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

Supplementary Material

1. More models

To further validate the effectiveness and general use of BIVDiff, we introduce more diffusion models into our proposed BIVDiff framework.

Additional Video Model. In addition to VidRD [3], we use another video diffusion foundation model ZeroScope [1] as our VDM to perform video temporal smoothing. Specifically, we perform controllable video generation with ControlNet [7] conditioned on depth maps, canny edge maps and human pose sequence, and video editing task with Instruct Pix2Pix [2]. As shown in Fig. 1 and Fig. 2, the generated videos keep temporal consistency well, demonstrating the flexity of model selections and general use of our BIVDiff framework.

Additional Image Model. We choose another popular controllable image generation model T2I-Adapter [5] as our IDM for controllable video generation, conditioned on depth maps. As shown in Fig. 3, the generated videos keep temporal consistency well and are consistent with the given controls, demonstrating the flexity of model selections and general use of our BIVDiff framework.

2. Effects of Video Temporal Smoothing

In Fig. 4, we show several cases to compare using image diffusion models (IDM) for frame-wise generation and our proposed BIVDiff which bridges image and video diffusion models, to validate the effectiveness of video temporal smoothing provided by VDM. Using IDM only produces temporally inconsistent videos for lacking temporal modeling. By bridging task-specific image models and video diffusion foundation models, we can produce temporally coherent videos (e.g., consistent appearance and structure of foreground objects and background across frames), while performing the target task well (e.g., well-matched with given controls for controllable video generation, and keeping good fidelity for video editing).

3. User Study Details

Following Dreamix [4], we invite 25 human raters working on AI, arts and other areas, to rate videos by quality, alignment, and fidelity on a scale of $1-5$ (1 is the lowest score and 5 is the highest score). The explanations of these metrics are as follows:

1. **Quality:** Rate the overall visual quality and smoothness of the edited video.
2. **Alignment:** How well does the edited video match the textual edit description provided?
3. **Fidelity:** How well does the edited video preserve unedited details of the original video?

For quantitative comparisons, we perform user study on DAVIS dataset in LOVEU-TGVE Benchmark [6], which consists of 16 videos and with 4 prompts per video. For ablation studies, we evaluate bridging strategies on four model pairs (ContorlNet and InstructPix2Pix as IDMs, and VidRD and ZeroScope as VDMs) with 10 videos and 16 text prompts. For mixing ratio, we use 8 videos, and test four IDMs (VidRD as VDM) with 5 prompts for each IDM.

4. Limitations

Fig. 5 shows a failure case of our method, where the edited video is inconsistent with the input video (i.e., low fidelity). This is due to the wrong editing results of Instruct Pix2Pix. When the results of frame-wise video generation with only image models are far away from expectations, our method may produce unsatisfied results. Luckily, due to the flexible image model selection brought by decoupling image and video models in our framework, we can tackle this problem by simply choosing another image diffusion model to generate correct results.

References

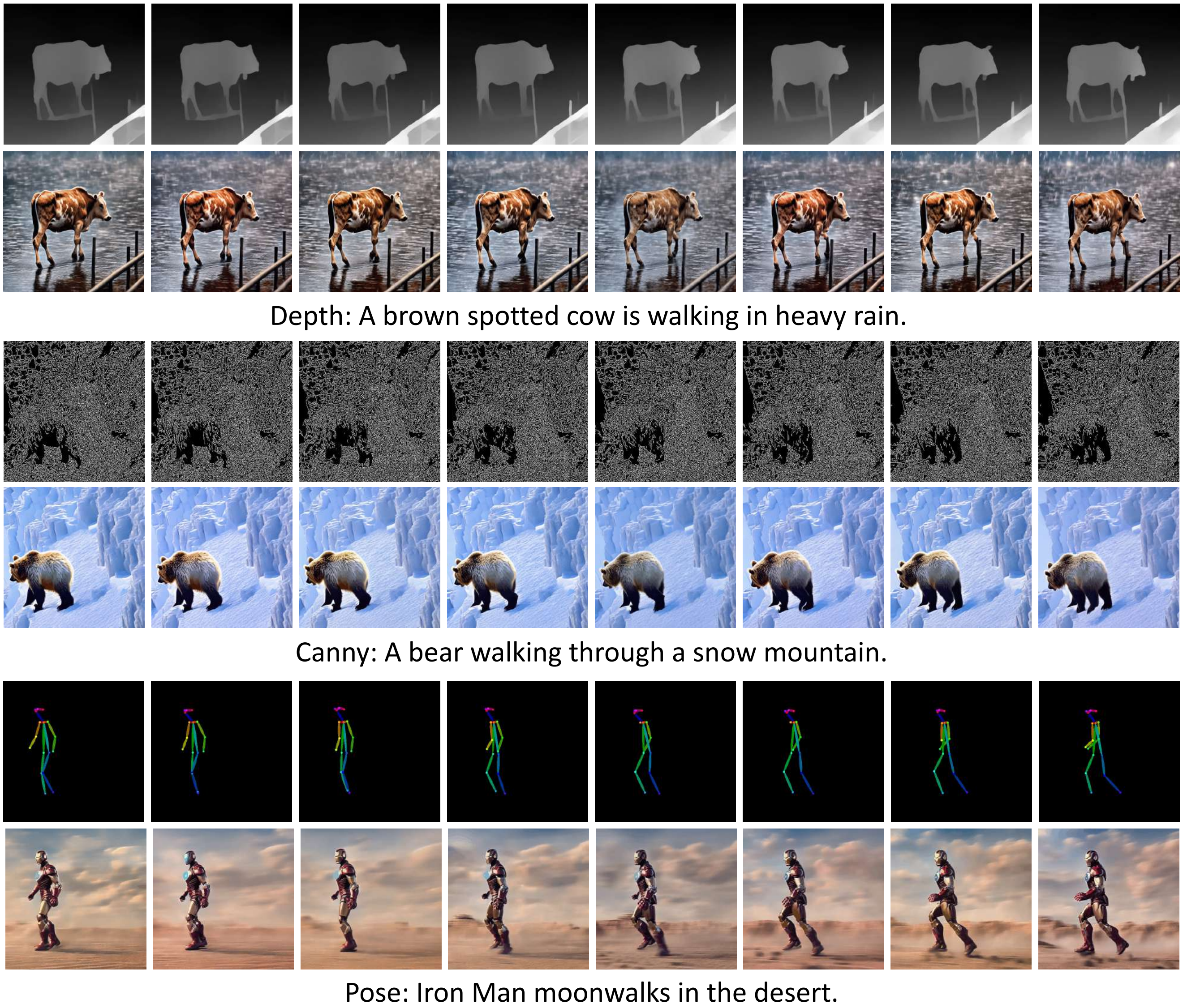
[1] Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 1

[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. 1, 3

[3] 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. 1, 3

[4] 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. 1

[5] 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. 1, 3



Depth: A brown spotted cow is walking in heavy rain.

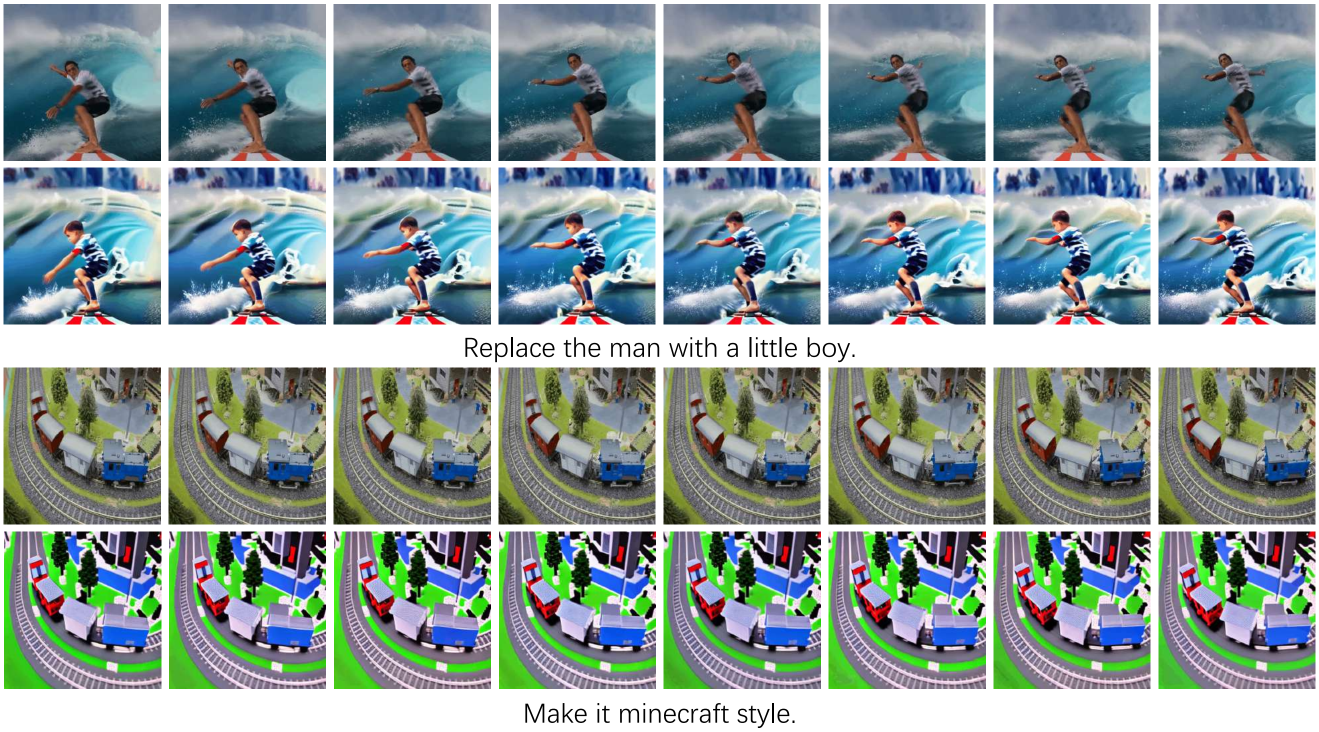
Canny: A bear walking through a snow mountain.

Pose: Iron Man moonwalks in the desert.

Figure 1. Qualitative results of our proposed BIVDiff on controllable video generation task, conditioned on depth maps, canny edges and human pose sequence. We choose ControlNet [7] as our image diffusion model and ZeroScope as our video diffusion model.

[6] 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. 1

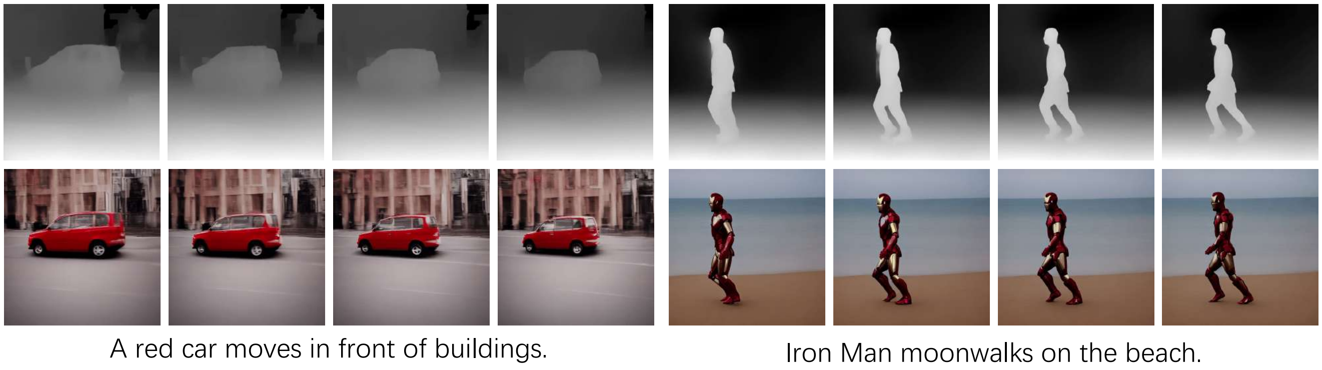
[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2



Replace the man with a little boy.

Make it minecraft style.

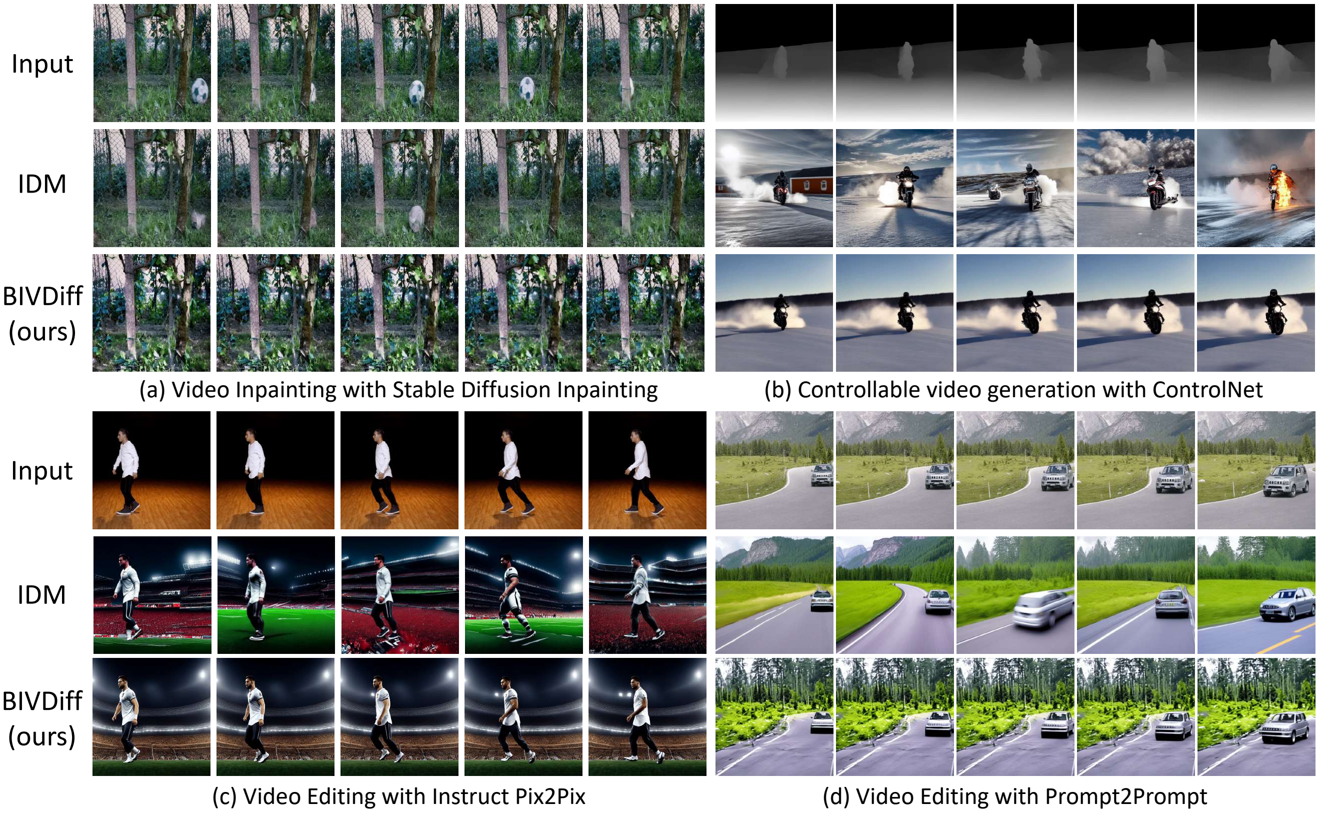
Figure 2. Qualitative results of our proposed BIVDiff on video editing task. We choose Instruct Pix2Pix [2] and ZeroScope as our video diffusion model.



A red car moves in front of buildings.

Iron Man moonwalks on the beach.

Figure 3. Qualitative results of our proposed BIVDiff on controllable video generation task, conditioned on depth maps. We choose T2I-Adapter [5] as our image diffusion model and VidRD [3] as our video diffusion model.



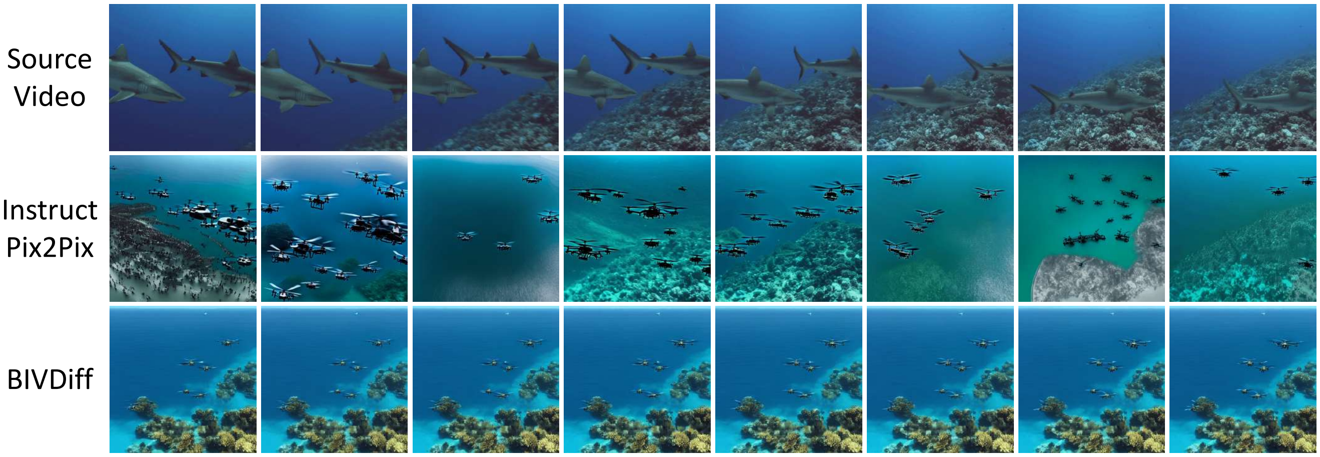
(a) Video Inpainting with Stable Diffusion Inpainting

(b) Controllable video generation with ControlNet

(c) Video Editing with Instruct Pix2Pix

(d) Video Editing with Prompt2Prompt

Figure 4. **Effects of Video Temporal Smoothing.** We compare IDM (using image models only) and our proposed BIVDiff (bridging image and video models), to validate the effectiveness of temporal smoothing power brought by VDM.



Replace sharks with quadrotor drones

Figure 5. **A failure case of video editing.** Our method may produce unsatisfied results when the results get by frame-wise video generation with only image models are far away from expectations.