DreamStyle: A Unified Framework for Video Stylization

Supplementary Material

A. Appendix

A.1. More Details about Datasets

The CT and SFT datasets are built on different data sources. For the CT dataset, we collect raw videos from an open-source dataset, Koala-36M [13], while a high-quality in-house dataset provides raw videos to make the SFT dataset. Approximately 60% of the videos across both datasets include at least one identifiable human subject, with the remaining encompassing diverse scenarios such as animals, buildings, and natural landscapes. Regarding styles, we leverage 10K diverse style images randomly selected from the Style30K [6] and WikiArt [9] datasets as style guidance for InstantStyle [12] to stylize the raw video frames in the CT dataset. As for the SFT dataset, we curate 200 style prompts crafted by designers, which are fed to Seedance 4.0 [8] to generate the stylized video frames. To accelerate the image-to-video process, our in-house I2V model with ControlNets is optimized via step distillation [7] and Classifier-Free Guidance (CFG) distillation [3], thus performing video generation with only 12 number of function evaluations (NFEs), which takes about 10 seconds on 8 NVIDIA A100 GPUs. The pass rate of video filtering in the SFT dataset is about 60%, thus in total, the image-to-video process requires about $2,400$ A100 GPU hours. Additionally, the specific VLM we use for data generation is Doubao VLM [10], and the system prompt to generate video captions is detailed in Fig. 1.

A.2. Inference Settings and Model Information

Table 1 details the inference configuration of DreamStyle. Currently, we use the original Wan14B-I2V [11] model without performance optimization, so the configuration largely adheres to the base model's settings. During inference, we merge the standard LoRA and the token-specific LoRA with token type $i = 1$ (corresponding to the video condition) into the base model, thus only the token-specific LoRA for image condition tokens incurs additional computational costs. We evaluate the inference speed using an 81-frame 480P video—its token length in DiT is $31,668$. The number of trainable parameters and inference speed of DreamStyle and the open-source competitors are illustrated in Table 2. StyleMaster [14], built on the smaller Wan1.3B-T2V, consequently achieves the fastest inference speed. Despite that, compared to VACE [4] and VideoX-Fun [1], DreamStyle has the fewest trainable parameters and introduces minimal extra computation overhead relative to the base model thanks to the efficient condition injection mechanism.

Table 1. Inference settings.

Parameter	Value
Sampler	UniPC [15]
Sampling Steps	40
CFG	3.0 for first-frame-guided task 5.0 for others
Timestep Shift	5.0
Data Type	BFloat16
Device	1 × NVIDIA A100

Table 2. Parameters and inference speed. The inference speed is measured using the average time of a denoising step on a single NVIDIA A100 GPU.

Method	Trainable Parameters	Inference Speed
StyleMaster	501M	4.83s / iteration
VACE	3049M	21.08s / iteration
VideoX-Fun	28579M	34.50s / iteration
Wan14B-I2V	-	18.33s / iteration
DreamStyle	561M	20.32s / iteration

Table 3. User study on three video stylization tasks.

Condition	Method	Metrics		
		Style Consistency	Content Consistency	Overall Quality
Text	Luma	2.05	2.58	2.24
	Pixverse	2.83	2.95	2.82
	Runway	2.52	2.80	2.59
	DreamStyle	**4.14**	**3.95**	**3.95**
Style Image	StyleMaster	1.17	-	1.31
	DreamStyle	**4.36**	3.87	**4.20**
First Frame	VACE	2.35	**4.30**	2.79
	VideoX-Fun	3.19	4.22	3.42
	DreamStyle	**4.37**	4.12	**4.24**

A.3. User Study

Human feedback serves as an important method for evaluating stylization performance, thus we conduct a user study focusing on three core metrics: style consistency, content consistency, and overall quality. Each metric is rated on a 1-5 scale, with the detailed evaluation criteria provided in Table 4. We recruited 20 professional data annotators as evaluators and randomly selected 10, 20, and 20 samples from the text-guided, style-image-guided, and first-frame-guided test sets, respectively, for blind evaluation. As shown in Table 3, DreamStyle outperforms other methods across all three stylization tasks, with a notable superiority in style consistency. Its overall quality score reaches approximately 4 or higher, reflecting user recognition of its performance.

Role
You are a video content analysis expert. Please generate descriptions (with and without style information) for user-uploaded video frames, and output them in JSON format.

Input
Video frames: Images extracted sequentially and evenly from a video.

Task
Describe the static and dynamic contents appearing in the video frames, including but not limited to static subjects and backgrounds, subject actions, and background changes.

Avoid descriptions starting with “subject is” or “background is”, and the total description length should not exceed 200 words.

Generate two versions of description: one with style information and the other without. Style information is defined as: artistic genre, subject and background colors, hue, texture pattern, and material.

Two versions of description should be as same as possible, except for the style infomation.

Output
Structured output in JSON format, including:
- caption_no_style: Video description without style information.
- caption_style: Video description with style information.

Figure 1. System prompt for Doubao VLM to generate video captions.

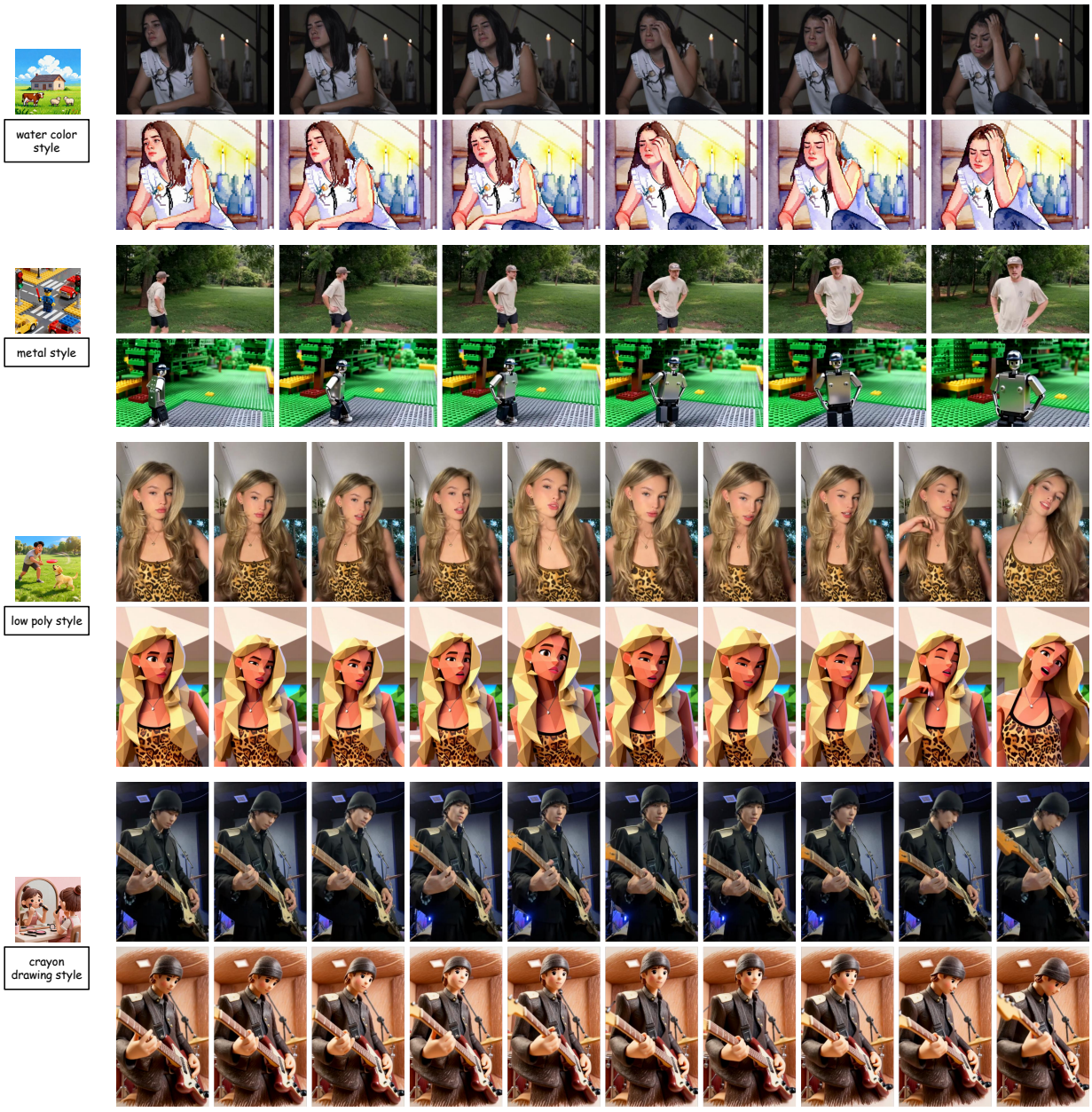

Figure 2. Visual results of multi-style fusion.

Table 4. Details of evaluation criteria.

Metric	Score	Description
Style Consistency	5	Both the main subject and background perfectly align with the style reference, with stable style throughout the entire video
	4	The main subject and background are relatively consistent with the style reference, or there are minor style degradation across the video
	3	Either the main subject or the background is somewhat inconsistent with the style reference, or the video exhibits noticeable style variations
	2	Neither the main subject nor the background aligns with the style reference, or the video has significant style inconsistencies
	1	The main subject and background are completely inconsistent with the style reference
Content Consistency	5	Both the main subject and background are highly consistent with the input video, and the motion of the main subject is also highly coherent
	4	Either the main subject or the background has slight discrepancies from the input video, or the motion of the main subject is somewhat inconsistent
	3	Either the main subject or the background has noticeable differences from the input video, or the motion of the main subject is highly inconsistent
	2	Both the main subject and background show obvious deviations from the input video
	1	The main subject and background are completely unrelated to the input video
Overall Quality	5	Excellent performance in both style consistency and content consistency, with aesthetically pleasing visuals and rational motion
	4	Either style consistency or content consistency needs improvement; or the visuals are generally aesthetically acceptable, with slight motion glitches
	3	Either style consistency or content consistency is poor; or the visuals have low aesthetic appeal, with noticeable motion issues
	2	Both style consistency and content consistency are poor, with unappealing visuals and significant motion issues
	1	Extremely poor performance in both style consistency and content consistency

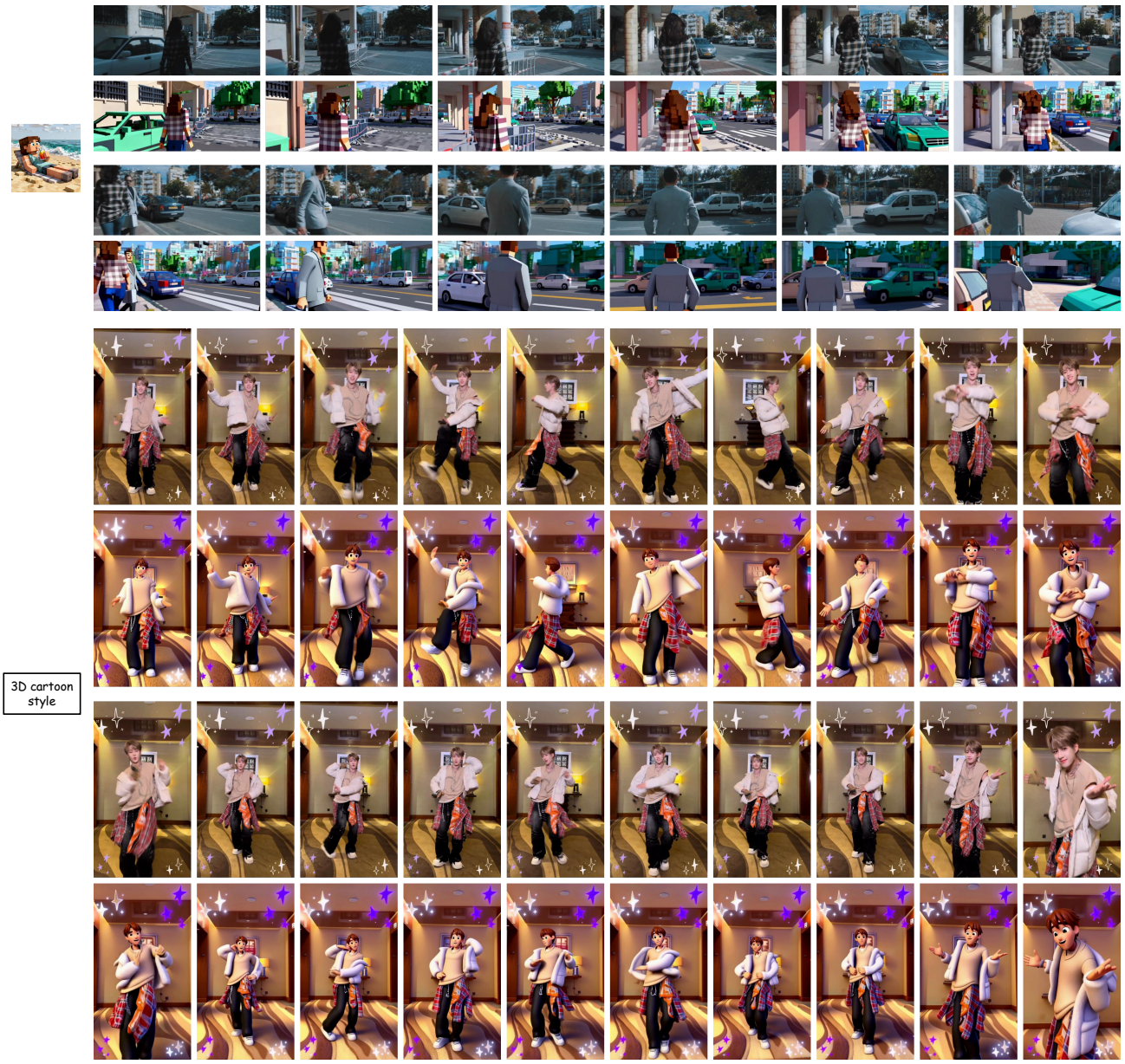

Figure 3. Visual results of long-video stylization.

A.4. Extended Applications

Although DreamStyle is trained with only a single condition type at a time, it still supports multiple guidance modalities during inference, thereby unlocking its potential to enable broader extended applications. Below, we highlight two representative scenarios:

Multi-Style Fusion. As shown in Fig. 2, DreamStyle can naturally integrates the style cues from both text prompts and style images, demonstrating its capability to fuse diverse style references and create a novel style. This flexibility allows for a creative combination of abstract textual description and precise visual reference, exhibiting the potential beyond single guidance.

Long-Video Stylization. By leveraging the last frame of a generated short video as the first frame condition for the next segment, we can seamlessly concatenate two short video clips. Thus, a combination of first frame guidance and text or style image enables DreamStyle to overcome the 5-second duration limit, supporting stylization for longer video sequences (except multi-shot video due to the inherent limitations of the base model and training data). Fig. 3 presents two long-video stylization examples, guided by style image and text, respectively.

A.5. Advantages of Data Curation Pipeline

The SOTA datasets (EditVerse [5], Senorita-2M [16]) rely on depth/HED/Canny control conditions, which have an intrinsic shortcoming that brings a strong structural constraint and thus fails at styles with large geometric deformation (Fig. 5 (a) shows our depth ControlNet also inherits this limitation). To address this, we further customize a pose ControlNet that enables stylization of geometrically deformed subjects. Note that the pose can not capture the dynamic of background regions, thus we leverage the same pose sequence to animate both stylized and raw videos to improve the success rate of generated paired data. The datasets made by depth and pose control conditions are complementary, making our model compatible with a larger range of styles. Moreover, as shown in Fig. 5 (b), we validate that our datasets enable DreamStyle's generalization ability that handles both geometric styles and background dynamics, simultaneously.

A.6. Architecture of I2V ControlNet

Our I2V ControlNet is built on the MM-DiT [2] backbone. As shown in Fig. 4, in contrast to standard ControlNet, we insert an extra MM-DiT block after each original block. The control conditions (e.g., depth, pose) are encoded by a convolutional network and aligned to the video token length. Instead of concatenation as in video-text MM-DiT, we add the Q, K, and V features from video and control condition, which enhances structure injection and improves consistency between generated videos and control conditions.

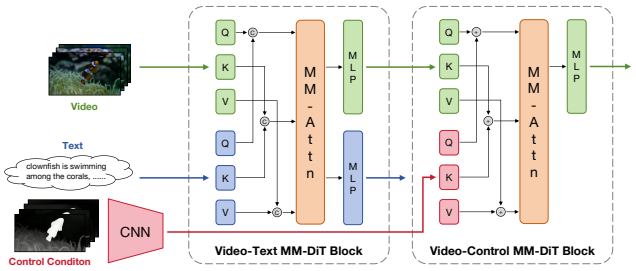

Figure 4. Architecture of our I2V ControlNet.

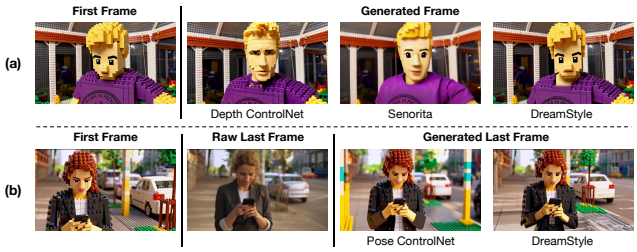

Figure 5. Visual comparison with I2V ControlNets and Senorita.

Table 5. Quantitative comparison.

	Depth CtrlNet	Pose CtrlNet	Senorita	DreamStyle
CSD Score	0.847	**0.865**	0.793	<u>0.851</u>
DINO Score	<u>0.627</u>	0.528	0.618	**0.640**

Table 6. Computational efficiency on 480P 5s video.

	Num. of Tokens	GPU Mem. Usage	Infer. Speed
Channel-wise	31,668	57.152GB	20.32s / step
Frame-wise	63,336	65.158GB	52.79s / step

A.7. More Comparisons

Comparison with I2V ControlNets and Senorita. Our I2V ControlNets and Senorita [16] both belong to first-frame propagation, so here we focus on the first-frame-guided setting. Table 5 shows that DreamStyle outperforms the others in content preservation (DINO score), and achieves the second-best style consistency (CSD score). Visual comparison in Fig. 5 (a) demonstrates the inherent flaw of our depth ControlNet and Senorita, alongside DreamStyle's ability to handle the styles with geometric deformation. Our pose ControlNet, while enabling geometric styles, fails to maintain the consistency of background regions, thus being inferior to DreamStyle (Fig. 5 (b)). By leveraging data cleaning and joint training on depth and pose datasets, DreamStyle exhibits overall superiority over our standalone I2V ControlNets.

Efficiency of Channel-wise Injection. Table 6 presents a comparison of the computational efficiency between channel-wise injection and another commonly used in-context frame-wise injection.

Figure 6. Visual example of the limitation.

A.8. Limitations

A major limitation of DreamStyle lies in its run-time performance. It requires several minutes to generate an 81-frame video on a NVIDIA A100 GPU, which falls far short of enabling real-time or interactive stylization for end users on consumer-grade GPUs. Besides, different style conditions may conflict with each other. Fig. 6 shows an example that the style image specifies a grayscale video, but the text prompt "green" has an opposite effect. Moreover, the upper bound of current stylization performance is largely constrained by the image stylization techniques used for data generation, which might be alleviated by advances in data pipelines and training paradigms in the future.

References

[1] AIGC-Apps. A more flexible framework that can generate videos at any resolution and creates videos from images. https://github.com/aigc-apps/VideoX-Fun. 1

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 4

[3] Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, and Ratheesh Kalarot. Plug-and-play diffusion distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13743–13752, 2024. 1

[4] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17191–17202, 2025. 1

[5] Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, Daniil Pakhomov, Zhe Lin, Soo Ye Kim, and Qiang Xu. Editverse: Unifying image and video editing and generation with in-context learning. *arXiv preprint arXiv:2509.20360*, 2025. 4

[6] Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 1

[7] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 1

[8] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 1

[9] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. 1

[10] ByteDance Seed Team. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1

[11] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[12] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 1

[13] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 1

[14] Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2630–2640, 2025. 1

[15] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023. 1

[16] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *arXiv preprint arXiv:2502.06734*, 2025. 4