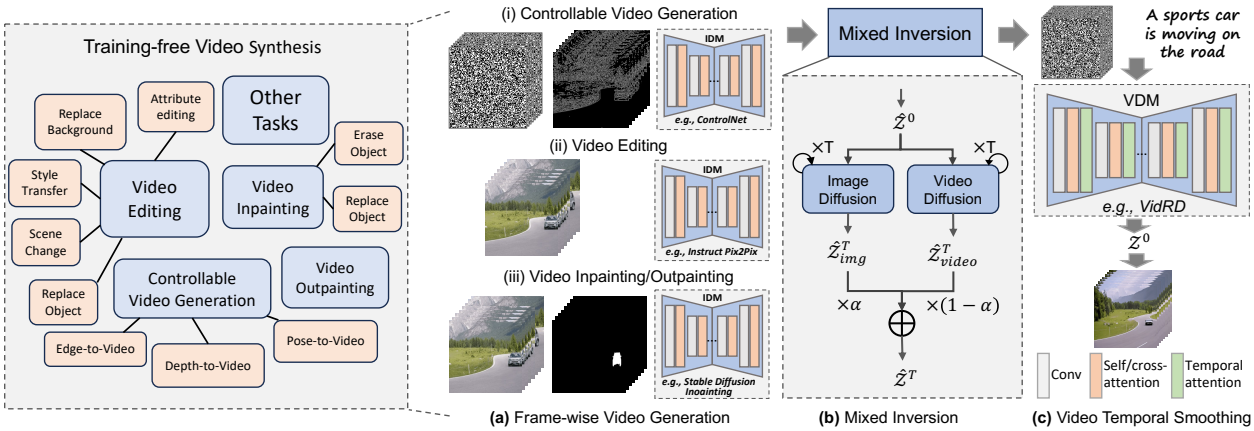


Figure 2. **BIVDiff pipeline.** Our framework consists of three components, including Frame-wise Video Generation, Mixed Inversion, and Video Temporal Smoothing. We first use the image diffusion model to do frame-wise video generation, then perform Mixed Inversion on the generated video, and finally input the inverted latents into the video diffusion model for video temporal smoothing.

image synthesis. Latent diffusion models (LDM) [25] perform diffusion and denoising processes in latent space, to increase training efficiency.

With the powerful pre-trained text-to-image diffusion foundation models, various downstream image synthesis tasks have also made great progress, such as controllable image generation [19, 37], image editing [2, 9, 17], personalized image generation [6, 26], image inpainting [25], etc. ControlNet [37] trains an auxiliary U-Net on image-control pairs to make models generate images conditioned on specific controls, such as depth, edge and human pose. Instruct Pix2Pix [2] is trained on generated training data to edit images from instructions. Textual Inversion [6] and DreamBooth [26] optimize a single word embedding using a few images of a user-provided concept for personalized image generation. Although effective, additional fine-tuning or optimization on input images is still required to transfer text-to-image foundation models into specific downstream image synthesis tasks, which is costly.

2.2. Diffusion Models for Video Synthesis

Inspired by text-to-image diffusion models [7, 8, 12, 13, 28], some works propose text-to-video diffusion models by adding extra temporal modules and train models on a large scale of video data. In addition to text-to-video generation, video diffusion models are also applied in various downstream video synthesis tasks, such as controllable video generation [3, 5, 32, 35] and video editing [15, 18].

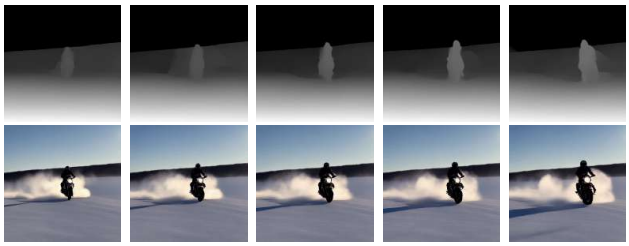
Training these video models is memory-hungry and computationally expensive. Some works attempt to adapt pre-trained image diffusion models to videos for efficient video synthesis. Tune-A-Video [33] adopts one-shot tuning on each input video for text-driven video editing. VideoP2P[16] is built on Tune-A-Video [33] and Prompt2Prompt [9], and introduce Null-Text Inversion [17] to improve the editing quality further. And there are also some training-free video synthesis methods, such as ControlVideo [38] and FateZero[21]. ControlVideo [38] proposes full-frame attention, i.e., concatenating all frames into a "big image" and performing self-attention on it, while Fate-Zero [21] fuses self-attention with a blending mask to ensure frame consistency. Although One-shot tuning and optimization make models generate high-fidelity videos, they suffer from poor generalization ability (e.g., cannot edit complex motion). Training-free adapting image diffusion models to videos provides an efficient solution to video synthesis, but they are less effective in maintaining cross-frame consistency at the level of texture and details [36], thus flickering artifacts are still severe.

Unlike previous works of adapting image models to videos by adding some modules or complex attention operations for a specific task, we present a simple method to bridge image and video models, and combine both advantages for training-free video synthesis. With a specific downstream image model (e.g., ControlNet [37]) and a general diffusion-based text-to-video foundation model (e.g., VidRD [8]), we can efficiently adapt to different video synthesis tasks (e.g., controllable video generation) in a training-free manner.

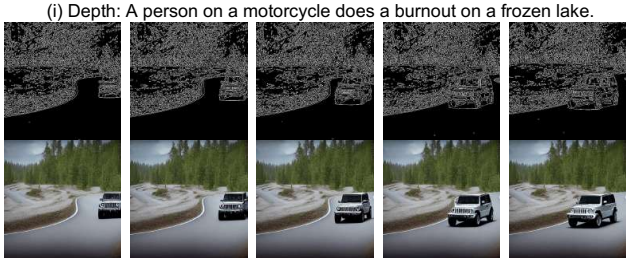
3. Method

Given a video synthesis task, we choose an image diffusion model **(IDM)** of its image task version and a text-to-video diffusion foundation model **(VDM)**. Let random latents $\mathcal{Z}^T = \{z_i^T\}_{i=1}^m$ or video $\mathcal{V} = \{v_i\}_{i=1}^m$ be the inputs, where T is the number of diffusion step, and m is frame number. Let $\mathcal{P}^*$ be the target prompt, and $\mathcal{C}$ be the conditions (e.g., depth maps and masks) according to target task. Our goal is to generate a temporally coherent video $\mathcal{V}^*$.

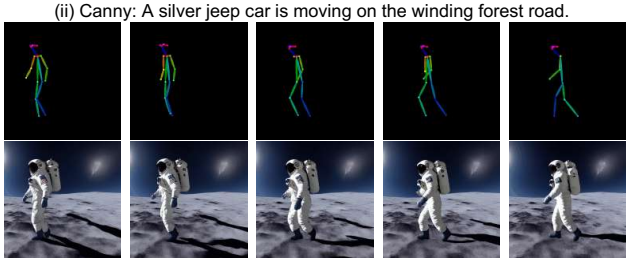
Our framework consists of three components, including Frame-wise Video Generation, Mixed Inversion, and Video



(i) Depth: A person on a motorcycle does a burnout on a frozen lake.



(ii) Canny: A silver jeep car is moving on the winding forest road.



(iii) Pose: An astronaut moonwalks on the moon.

Figure 3. Qualitative results of our proposed BIVDiff on controllable video generation task, conditioned on depth maps, canny edges and human pose sequence. We choose ControlNet [37] as our image diffusion model.

Temporal Smoothing. As shown in Fig. 2, we first use the image diffusion model to perform frame-wise video generation, then perform Mixed Inversion on the generated video, and finally input the inverted latents into the video diffusion model for video temporal smoothing.

3.1. Frame-wise Video Generation

The first step of our proposed framework is to perform frame-wise video generation with image diffusion models. For the given video synthesis task, we can choose an image counterpart. For instance, if we want to do controllable video generation, then we can use ControlNet [37] to generate the frames under the conditioning controls (e.g., edges, depth, etc.) independently. As for video editing, we can select one image editing model, such as Instruct Pix2Pix [2], to edit each frame in the video according to the target prompt independently. The generation process can be formulated as:

$$\hat{\mathcal{Z}}^0 = \{\hat{z}_i^0 = \text{IDM}(f_i, \mathcal{C})\}_{i=1}^m, \tag{1}$$

where f_i is the i-th random latent or frame in the given video, and $\mathcal{C}$ are conditions (e.g., text prompt, depth maps, and masks). Due to this decoupled design, our framework gains great flexibility and strong generalization ability. That is to say that we can choose arbitrary downstream image diffusion models for general-purpose video synthesis.

3.2. Mixed Inversion

The key of bridging image and video diffusion models is DDIM Inversion. After IDM denoising, we need to conduct DDIM Inversion to convert denoised latents to initial noisy latents as the input to the subsequent VDM. By DDIM Inversion, we can preserve the information that IDM generates, and make VDM synthesized videos consistent with the results of IDM but temporally coherent, instead of free generation. The frame-wise DDIM Inversion process can be formulated as:

$$\hat{\mathcal{Z}}^T = \{\hat{z}_i^T = \text{DDIM}_{\text{inv}}^{\text{img}}(\hat{z}_i^0)\}_{i=1}^m,$$

where $\text{DDIM}_{\text{inv}}^{\text{img}}$ means DDIM Inversion with image diffusion models. It is worth noting that we choose an image diffusion foundation model (e.g., Stable Diffusion [25]) for DDIM Inversion instead of the same model for frame-wise video generation, and the prompt is ϕ for DDIM Inversion.

Despite using inverted latents by image DDIM Inversion, VDM tends to generate content inconsistent with IDM in some cases, due to the distribution shifts. Moreover, when the gap between the latent distributions of image and video diffusion models is big, VDMs will fail to generate correct videos. For example, in the cases of inputting source videos, the initial noised latents of frames obtained by frame-wise DDIM Inversion with image diffusion models are highly correlated, making some VDMs (e.g., VidRD [8]) requiring i.i.d. random latents as inputs collapse and generate meaningless noises (Fig. 10).

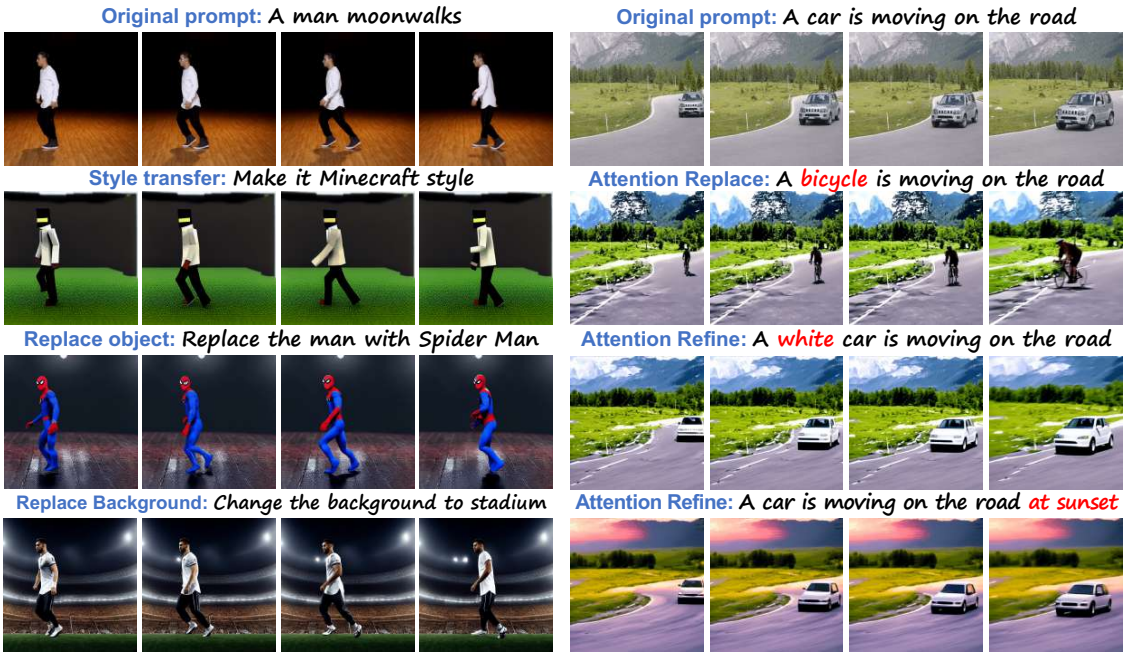
To solve these problems, we introduce Mixed Inversion. As shown in Fig. 2, we perform DDIM Inversion with both image and video diffusion models. Both latents by Image and Video DDIM Inversion keep the contents of videos. The former can be temporally smoothed by VDM but the distribution may be different from the distributions required by VDM. The latter cannot be further temporally smoothed by VDM distribution but the distribution is consistent with VDM. We can weighted-sum these two latents to adjust the distribution of initial latents fed into VDM. The latents mixing process is as follows:

$$\hat{\mathcal{Z}}_{img}^T = \{\hat{z_i^T} = \text{DDIM}_{\text{inv}}^{\text{img}}(\hat{z_i^0})\}_{i=1}^m, \tag{2}$$

$$\hat{\mathcal{Z}}_{video}^T = \text{DDIM}_{\text{inv}}^{\text{video}}(\hat{\mathcal{Z}}^0), \tag{3}$$

$$\hat{\mathcal{Z}}^T = \alpha \cdot \hat{\mathcal{Z}}_{img}^T + (1-\alpha) \cdot \hat{\mathcal{Z}}_{video}^T, \tag{4}$$

where $\text{DDIM}_{\text{inv}}^{\text{video}}$ means DDIM Inversion with our video diffusion model [8] and α is the mixing ratio used to adjust the ratio of the image and video latent components. With Mixed Inversion, we can adjust the latent distribution



(a) With Instruct Pix2Pix (b) With Prompt2Prompt

Figure 4. Qualitative results of our proposed BIVDiff on video editing task. We select two popular image editing methods, Instruct Pix2Pix [2] and Prompt2Prompt [9] as image models, and test a wide range of editing types.

to make VDMs produce correct results, and trade eoff between temporal smoothing and open generation capability of VDMs.

3.3. Video Temporal Smoothing

Although we can resort to image diffusion models for video synthesis tasks by frame-wise generation, temporal consistency is ignored, leading to visible flickers (e.g., Fig. 6). Video generation foundation models learn temporal consistency and can generate temporally coherent videos. Therefore, we do temporal smoothing on the video generated by IDM, by feeding the inverted latents into VDM. VDM can effectively capture the information stored in the inverted latents, and make the input videos consistent in temporal dimension, without destroying the contents created by IDM. The temporal smoothing process is formulated as:

$$\mathcal{Z}^0 = \mathrm{VDM}(\hat{\mathcal{Z}}^T, \mathcal{P}^*). \tag{5}$$

After temporal smoothing, we use vae decoder to decode the denoised latents to the target video.

4. Experiment

4.1. Implementation Details

To validate the effectiveness of our framework, we perform experiments on four representative video synthesis tasks, including 1) controllable video generation with ControlNet [37], 2) video editing with Instruct Pix2Pix [2] and Prompt2Prompt [9], 3) video inpainting with Stable Diffusion Inpainting [25] and 4) video outpainting with Stable Diffusion Inpainting [25]. For the video diffusion foundation model, we choose VidRD [8]. In the case of models for DDIM Inversion, we use Stable Diffusion 1.5 for frame-wise inversion, and VidRD for video-level inversion.

In our experiments, we generate 8 frames with 512 $\times$ 512 resolution for each video. The classifier-free guidance scale is 7.5 and the total timestep is 50. For the mixing ratio α in Mixed Inversion, we set 1.0, 1.0, 0.25, and 0.1 for BIVDiff with ControlNet, Instruct Pix2Pix, Prompt2Prompt and Stable Diffusion Inpainting as the default settings, respectively. And there is no per-video optimization (e.g. Null-text Inversion [17]) in our experiments.

4.2. Qualitative Results

Controllable Video Generation. By bridging pre-trained controllable image generation model ControlNet [37] and text-to-video foundation model VidRD [8], our framework BIVDiff supports zero-shot controllable video generation. Fig. 3 shows the generated videos conditioned on depth maps, canny edge maps, and human pose sequences. As shown in Fig. 3, the generated videos are well-matched with the conditions and keep significant temporal consistency, such as backgrounds, and both the appearance and structure of foreground objects.

Video Editing. Video editing is another important application in video synthesis. We choose two representative image editing models Instruct Pix2Pix [2] and Prompt2Prompt [9]

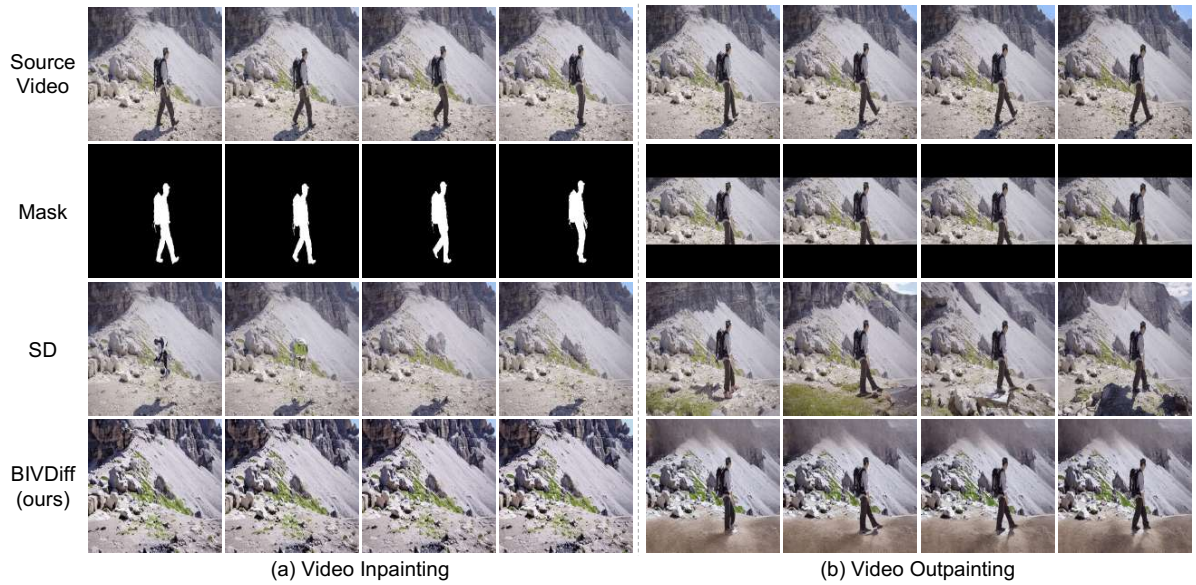


Figure 5. Qualitative results of our proposed BIVDiff on video inpainting and outpainting task. We adopt Stable Diffusion Inpainting [25] as our image model. Our method can erase objects and complete the masked regions well.

for zero-shot video editing. For video editing with Instruct Pix2Pix, we test various editing types, including style transfer, object replacement, and background replacement, as shown in Fig. 4 (a). As for Prompt2Prompt, we follow the paper to do attention replacement and attention refinement. As shown in Fig. 4 (b), our framework can replace the object, edit attribute, and do global editing, which are inherited from Prompt2Prompt in a zero-shot manner.

Video Inpainting and Outpainting. Additionally, we introduce an image inpainting model Stable Diffusion Inpainting [25] for video inpainting. For video outpainting, we can transfer the inpainting model to outpainting easily, by making masked regions of outpainting be the erased regions of inpainting. As shown in Fig. 5, independently processing each frame makes imperfect shadows that have not been completely erased and inconsistent areas to be filled in. We can eliminate these temporal inconsistencies by combining image and video diffusion models.

Additional Models. To further validate the effectiveness and general use of BIVDiff, we introduce more diffusion models, including another video diffusion model ZeroScope [1] and image diffusion model T2I-Adapter [19]. The qualitative results are in Supplementary Material.

4.3. Comparison with Baselines

We quantitatively and qualitatively compare our method with some baselines on controllable video generation (Text2Video-Zero [14], FateZero [21] and Tune-A-Video[38]) and video editing (Text2Video-Zero [14] and ControlVideo [14]). For quantitative comparison, we use DAVIS dataset in LOVEU-TGVE Benchmark [34], which consists of 16 videos and with 4 prompts per video, for automatic metrics and user study evaluation. Following Tune-A-Video [33], we adopt CLIP [22] to calculate frame consistency and textural alignment score. For user study, we follow Dreamix [18] to invite 25 human raters working on AI, arts and other areas, to rate videos by quality, fidelity, and alignment score on a scale of $1-5$. We also test the practical running time to compare inference speed.

Quantitative Comparison. Table 1 shows the quantitative results. For automatic metrics, our method has the best frame consistency due to the strong temporal modeling of VDM and comparable textual alignment. And our method is most favored by participants in the user study experiment since we can generate temporally coherent and realistic high-quality videos. Moreover, BIVDiff achieves a comparable inference speed in practice. Without modifying structures and inference pipelines inside IDM and VDM, we avoid time-consuming attention operations [21] or training [33] and benefit from parallel GPU computing.

Qualitative Comparison. We present visual comparisons in Fig. 6. Fig. 6(a) shows that ControlNet generates high-quality frames matched with controls (e.g., depth maps), but has severe frame inconsistency (e.g., the background is inconsistent across frames). Text2Video-Zero and ControlVideo generate temporally smooth videos, but there are still some slight flickers due to weak temporal modeling, and they struggle to accurately match the given controls (e.g., the lane lines disappear). In contrast, our method can generate temporally coherent videos well-matched with the conditions. Similar results can be found in video editing (Fig. 6(b)). Our method can keep more details in the input video (e.g., floor and shadows) and the generated videos are more realistic (e.g., the body of Spider Man).

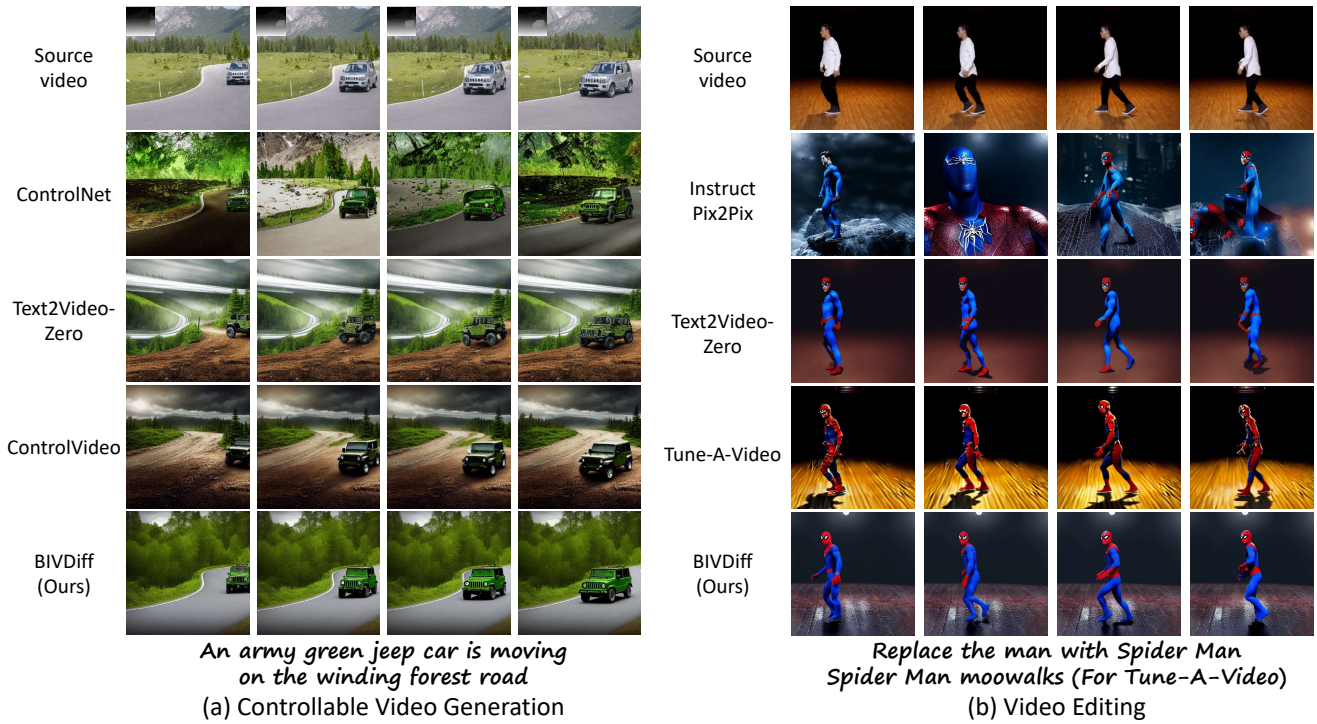


Figure 6. Qualitative comparison with baselines on controllable video generation and video editing task. Our BIVDiff generates high-quality and temporally coherent videos, and shows better (a) control and temporal consistency and (b) fidelity and realness.

Method	Automatic Metrics		User Study				Inference Time
	Frame Consistency	Textual Alignment	Quality	Alignment	Fidelity	Avg.	(per video)
Text2Video-Zero	91.69	**26.85**	2.74	3.16	**2.98**	2.96	**25s**
ControlVideo	92.63	26.12	2.61	3.12	2.54	2.76	57s
BIVDiff (Ours)	**92.67**	26.25	**3.38**	**3.24**	2.72	**3.11**	61s
Text2Video-Zero	91.57	25.37	2.26	2.23	2.46	2.32	**56s**
FateZero	90.75	26.42	2.38	1.7	**3.05**	2.38	221s
Tune-A-Video	90.46	**28.33**	2.30	2.23	2.35	2.29	11min + 26s
BIVDiff (Ours)	**93.50**	26.16	**2.98**	**2.30**	2.68	**2.65**	64s

Table 1. Quantitative comparison with baselines. The upper part is the result of controllable video generation with depth control. The bottom part is the result of video editing. Tune-A-Video adopts null-text inversion and one-shot tuning, while Text2Video-Zero and our BIVDiff are based on InstructPix2Pix and training-free. Our method has the best temporal consistency and is most favored by humans.

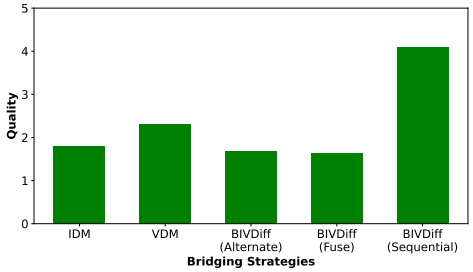


Figure 7. User study for bridging strategies ablation study.

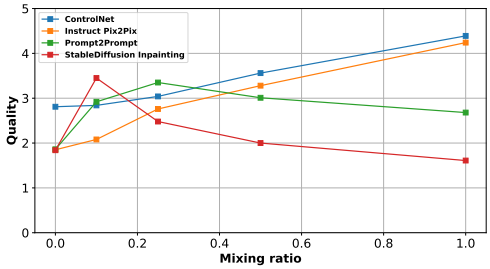


Figure 8. User study for mixing ratio ablation study.

4.4. Ablation Study

In this section, we study several key designs of our method, including the strategies of bridging image and video diffusion models, and the mixing ratio α in Mixed Inversion.

Ablation on bridging strategies. To validate the effectiveness of our bridging framework, we realize different strategies for comparisons, including 1) **IDM**. We use ControlNet [37] to do frame-wise video generation under the guidance of depth control. 2) **VDM**. We adopt VidRD [8] for text-to-video generation without depth control. 3) **IDM and VDM Alternate**. We use IDM and VDM for alternate denoising, i.e. one IDM denoising step by one VDM denoising step. 4) **IDM and VDM Fuse**. We perform IDM and VDM denoising simultaneously, and average these two latents. 5) **IDM and VDM Sequential**, i.e. our proposed

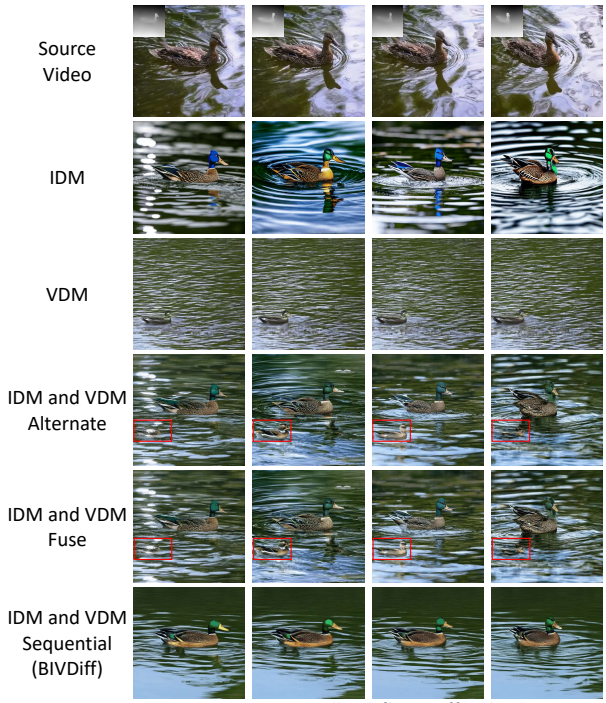


Figure 9. Ablation on strategies of bridging image and video diffusion models. Our sequential strategy can temporally smooth the videos (e.g., consistent appearance, structure, and background), and limit the open generation ability of VDM (e.g., the generated mallard of VDM is not in the final result of our method.)

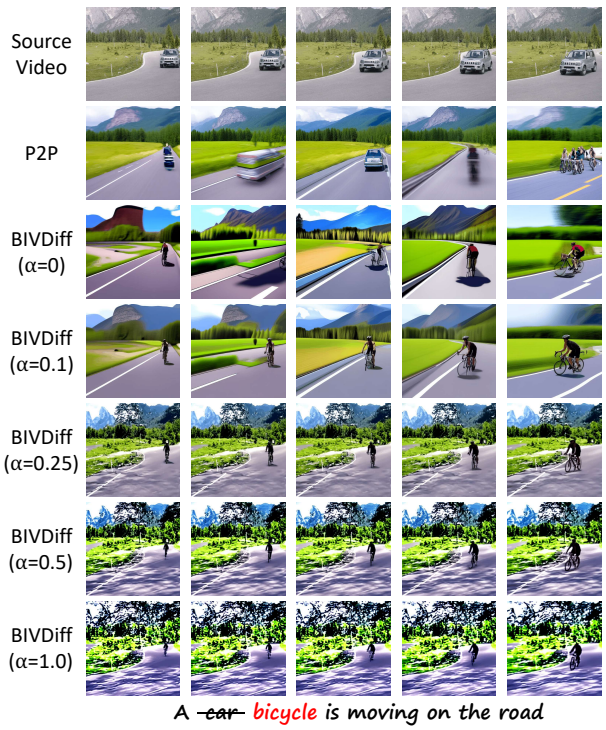


Figure 10. Ablation on the mixing ratio α in Mixed Inversion. Larger α leads to more temporally consistent videos, and smaller α makes the distribution of latents fed into VDM closer to VDM's and generates higher quality videos.

BIVDiff. The user study in Fig. 7 shows our sequential strategy works best. As shown in Fig. 9, videos generated by ControlNet are temporally inconsistent and VDM produces temporally consistent videos but unmatched with the given depth control. Bridging IDM and VDM during the denoising process ("Alternate" and "Fuse") tries to combine the results of IDM and VDM (e.g., there are two mallards in the videos). In contrast, our proposed BIVDiff bridges IDM and VDM in a sequential way, and generates temporally coherent videos consistent with depth control.

Ablation on mixing ratios. We also conduct an ablation study on video editing with Prompt2Prompt [9] to analyze the effects of mixing ratio. As shown in Fig. 10, the larger α is, the more temporally consistent the generated videos are. For example, there is a car and multiple bicycles that should not have appeared in the edited videos of Prompt2Prompt [9]. In contrast, videos generated by our method are more consistent with the input video and text prompt, and temporally coherent when α is 0.25. However, with α increasing, the quality of synthesized videos degrades and there are a lot of noises and artifacts in the videos. This is because the frames in the edited videos are similar (e.g., similar large areas of background) and latents by frame-wise DDIM Inversion with image diffusion models are highly correlated. When the video diffusion models, such as VidRD [8], require i.i.d. random latents as input, models will corrupt and produce noised videos. Fig. 8 shows video quality under different mixing ratios for each IDM and VDM pair. In practice, we can use small α to bridge latent distribution gaps and generate correct videos.

5. Conclusion

In this paper, we present a training-free framework for general-purpose video synthesis, coined as BIVDiff, via bridging downstream image diffusion models and text-to-video foundation diffusion models. We first use an image diffusion model (e.g., ControlNet [37]) for frame-wise video generation, then perform Mixed Inversion on the generated video, and finally input the inverted latents into the video diffusion model (e.g., ViDRD [8]) for temporal smoothing. We introduce Mixed Inversion to adjust the latent distribution to make VDMs produce correct results, and balance between temporal smoothing and open generation capability of VDMs. Extensive experiments on a wide range of video synthesis tasks demonstrate the effectiveness and generalization power of our method.

Acknowledgements. This work is supported by National Key R&D Program of China (No. 2022ZD0160900), National Natural Science Foundation of China (No. 62076119, No. 61921006)), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 6

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 3, 4, 5

[3] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 3

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 2, 3

[7] 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. 2, 3

[8] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 2, 3, 4, 5, 7, 8

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 2, 3, 5, 8

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 3

[14] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 6

[15] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 2, 3

[16] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3

[17] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3, 5

[18] 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. 2, 3, 6

[19] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3, 6

[20] 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

[21] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 3, 6

[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. 6

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 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*, 1 (2):3, 2022. 1, 2

[25] 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, 3, 4, 5, 6

[26] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine

tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[28] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 2, 3

[29] 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. 1, 2

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2

[32] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 3

[33] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3, 6

[34] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023. 6

[35] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 3

[36] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 4, 5, 7, 8

[38] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3, 6