Introduction to I2VGen-XL repository

Alibaba Group

1. General Introduction

We are excited to announce the **VGen** project, a holistic video generation ecosystem for video generation building on diffusion models, integrated into our I2VGen-XL repository (link). This repository serves as a comprehensive hub for academic research, offering a robust codebase centered around diffusion models for video generation. **VGen** is structured around three key domains: fundamental models, creative synthesis, and efficient synthesis, as depicted in Fig. 1. Our initial efforts encompass the open-sourcing of our essential work I2VGen-XL, including the pre-trained models and the code for its training and inference. This release aims to facilitate the entry of new researchers into the field, fostering the development of innovative algorithms within this field. Our ongoing efforts encompass the release of a suite of works foundational to diverse research areas, spanning image-to-video synthesis, text-to-video synthesis, video editing, customization, and learning from human feedback. In the preliminary version of **VGen**, we plan to include the following works:

I2VGen-XL [11]. This is a two-stage high-definition video generation model capable of producing 720P or higher resolution videos from a single static image, offering improved semantic and more realistic visual quality. Within this repository, we release a single-stage model capable of directly generating 720P videos while fully preserving all the details. This expansion will provide the academic community and diverse application fields with a foundational video generation model.

VideoComposer [6]. This algorithm enables us to control video generation in a compositional manner through textual, spatial, and temporal conditions, significantly enhancing the flexibility of video generation. Within this repository, we enhance the model performance in both quality and resolution by using the TF-T2V method mentioned below, with the aim of broadening its range of application scenarios.

TF-T2V [7]. The approach allows for scaling up video generation by utilizing readily available text-free videos. Consequently, we augment the video dataset by 10 million to train the fundamental text-to-video synthesis and compositional video synthesis paradigms, elevating video resolution and visual quality, ultimately leading to a significant performance enhancement. Here, you can access upgraded versions of the T2V and VideoComposer models. In this repository, you can access the upgraded versions of ModelScopeT2V [5] and VideoComposer by utilizing this scaling recipe.

InstructVideo [10]. The goal of InstructVideo is to instruct text-to-video diffusion models with human feedback through reward fine-tuning. This approach allows us to tailor the video diffusion models to human preferences, leading to significant enhancements in terms of visual quality and video-text alignment. To the best of our knowledge, this represents the first research effort to apply human feedback in the realm of video generation.

HiGen [3]. This approach enhances performance by decoupling the generation of the spatial and temporal elements of videos from two perspectives: the structural level and the content level. As a result, the model can generate videos with more precise semantics, improved quality, and seamless continuity from input texts. Furthermore, it effortlessly regulates the speed of motion and content changes using a single scalar.

DreamVideo [9]. The objective of this method caters to personalized video generation. It accomplishes this by employing two meticulously crafted adapters, enabling the creation of personalized videos from a small number of static images of the desired subject and a few videos of the target motion. The integration of these two lightweight and efficient adapters allows for versatile customization of any subject with any motion effortlessly.

VideoLCM [8]. The prohibitively large computation of video generation often impedes its broader application and research. In this repository, VideoLCM builds upon our video diffusion models and incorporates distillation techniques for training a latent consistency model [2] tailored for video generation. VideoLCM enables high-quality video generation in just a few steps. With this approach, we aim to provide a more flexible user experience and unlock additional creative possibilities across a broader spectrum of video creation scenarios.

2. Comparison with SVD

Currently, Stable Video Diffusion(SVD) [1] has made significant progress as a powerful open-source model for

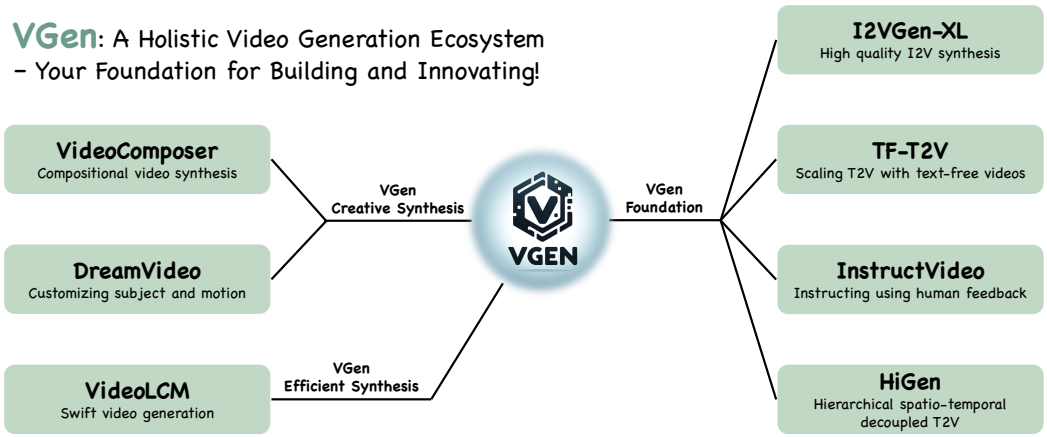


Figure 1. Main content of VGen project.

high-definition video generation. In this report, we will provide a detailed comparison between our I2VGen-XL and SVD for the image-to-video task, and we will reveal their differences using a modest number of samples.

Overall Comparison. Tab. 1 presents a comprehensive comparison of the two models, taking into account aspects such as the models, data, and the number of parameters. From this comparison, we can draw several conclusions: *i)* The main ideas are similar, such as framework design, training strategies, and data construction. To elaborate further, SVD integrates a noise-augmented version of the conditioning frame with the input of the UNet on a channel-wise basis. We simultaneously concatenate this local condition with the input and employ a global encoder to extract complementary features. These features are then fused with the original CLIP [4] image embedding and embedded into the UNet using cross-attention. Through this process, we find that the model can be more stable, leading to a significant reduction in video distortion. *ii)* Different inputs. In comparison to SVD, I2VGen-XL has the added capability of accepting text input, enabling our model to possess both text-to-video and image-to-video capabilities.

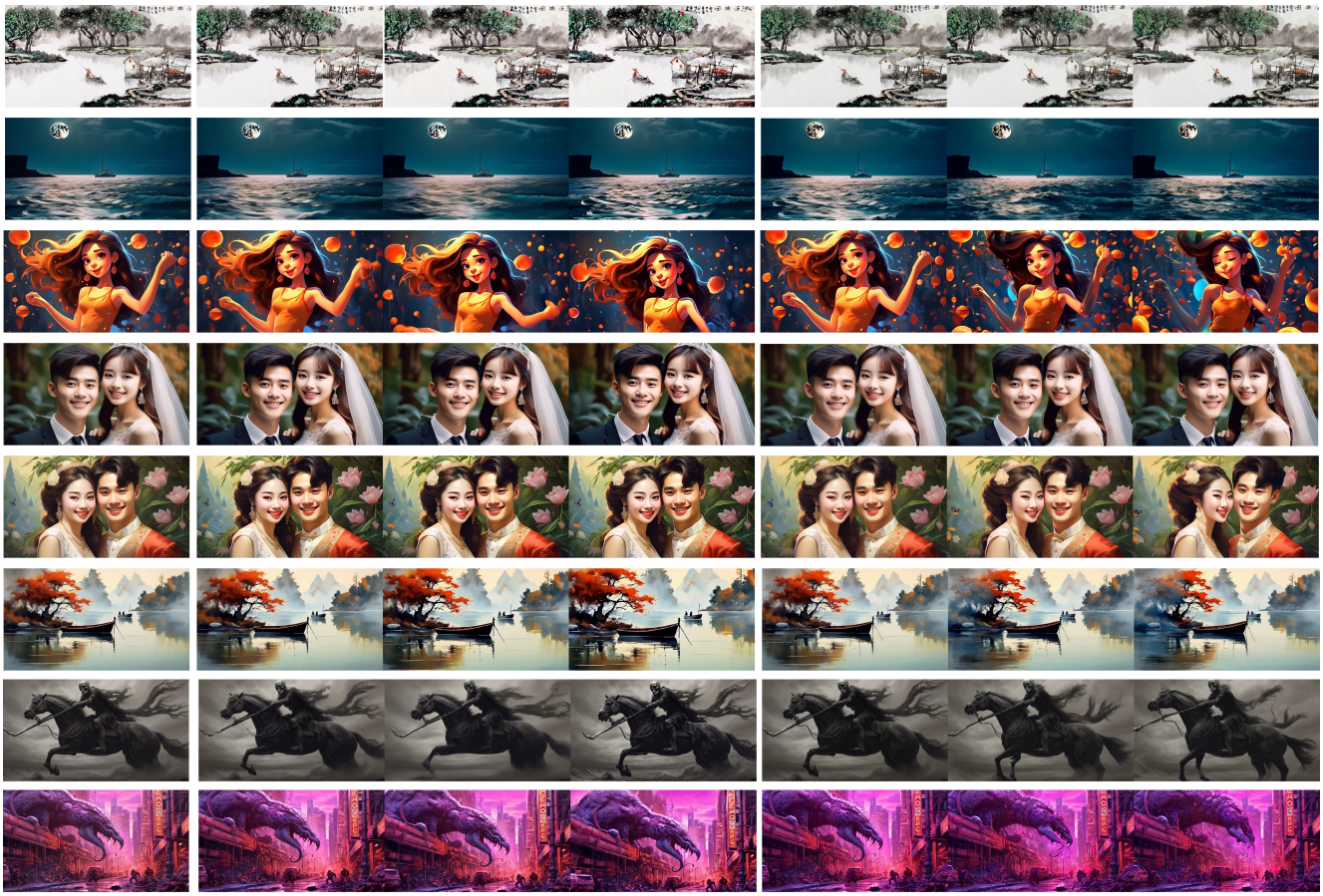
Qualitative Comparison. Fig. 2 illustrates the qualitative comparison between our approach and SVD. From these results, it can be seen that the motions generated by our method are more realistic, and the magnitude of the motions is larger. In comparison, the motions generated by SVD are more similar to a single image or a 3D object transformation. In our internal evaluation experiments, we discovered that 26% of the videos generated by SVD can be achieved through image transformation, such as translation, rotation, and so on. Furthermore, approximately 30% of the videos demonstrate a notable layered phenomenon, in which the foreground remains constant while only the background changes. In contrast, I2VGen-XL for these two aspects are 0% and 4%, respectively, which may be attributed to the different data distribution. Considering motion rationality from a statistical standpoint, I2VGen-XL has a clear advantage over SVD. We will also publish our more detailed quantitative results, which can better demonstrate our superiority over SVD. Consequently, we aspire for I2VGen-XL to become another effective foundational model accessible to the community and academic realm.

Items	SVD	I2VGen-XL
Framework	• Latent Diffusion • Global, Local Conditions	• Latent Diffusion • Global, Local Conditions
Data	• 600M Videos • With Motion Annotations	• 35M Videos +6B Images • With Category Annotations
Training	• Step1: Image pertaining • Step2: Video pertaining • Step3: High-Quality finetuning	• Step1: Image pertaining • Step2: Joing image and video pertaining • Step3: High-Quality finetuning
Inputs	• Image	• Image + Text
Resolutions	• 720P	• 720P
Frames	• 14 or 27	• 16 or 32 or 64
Parameter	• 1.5B	• 1.7B

Table 1. Overall comparison with SVD.

3. Conclusion

This report provides a brief overview of the **VGen** project, which showcases some of our latest developments in the field of video generation. It not only features robust foundational models in the T2V and I2V domains but also includes endeavors to enhance controllability, personalization, human preferences, and other areas. Moving



(a) Input Images (b) SVD (b) I2VGen-XL

Figure 2. Comparison of results between I2VGen-XL and SVD.

forward, we will persist in tackling the challenges present in current video generation and make these advancements available as open source. Furthermore, we have conducted a comparison between I2VGen-XL and one of the currently most powerful open-source models, SVD, highlighting the advantages and distinctive characteristics of our approach.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[2] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1

[3] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312*, 2023. 1

[4] 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 *ICML*, pages 8748–8763, 2021. 2

[5] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1

[6] 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. *NeurIPS*, 2023. 1

[7] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. *arXiv preprint arXiv:2312*, 2023. 1

[8] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312*, 2023. 1

[9] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream

videos with customized subject and motion. *arXiv preprint arXiv:2312*, 2023. 1

[10] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312*, 2023. 1

[11] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1