Overcoming Data Limitations for High-Quality Video Diffusion Models

Supplementary Material

1. Quantitative Evaluation on MSR-VTT.

In the manuscript, we used EvalCrafter [5] as the benchmark in Sec. 4. Here, we also compare our method with others on the MSR-VTT dataset [8], which is a large-scale dataset for open-domain video captioning. We follow the zero-shot test setting in Show-1 [9] to evaluate our two models. One is the fully trained T2V base model (F-base) with WebVid-10M [2] and LAION COCO [1]. The other is the model obtained by directly fine-tuning the spatial modules of the T2V base model using JDB [6], *i.e.,* F-Spa-DIR.

The results are shown in Tab. 1. Our F-base model achieves the best FVD, and its CLIPSIM is comparable to other models. After finetuning on the image dataset, JDB, the FVD of F-Spa-DIR becomes higher compared to F-base. One reason is the distribution shift when training with JDB. The picture quality of the generated videos is greatly improved, which is significantly different from that of WebVid-10M and MSR-VTT. The aesthetics is more similar to the results of Midjourney, rather than WebVid-10M and MSR-VTT. In terms of CLIPSIM, the performance of F-Spa-DIR is comparable to other models.

	Resolution	FVD (↓)	CLIPSIM (↑)
Make-A-Video [4]	256x256	-	0.3049
ModelScope [7]	256x256	550	0.2930
VideoLDM [3]	320x512	-	0.2929
Show-1 [9]	256x256	538	0.3072
Ours(F-base)	320x512	485	0.3005
Ours(F-Spa-DIR)	512x512	653	0.2962

Table 1. Comparison on the MSR-VTT dataset.

2. Image Data Influence

In Sec. 3.3 of the manuscript, we presented the influence of high-quality image data on concept composition. Here, we illustrate more visual examples in Fig. 1.

In most cases, the model trained with JDB (F-Spa-DIR) achieves better performance than the model trained with LAION Aesthetics V2 (F-Spa-DIR-LAION) in terms of accuracy in covering concepts, image structure, and artifacts. F-Spa-DIR is significantly better, especially when the concepts contain style. For example, in the first row of Fig. 1, the result of F-Spa-DIR reflects the concepts such as *'koala'*, *'wearing a leather jacket'*, and *'walking down a street'*. Meanwhile, F-Spa-DIR-LAION misses *'wearing a leather jacket,'*. The third row shows another example. F-Spa-DIR not only captures the style *'sketch'* but also *'blue'*, *'riding a scooter'*, and *'the sun in the sky'*. However, F-Spa-DIR-LAION only shows the blue cat in the sketch.

Methods	Text-Video Alignment	Visual Quality
F-Spa-DIR vs F-Spa-DIR-LAION	65%	80%

Table 2. Human preference. The numbers represent the probability of users choosing our method.

Moreover, we also conduct a user study to compare F-Spa-DIR-LAION and F-Spa-DIR in two aspects, *i.e.,* concept composition (text-video alignment) and visual quality. We use the 50 prompts in Sec. 4.2 of the manuscript and ask three participants to rate the generated videos. The results are shown in Table 2. F-Spa-DIR performs better than F-Spa-DIR-LAION in both concept composition and visual quality.

3. Visual Examples

In the manuscript, we showed a few examples in Fig.1 and Fig. 6. Here we present more visual examples generated by our model (F-Spa-DIR). The results are shown in Fig. 2 and Fig. 3.

References

[1] Laion-coco. Accessed October 22, 2023 [Online] `https://laion.ai/blog/laion-coco/`. 1

[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*, 2021. 1

[3] 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*, 2023. 1

[4] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 2022. 1

[5] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2023. 1

[6] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *arXiv preprint arXiv:2307.00716*, 2023. 1

[7] 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

[8] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In

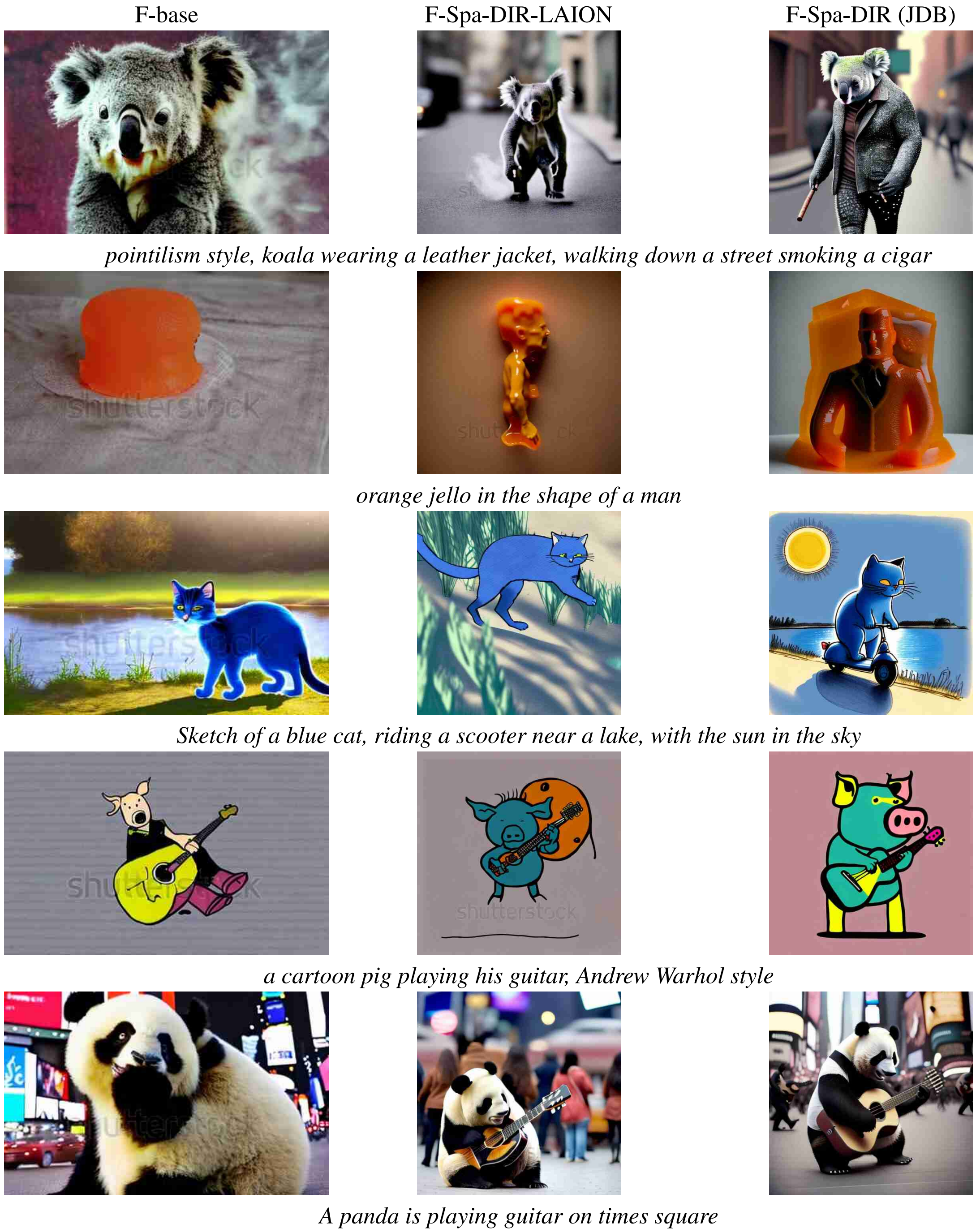

Figure 1. Influence of the high-quality image data. *Best viewed with Acrobat Reader. Click the images to play the video clips.*

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5288–5296, 2016. 1

[9] 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, 2023. 1

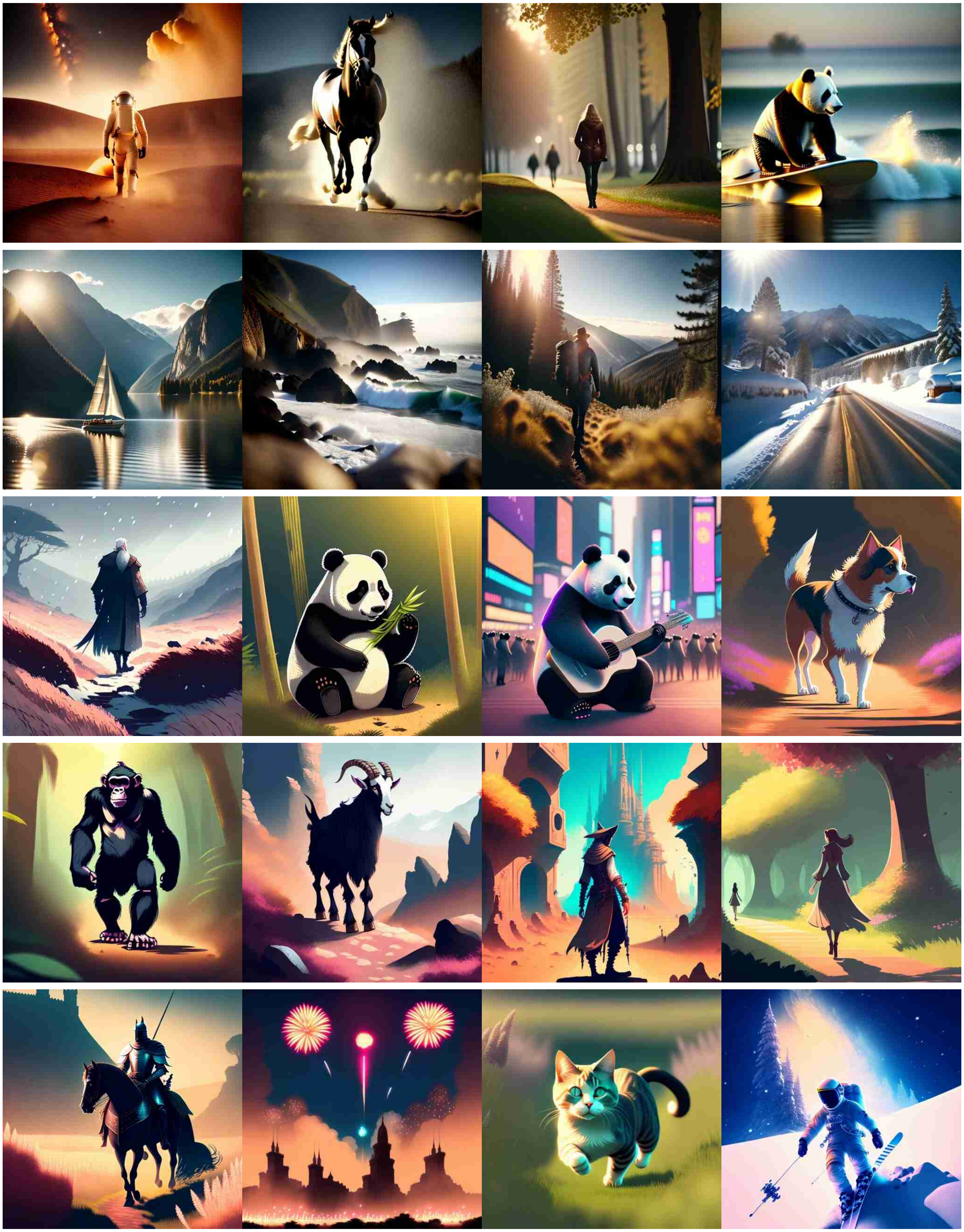

Figure 2. More visual examples of F-Spa-DIR. *Best viewed with Acrobat Reader. Click the images to play the video clips.*

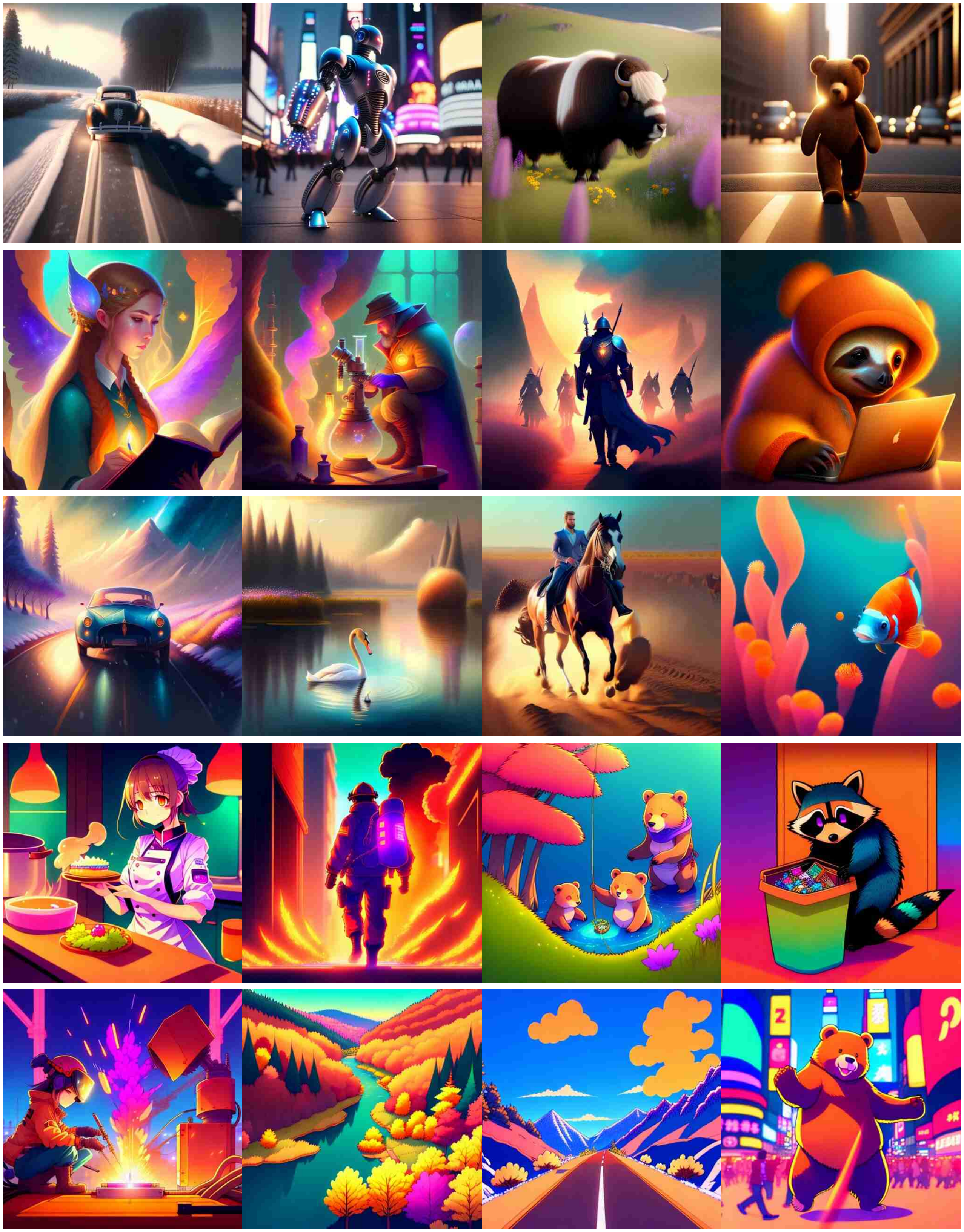

Figure 3. More visual examples of F-Spa-DIR. *Best viewed with Acrobat Reader. Click the images to play the video clips.*