

VideoLCM: Video Latent Consistency Model

Xiang Wang^{1*} Shiwei Zhang^{2†} Han Zhang³ Yu Liu² Yingya Zhang² Changxin Gao¹ Nong Sang^{1†}

¹Key Laboratory of Image Processing and Intelligent Control,

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

²Alibaba Group ³Shanghai Jiao Tong University

{wxian, cgao, nsang}@hust.edu.cn, {zhangjin.zsw, ly103369, yingya.zyy}@alibaba-inc.com, hzhang9617@gmail.com

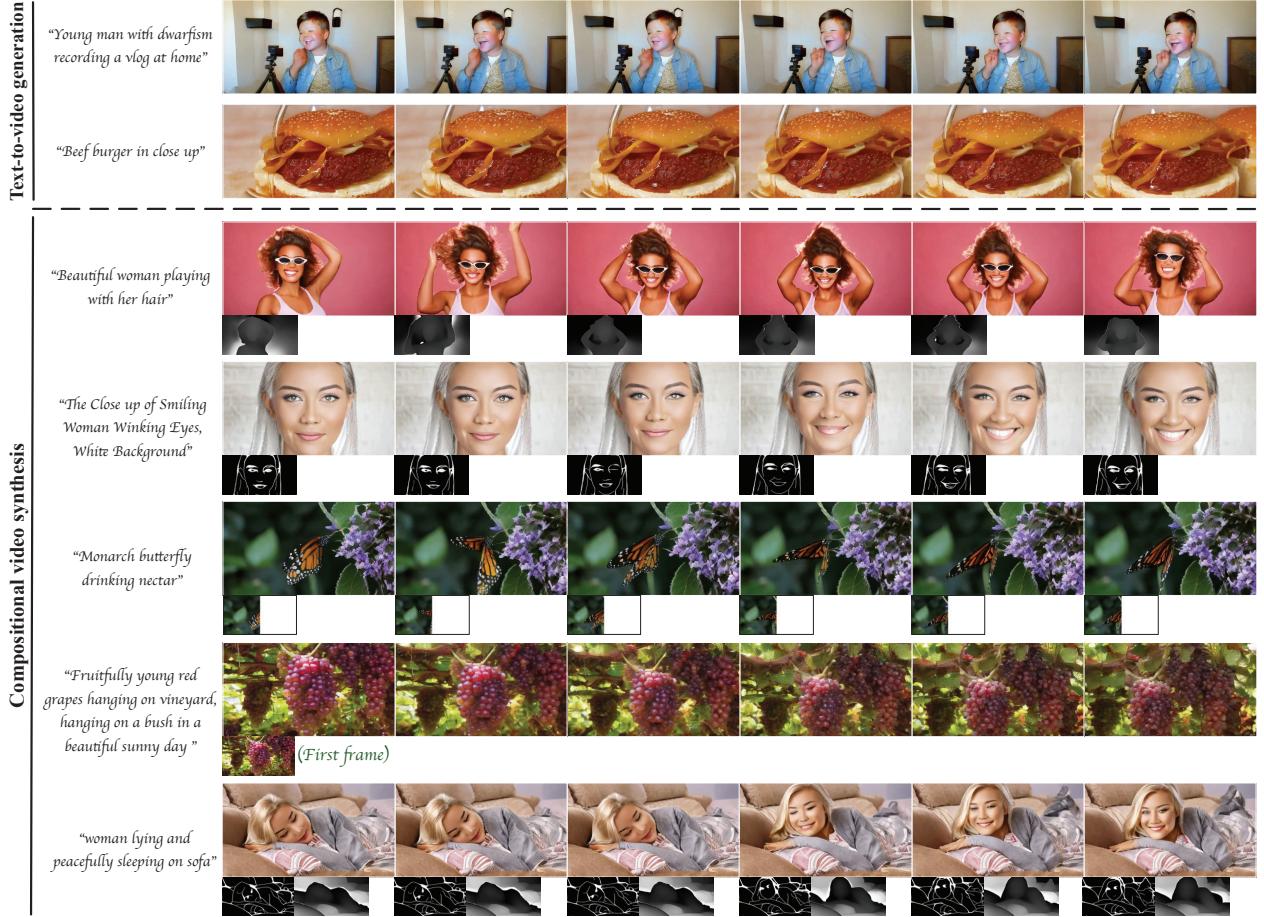


Figure 1. **Video examples** synthesized by the proposed VideoLCM with 4 inference steps. VideoLCM is a plug-and-play technique and can be integrated into text-to-video generation and compositional video synthesis paradigms.

Abstract

Consistency models have demonstrated powerful capability in efficient image generation and allowed synthesis within a few sampling steps, alleviating the high computational cost in diffusion models. However, the consistency

model in the more challenging and resource-consuming video generation is still less explored. In this report, we present the VideoLCM framework to fill this gap, which leverages the concept of consistency models from image generation to efficiently synthesize videos with minimal steps while maintaining high quality. VideoLCM builds upon existing latent video diffusion models and in-

* Intern at Alibaba Group. † Corresponding authors.

corporates consistency distillation techniques for training the latent consistency model. Experimental results reveal the effectiveness of our VideoLCM in terms of computational efficiency, fidelity and temporal consistency. Notably, VideoLCM achieves high-fidelity and smooth video synthesis with only four sampling steps, showcasing the potential for real-time synthesis. We hope that VideoLCM can serve as a simple yet effective baseline for subsequent research. The source code and models will be publicly available.

1. Introduction

Recently, the field of video generation has witnessed tremendous advancements in synthesizing photo-realistic and temporally coherent video content, especially with the development of diffusion models [4, 6, 14, 17, 38, 44, 46, 47, 54, 58, 60]. Traditional diffusion-based methods such as videoLDM [5], Make-A-Video [38] and ModelScopeT2V [38], have achieved significant performance by incorporating additional temporal layers into existing image diffusion models [30, 31] to handle the temporal dynamics in videos. Nevertheless, these diffusion-based approaches inevitably require substantial sampling steps to synthesize videos during inference, *e.g.*, 50-step DDIM sampling [40]. This limitation hinders the efficient and rapid synthesis of high-quality videos.

To address the challenge of high sampling cost in diffusion models, the concept of consistency models has been introduced in image generation [23, 24, 41, 53], achieving remarkable progress by enabling efficient image synthesis with a minimal number of steps (*e.g.*, 4 steps *vs.* 50 steps). Despite its success, the application of the consistency model in the domain of video synthesis still remains largely unexplored.

To fill this research gap, we propose the VideoLCM framework. Our method builds upon existing latent diffusion models in video generation and leverages the idea of consistency distillation to train a video latent consistency model. By incorporating the VideoLCM framework, we aim to alleviate the need for extensive sampling steps while maintaining high-quality video synthesis. The quantitative and qualitative results demonstrate the effectiveness of our approach. Remarkably, our method achieves high-fidelity video synthesis with only 4~6 sampling steps, showcasing its potential for fast and real-time synthesis. In comparison, previous methods such as ModelScopeT2V [44] and VideoLDM [5] typically require approximately 50 steps based on the DDIM solver to achieve similarly satisfactory results. In addition to text-to-video generation, we further extend the consistency model to compositional video synthesis. Experimental results indicate that in compositional video synthesis tasks, such as compositional depth-to-video synthesis, VideoLCM can produce visually satisfactory and

temporally continuous videos with even fewer steps, such as just 1 step.

In summary, the proposed VideoLCM bridges the gap between diffusion models and consistency models in video generation, enabling efficient synthesis of high-quality videos. By exploring the potential of consistency models in video generation, we aim to contribute to the field of fast video synthesis and provide a simplified and effective baseline for future research.

2. Related Work

The relevant fields related to this work include text-to-image generation, consistency model, and video generation. Next, we provide a brief review of these fields.

Text-to-image generation. In recent years, significant progress has been made in image generation with the development of generative models [7, 10, 13, 15, 19, 22, 34, 37], especially with the emergence of diffusion models [20, 27, 29, 31–33]. Typical methods for text-to-image generation, such as DALLE-2 [30], propose a two-stage approach where the input text is first converted into image embeddings using a prior model, followed by the generation of images based on these embeddings. Stable Diffusion [31] introduces a VAE-based approach in the latent space to decrease computational demand and optimizes the model with large-scale datasets [36]. Subsequent methods like ControlNet [57] and Composer [18] have incorporated additional conditional inputs, such as depth maps or sketches, for spatially controllable image synthesis.

Consistency model. To alleviate the limitation of requiring a large number of inference steps in diffusion models, the consistency model [41] has been developed. Built upon the probability flow ordinary differential equation (PF-ODE), consistency models learn to map any point at any time step to the starting of the trajectory, *i.e.*, the original clean image. The consistency model facilitates efficient one-step image generation without sacrificing the advantages of multi-step iterative sampling, thereby enabling more high-quality results through multi-step inference. Building upon the foundation of the consistency model, LCM [23] explores consistency models in the latent space to save memory consumption and improve inference efficiency. Subsequently, several methods [24, 35, 53] have also investigated efficient generation techniques and achieved impressive results. Inspired by the efficiency of the consistency model and its significant success in image generation, we extend the application of the consistency model to the domain of video generation.

Video generation. The pursuit of visually appealing and temporally coherent synthesis is crucial in video generation. Early methods [3, 39, 42, 48–50] primarily relied on generative adversarial networks (GANs), resulting in

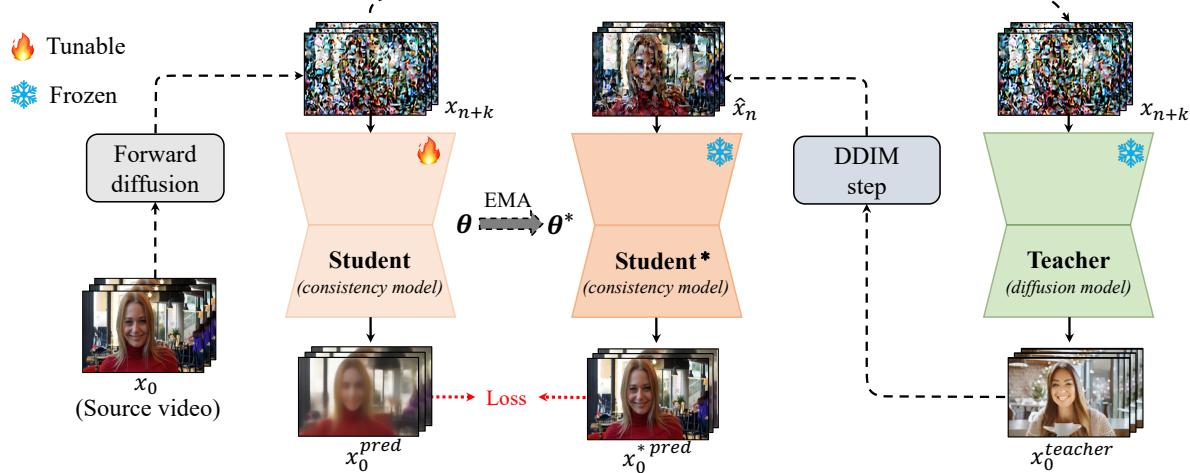


Figure 2. **The overall pipeline** of the proposed VideoLCM. Given a source video x_0 , a forward diffusion operation is first performed to add noise to the video. Then, the noised x_{n+k} is entered into the student and teacher model to predict videos. \hat{x}_n is estimated by the teacher model and fed into the EMA student model. To learn self-consistency, a loss is imposed to constrain the output of the two student models to be consistent. The whole consistency distillation is conducted in the latent space, and conditional guidance is omitted for ease of presentation. The teacher model is a video diffusion model, and the student model shares the same network structure as the teacher model and is initialized with the parameters of the teacher model.

poor video quality and limited generalization to unseen domains [5]. With the rise of diffusion models [31], which provide stable training and impressive results, recent approaches [1, 4, 6, 12, 14, 16, 21, 25, 26, 38, 44, 45, 51, 52, 55, 56, 58, 61] have started exploring diffusion-based video generation. To generate temporally consistent videos, mainstream text-to-video methods such as ModelScopeT2V [44] and VideoLDM [5] achieve long-term temporal modeling by inserting temporal modules into a 2D UNet and training on large-scale datasets [2]. There are also some methods [26, 38, 43, 58, 61] focusing on using super-resolution and frame interpolation models to generate more realistic videos. To achieve spatiotemporal controllability, controllable video synthesis methods [8, 9, 11, 28, 46, 59, 60] have been proposed. In particular, VideoComposer [46] presents a compositional video synthesis paradigm that enables flexible control over the generated videos through textual, spatial, and temporal conditions. Despite significant advancements, these methods rely heavily on extensive iterative denoising processes to obtain satisfactory results, posing challenges for fast video generation. To address this issue, in this work, we propose the VideoLCM framework based on consistency models for fast and efficient video synthesis.

3. Method

The proposed VideoLCM builds upon the foundations of latent consistency models. We first briefly describe the preliminaries about latent consistency models. Then, we will present the details of the proposed VideoLCM. The overall structure of VideoLCM is displayed in Fig. 2.

3.1. Preliminaries

To achieve fast image generation, song *et al.* [41] brings into the conception of the consistency model, which aims to optimize a model that learns to map any point at any time step to the starting of the PF-ODE trajectory. Formally, the self-consistency property can be formulated as:

$$f_\theta(x_t, t) = f_\theta(x_{t'}, t'), \forall t, t' \in [\epsilon, T] \quad (1)$$

where ϵ is a time step, T is the overall denoising step, and x_t denotes the noised input.

To accelerate the training and extract the strong prior knowledge of the established diffusion model [31], consistency distillation is usually adopted:

$$\mathcal{L}(\theta, \theta^*; \Phi) = \mathbb{E}[d(f_\theta(x_{t_{n+1}}, t_{n+1}), f_{\theta^*}(\hat{x}_{t_n}, t_n))] \quad (2)$$

where Φ means the applied ODE solver and the model parameter θ^* is obtained from the exponential moving average (EMA) of θ . \hat{x}_{t_n} is the estimation of x_{t_n} :

$$\hat{x}_{t_n} \leftarrow x_{t_{n+1}} + (t_n - t_{n-1})\Phi(x_{t_{n+1}}, t_{n+1}) \quad (3)$$

LCM [23] conducts the above consistency optimization in the latent space and applies classifier-free guidance [15] in Eq. (3) to inject control signals, such as textual prompts. For more details, please refer to the original works [23, 41].

3.2. VideoLCM

Following LCM, the proposed VideoLCM is also established in the latent space to reduce the computational burden. To leverage the powerful knowledge within large-scale

pretrained video diffusion models and speed up the training process, we apply the consistency distillation strategy to optimize our model. Note that the pretrained video diffusion model can be the text-to-video generation model (*e.g.*, ModelScopeT2V [44]) or the compositional video synthesis model (*e.g.*, VideoComposer [46]).

In `VideoLCM`, we apply DDIM [40] as the basic ODE solver Ψ to estimate \hat{x}_{t_n} :

$$\hat{x}_{t_n} \approx x_{t_{n+1}} + \Psi(x_{t_{n+1}}, t_{n+1}, t_n, c) \quad (4)$$

where c means the conditional inputs, which can be textual prompts in text-to-video generation or multiple combined signals in compositional video synthesis task.

Since classifier-free guidance is pivotal in synthesizing high-quality content [15], we also leverage classifier-free guidance in the consistency distillation stage and use a factor w to control the guidance scale:

$$\begin{aligned} \hat{x}_{t_n} \approx & x_{t_{n+1}} + (1+w)\Psi(x_{t_{n+1}}, t_{n+1}, t_n, c) \\ & - w\Psi(x_{t_{n+1}}, t_{n+1}, t_n, \phi) \end{aligned} \quad (5)$$

In LCM [23], w is variable and can be fed into the network for modulation, but this changes the structure of the initial network because a module encoding w needs to be added. In order to keep the initial parameters and design of the consistency model consistent with the teacher diffusion model, we train the consistency model with a fixed value of w , such as 9.0. Note that classifier-free guidance is only applied to the teacher diffusion model in training and is not required during the inference process of the consistency model.

`VideoLCM` is a plug-and-play technique compatible with text-to-video generation and compositional video synthesis. During the inference phrase, we can sample 4~6 LCM steps to produce plausible results on text-to-video generation. For compositional video synthesis, take the compositional depth-to-video task as an example, 2~4 steps are sufficient, and sometimes even 1 step.

4. Experiments

In this section, we first introduce the details of the experimental setup. Then, quantitative and qualitative comparisons will be represented to evaluate the effectiveness of the proposed `VideoLCM` framework.

4.1. Experimental setup

Datasets. We train our video consistency model on two widely used datasets, *i.e.*, WebVid10M [2] and LAION-5B [36]. WebVid10M is a video-text dataset containing approximately 10.7M videos. To further enhance temporal diversity and improve visual quality, we additionally utilize about 1M internal video-text data to train `VideoLCM`. LAION-5B is an image-text dataset that is used to provide high-quality visual-text correspondence.

Table 1. **Inference latency** on text-to-video generation task. All experiments are performed on an NVIDIA A100 GPU. The inference overhead of generating eight videos at a time is reported.

Method	Step	Resolution	Latency
Baseline	DDIM 50-step	$16 \times 256 \times 256$	60s
<code>VideoLCM</code>	LCM 4-step	$16 \times 256 \times 256$	10s
Baseline	DDIM 50-step	$16 \times 448 \times 256$	104s
<code>VideoLCM</code>	LCM 4-step	$16 \times 448 \times 256$	16s

Implementation details. For convenience, we directly leverage the existing pretrained video diffusion models in TF-T2V [47] as the teacher model and fix the network parameters in the consistency distillation process. TF-T2V [47] is a technique that exploits text-free videos to scale up video diffusion models and can be applied to mainstream text-to-video generation and compositional video synthesis framework. The model structure of the consistency model is the same as the teacher diffusion model and is initialized with the teacher’s model parameters. AdamW optimizer with a learning rate of 1e-5 is adopted to train `VideoLCM`. The EMA rate used in our experiments is 0.95. We sample 16 frames and crop a 448×256 center region from each source video as model input. The training loss used in `VideoLCM` is a Huber loss by default. The entire network is trained on 4 NVIDIA A100 GPUs (one for image and three for video), requiring approximately 4k training iterations to produce relatively reliable video generation results.

4.2. Time efficiency

We measure the inference time required for text-to-video synthesis using our proposed `VideoLCM` and compare it with the baseline method [47]. The comparative results are exhibited in Tab. 1. From the results, we can observe that since our method merely requires 4 LCM steps for inference, it is much faster than the baseline method with 50 DDIM steps. It should be noted that in addition to iterative denoising, the inference cost also includes text feature encoding, latent code decoding, *etc.* In addition, we notice that `VideoLCM` saves more time on high-resolution generation compared to baseline. For example, generating a video of $16 \times 256 \times 256$ size saves 50 seconds (*i.e.*, 60s – 10s), while $16 \times 448 \times 256$ saves 88 seconds. The above comparison demonstrates the high efficiency of our approach.

4.3. Ablation study on inference steps

In Fig. 3, we present the experimental visualization of varying inference steps in the text-to-video task. The results indicate that when the sampling step is too small, such as step = 1, the generated videos suffer from blurriness, with many details being inaccurately represented, and the temporal structure of objects cannot be preserved. As the number of iteration steps increases, the visual quality gradually

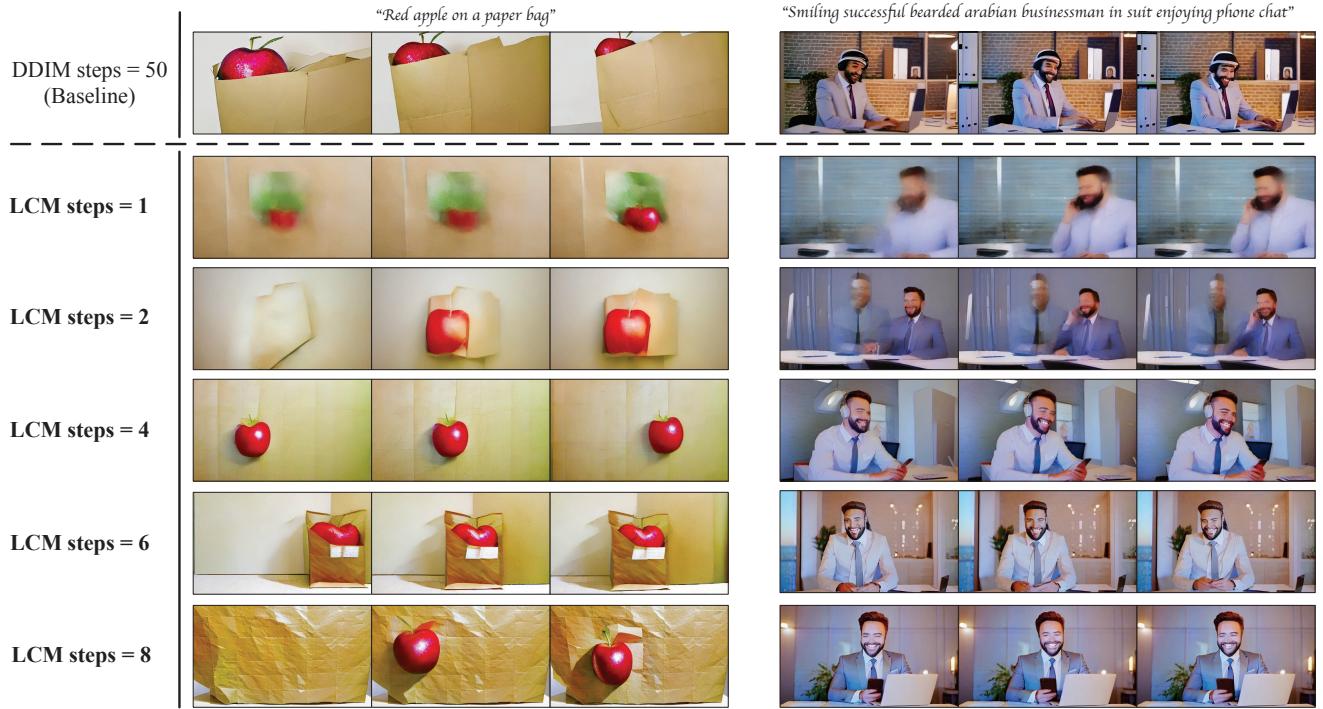


Figure 3. **Ablation study on text-to-video task under varying steps.** Larger steps generally yield better visual quality and time continuity.

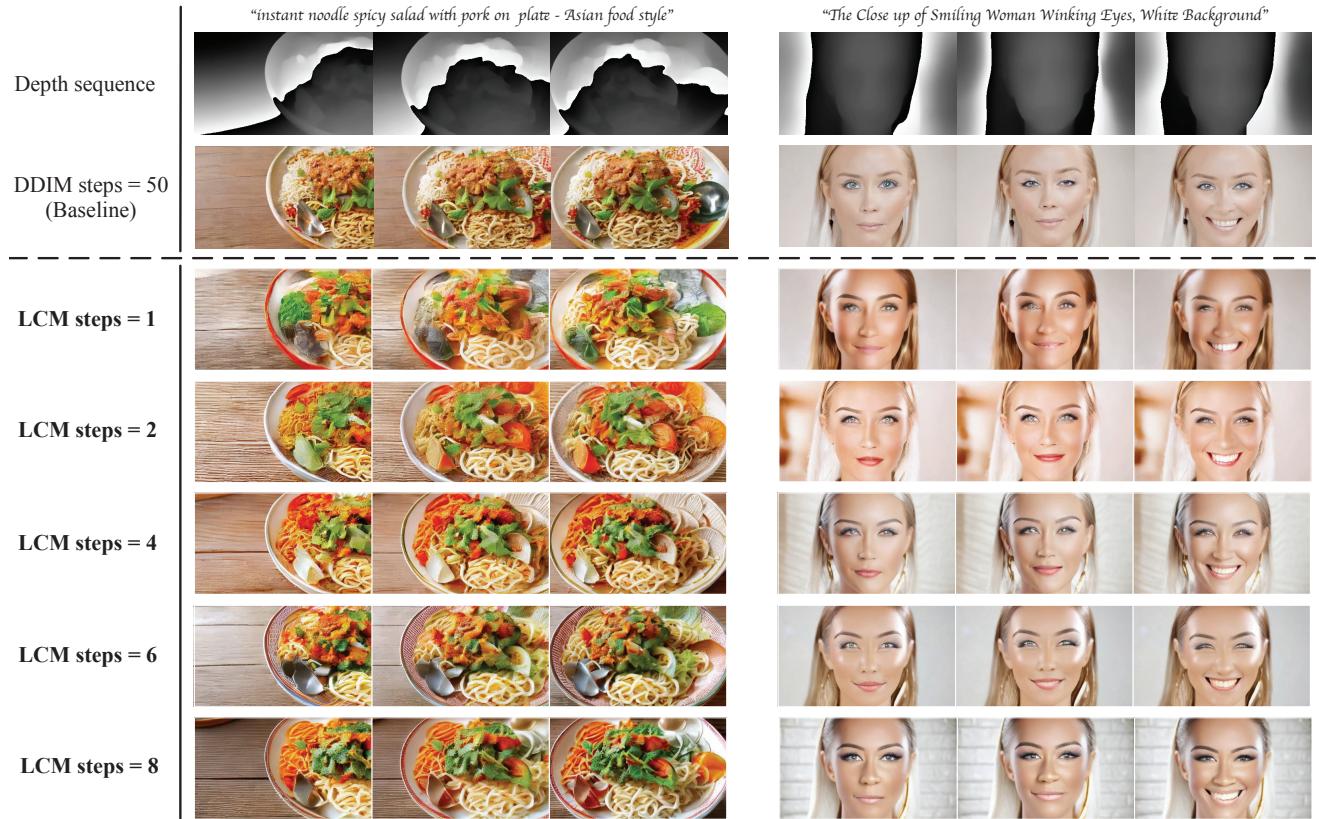


Figure 4. **Ablation study on compositional depth-to-video synthesis task under different inference steps.** Since the additional depth sequence can provide prior guidance about structure and temporal, our method can produce plausible results with fewer steps or even only one step.

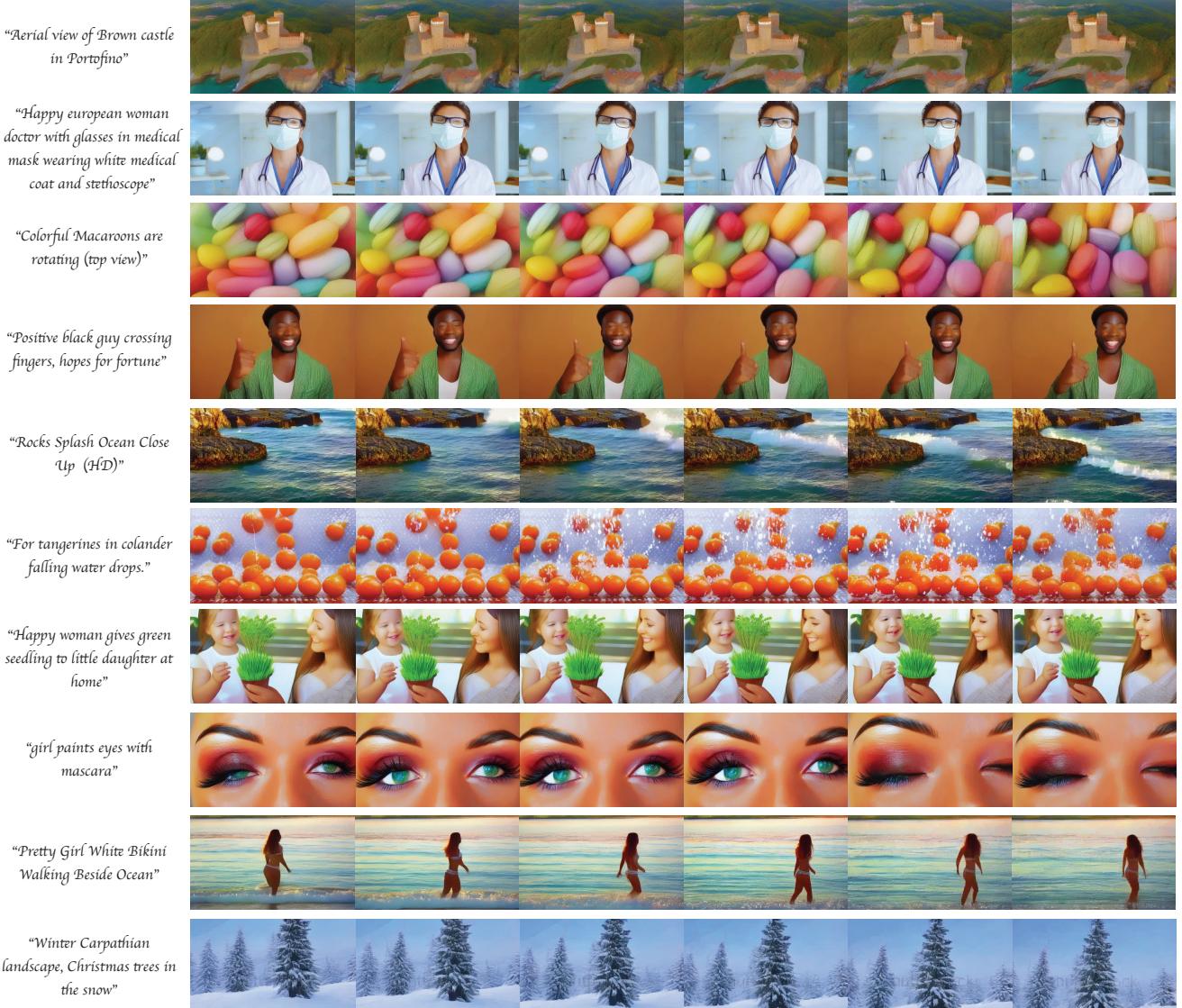


Figure 5. Qualitative visualization results on text-to-video generation task. Videos are synthesized by performing 4 denoising steps.

improves, and the temporal structure can be better maintained. For example, when using 4~6 steps, comparable results to DDIM’s 50 steps can be achieved, while significantly reducing the sampling steps and improving the generation speed. Moreover, we additionally perform an ablation study on compositional depth-to-video synthesis. As illustrated in Fig. 4, we observe that with only a few steps, such as 1 step, the generated results can display good visual quality. With an increase in the step size, many details became even more apparent. We attribute this phenomenon to the fact that compared to text-to-video generation, compositional depth-to-video synthesis can leverage additional structured control information, which reduces the difficulty of predicting videos from pure noise and enables achieving high-quality results with fewer steps. Empirically, we find

that 2~4 steps can produce relatively stable compositional synthesis contents. The results on both text-to-video and compositional video generation demonstrate the effectiveness of the proposed **VideOLCM**, achieving a good balance between quality and speed.

4.4. More qualitative results

To comprehensively showcase the capabilities of our method in handling various video generation tasks, we present additional visualization examples in Fig. 5. These results are generated with 4 LCM steps. From the visualization results, it is evident that our method achieves impressive generation outcomes in terms of visual quality and temporal continuity.

In Figs. 6 to 8, we present more results of compositional

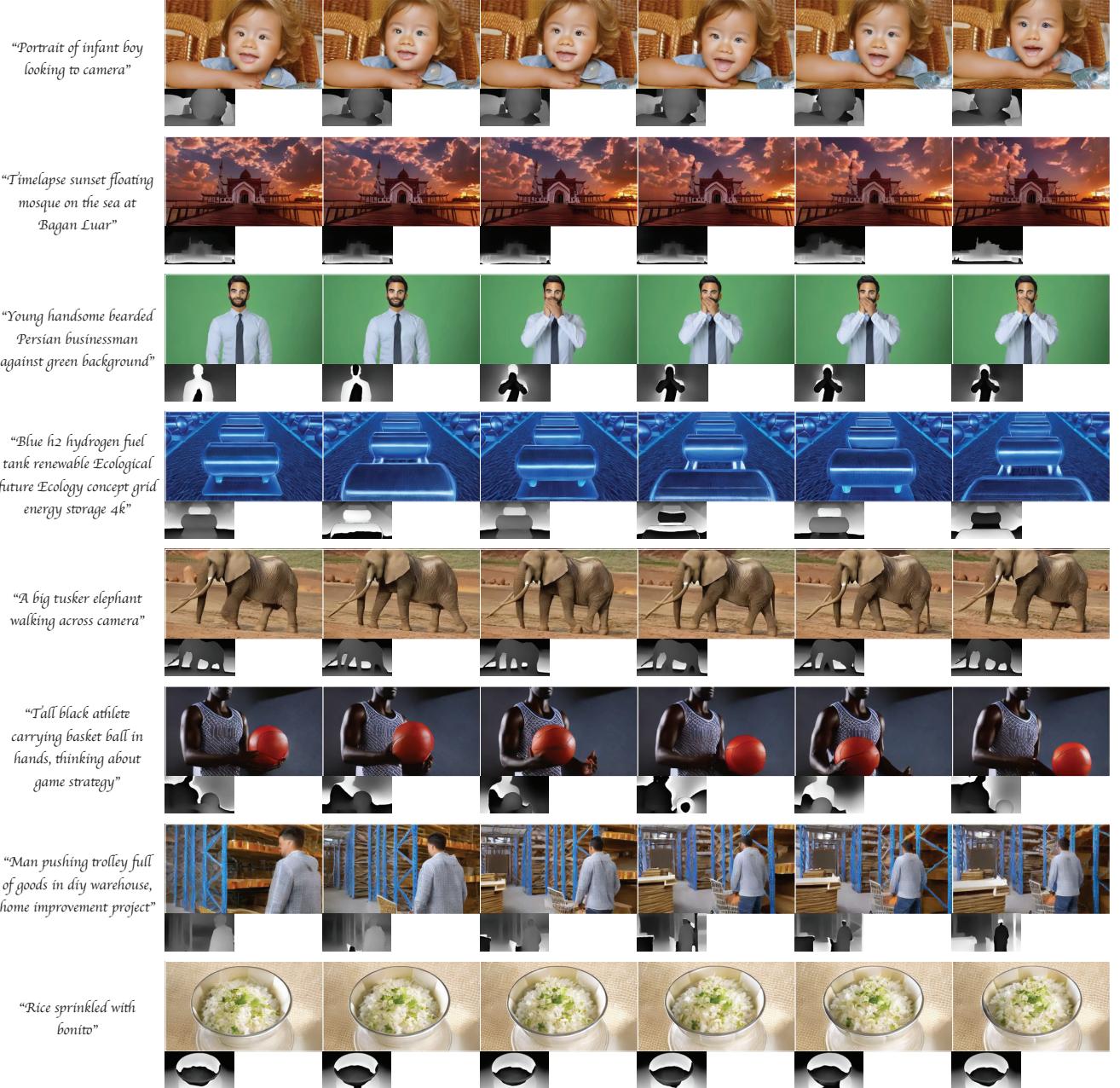


Figure 6. Qualitative results on compositional depth-to-video synthesis task. Videos are synthesized by performing 4 denoising steps.

video generation using 4 sampling steps, including compositional depth-to-video synthesis (Fig. 6), compositional sketch-to-video synthesis (Fig. 7), and compositional video inpainting (Fig. 8). These qualitative results demonstrate stable generation quality and highlight the controllability of our method in accordance with the input conditioning. Our method can be applied to multiple mainstream video generation tasks mentioned above, revealing the generality of the VideoLCM framework and the vast potential for other future applications.

4.5. Limitations

In our VideoLCM, we explore the application of the consistency model in video generation tasks, including text-to-video generation and compositional video synthesis. However, there are certain limitations: 1) Our method relies on a strong teacher model as the distillation target. 2) The consistency distillation process requires finetuning the model. While consistency distillation only requires a small number of training steps, it may lead to unsatisfactory results when the training data for the teacher model is unavail-

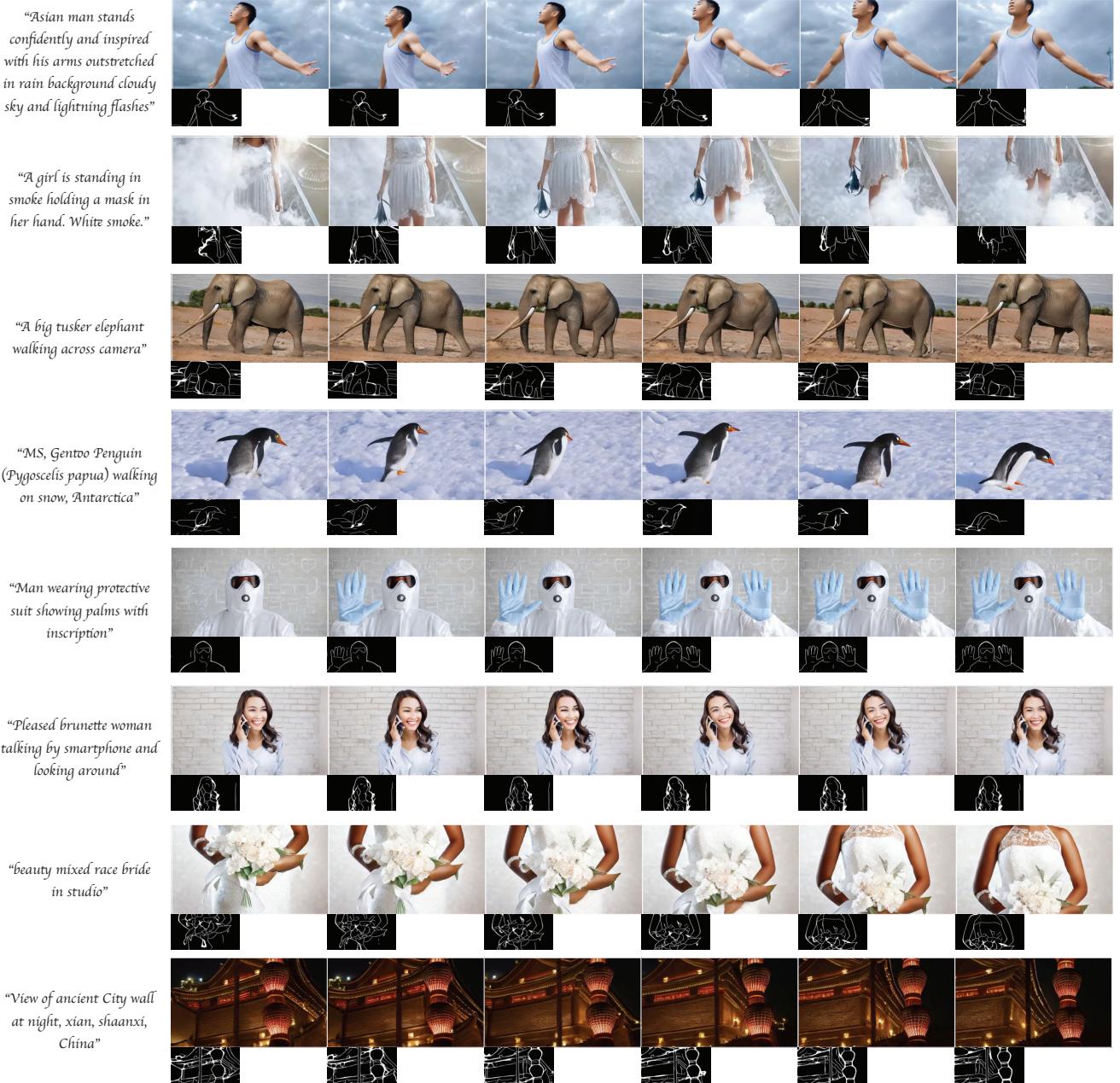


Figure 7. Qualitative results on compositional sketch-to-video synthesis task. Videos are synthesized by performing 4 denoising steps.

able or from different domains. 3) Even though our method reduces the inference steps to 4~6, real-time video generation, like image generation, is still not achieved. Exploring more stable and efficient video generation algorithms while ensuring the high quality of the generated video content is a promising future direction.

5. Conclusion

In this work, we propose the VideoLCM framework that extends latent consistency models to the video gener-

ation field. Our approach leverages existing latent video diffusion models and employs the consistency distillation technique to enable efficient and fast video synthesis. Experimental results demonstrate the effectiveness of our approach, with high-fidelity video synthesis achieved in just four steps, showcasing its real-time synthesis potential compared to prior methods requiring approximately 50 DDIM steps. We hope that VideoLCM can serve as a simple yet effective baseline for subsequent research work.

Acknowledgements. This work is supported by the Na-

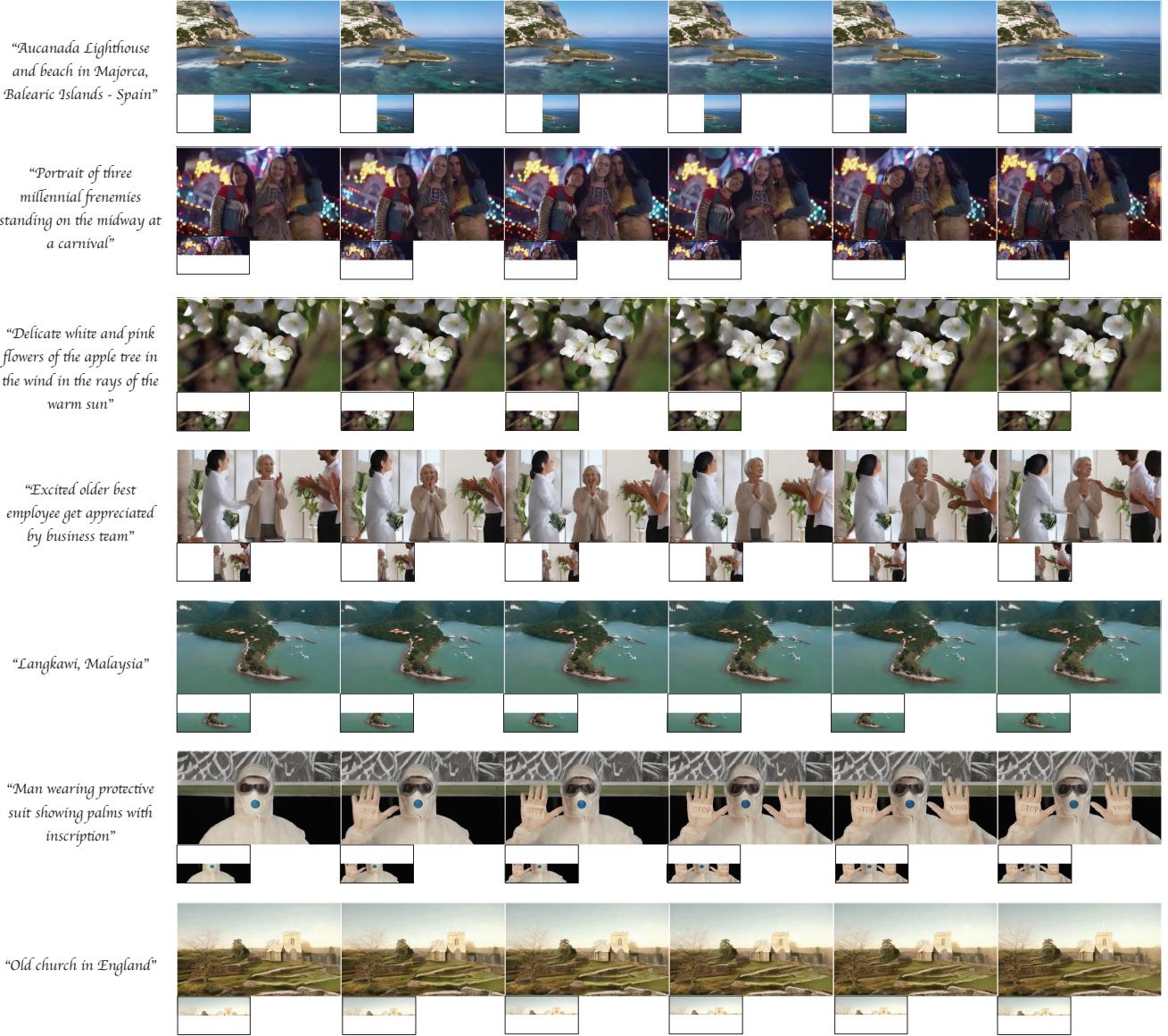


Figure 8. Qualitative visualizations on compositional video inpainting task. Videos are synthesized by performing 4 denoising steps.

tional Natural Science Foundation of China under grant U22B2053 and Alibaba Group through Alibaba Research Intern Program.

References

- [1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. [3](#)
- [2] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. [3, 4](#)
- [3] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional GAN with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. [2](#)
- [4] 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. [2, 3](#)
- [5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. [2, 3](#)
- [6] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stable-video: Text-driven consistency-aware diffusion video editing. In *ICCV*, pages 23040–23050, 2023. [2, 3](#)

- [7] 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. 2
- [8] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3
- [9] 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
- [10] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2
- [11] Patrick Esser, Johnathan Chiu, Parmida Atighchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pages 7346–7356, 2023. 3
- [12] 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 *ICCV*, pages 22930–22941, 2023. 3
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. 2
- [14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3
- [15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 4
- [16] 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. 3
- [17] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 2
- [18] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *ICML*, 2023. 2
- [19] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 2
- [20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 2
- [21] Levon Khachatryan, Andranik Mojsisyan, 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. 3
- [22] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2023. 2
- [23] 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. 2, 3, 4
- [24] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 2
- [25] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, pages 10209–10218, 2023. 3
- [26] Eyal Molad, Eliah Horwitz, Dani Vavlevski, 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
- [27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhonggang 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
- [28] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, pages 18444–18455, 2023. 3
- [29] Alexander Quinn 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. In *ICML*, pages 16784–16804. PMLR, 2022. 2
- [30] 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. 2
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3
- [32] 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 *CVPR*, pages 22500–22510, 2023.
- [33] 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. *NeurIPS*, 35:36479–36494, 2022. 2
- [34] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

- [35] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 2
- [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 2, 4
- [37] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *CVPR*, pages 1532–1540, 2021. 2
- [38] 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. *ICLR*, 2023. 2, 3
- [39] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-v: A continuous video generator with the price, image quality and perks of StyleGAN2. In *CVPR*, pages 3626–3636, 2022. 2
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 4
- [41] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 2, 3
- [42] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MocoGAN: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018. 2
- [43] Ruben Villegas, Mohammad Babaizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3
- [44] 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. 2, 3, 4
- [45] Wenjing Wang, Huan Yang, Zixi Tu, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 3
- [46] 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. 2, 3, 4
- [47] 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*, 2023. 2, 4
- [48] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitz Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, pages 5264–5273, 2020. 2
- [49] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitz Dantcheva. Imaginator: Conditional spatio-temporal GAN for video generation. In *WACV*, pages 1160–1169, 2020.
- [50] Yuhan Wang, Liming Jiang, and Chen Change Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, pages 22851–22861, 2023. 2
- [51] 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:2312.04433*, 2023. 3
- [52] 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 *ICCV*, pages 7623–7633, 2023. 3
- [53] Jie Xiao, Kai Zhu, Han Zhang, Zhiheng Liu, Yujun Shen, Yu Liu, Xueyang Fu, and Zheng-Jun Zha. Ccm: Adding conditional controls to text-to-image consistency models. 2023. 2
- [54] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2
- [55] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, pages 18456–18466, 2023. 3
- [56] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 3
- [57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2
- [58] 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. 2, 3
- [59] 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
- [60] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 2, 3
- [61] 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