LongAnimation: Long Animation Generation with Dynamic Global-Local Memory (Supplementary Material)

Nan Chen[1], Mengqi Huang[1], Yihao Meng[2], Zhendong Mao[1†]
[1]University of Science and Technology of China
[2]Hong Kong University of Science and Technology
{chen_nan,huangmq}@mail.ustc.edu.cn, ymengas@connect.ust.hk, {zdmao}@ustc.edu.cn

In this appendix, we will first provide detailed implementation information in Appendix A and present additional experimental results in Appendix B. Qualitative analysis of images may not be as compelling as directly comparing videos. Watching videos at https://cn-makers.github.io/long_animation_web/ will offer a more direct and clear understanding.

A. More Implementation Details

A.1. Implementation Details

During training, we set the layer numbers of cross attention in Dynamic Global-Local Memory $M = 3$. Specifically, we extract the KV cache of the Long Video Understanding model from layers 20, 24, and 28 to obtain global and local features. During training, we randomly extract a frame in the video sequence as the reference image to improve the robustness. The closer the frame is to the first frame, the greater the probability of being selected. During inference, we use CogVideoXDPMScheduler to denoise 50 steps.

A.2. Training Strategies

In the first training stage, we randomly select 81 frames from animation videos for training and randomly choose reference images from these frames. In the second and third training stages, we select videos with more than 162 frames (some videos have up to 1000 frames) and randomly extract 81 frames from the segment following the initial 81 frames for training, with a random key frame from the historical segments as the reference frame. The frames before the chosen 81 frames are processed by the long video understanding model to extract global memory (all historical videos) and local memory (the previous 81 frames). In the third training stage, to accelerate training, we optimize the non-gradient color consistency reward during the 30-step denoising process.

Method	Base Model	Type	Local Strategy	Global Strategy	Color Reward
ToonCrafter [4]	DynamiCrafter [5]	UNet	✗	✗	✗
Layeranimate [6]	ToonCrafter [5]	UNet	✗	✗	✗
LVCD [2]	SVD [1]	UNet	✓	✗	✗
AniDoc [3]	SVD [1]	UNet	✗	✗	✗
LongAnimation	CogvideoX [7]	**DiT**	✓	✓	✓

Table 1. An overview of previous SOTAs on the animation generation task.

A.3. Prior SOTAs

We compared all the open-source methods, performing inference using their default settings. It should be noted that ToonCrafter[4] and Anidoc[3] do not have the capability to generate long videos with their initial settings. Therefore, we use frame segment overlap to improve the consistency of generating multiple frame segments and finally splice these frame segments together.

In Tab. 1, we provide a comprehensive comparison between our method and previous methods. As shown in Tab. 1, previous methods do not consider global consistency. Our method enhances global consistency through Dynamic Global-Local Memory and further refines the model's colorization capabilities using the color consistency reward.

B. More Experiments Results

B.1. Reference Image Selection

During inference, we use the same reference image to generate all animation segments, rather than autoregressively using the last frame of the generated segments as the reference image for the next segment, as there is a gap between the generated video and the reference video. Autoregressively using the last generated frame as the reference image for the next segment leads to error accumulation. Although this issue is not noticeable in the short-term generation, it results in noise in the long-term generation, as shown in

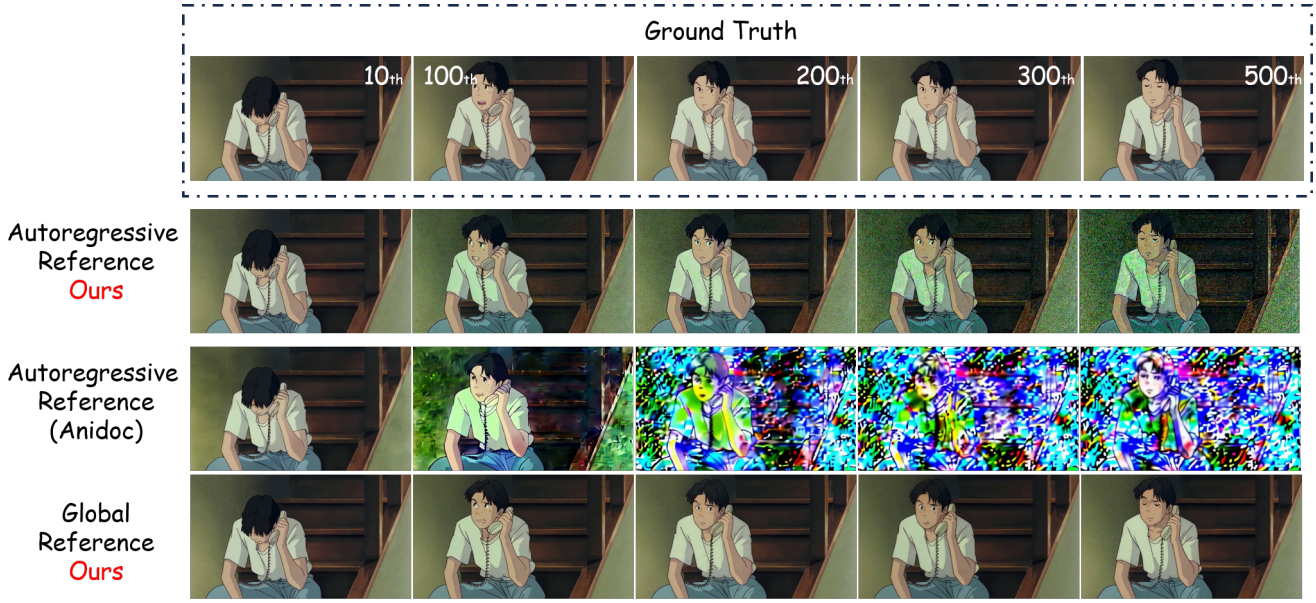

Figure 1. **Qualitative comparison** of global reference (*i.e.*, all frame segments are generated using the same reference image) and autoregressive reference (*i.e.*, autoregressively using the last frame of the previously generated segment as the reference image for the next generation). The animation generated by autoregressive reference gradually becomes noisy after 200 frames. The noise at the 500th frame is very obvious.

ID	Control Type	LPIPS ↓	SSIM ↑	PSNR ↑	FVD ↓
0	Full Control	0.097	0.812	23.82	333.09
1	Latter Half	0.093	0.817	23.75	328.89
2	Former Half	0.091	0.815	23.66	323.95
3	Skip-Layer	**0.086**	**0.838**	**24.46**	**321.62**

Table 2. Ablation studies on the effectiveness of the feature injection mode for SketchDiT. Compared to other control styles, the skip-layer achieves the best performance.

Method	FCC ↑ (0 ∼ 10)	VC (0 ∼ 10) ↑	COV (0 ∼ 10) ↑
LVCD[2]	4.20	5.11	3.63
LVCD*	5.10	5.37	4.63
Anidoc[3]	3.34	4.47	3.39
Ours	8.55	9.00	8.40

Table 3. Human perception evaluation for different methods. Our method outperforms previous methods in all aspects, achieving superior long-term colorization effects.

Fig. 1. This phenomenon also occurs in the previous advanced models (*e.g.*, Anidoc[3]), so we use the same reference image for all these models during inference.

B.2. More qualitative comparisons.

In the supplementary material, we present more qualitative comparisons, as shown in Fig. 2 and Fig. 3. Compared to previous methods, our approach better preserves the color details of the characters' clothing. This indicates that our Dynamic Global-Local Memory mechanism facilitates global color consistency. More direct qualitative comparisons can be viewed through the project page.

B.3. More Ablation Studies

In this section, we will present more ablation experiments. All ablation experiments are tested on long video (average 500 frames).

Effectiveness of feature injection mode. As shown in Tab. 2, our skip-layer control for SketchDiT achieves the best performance. Since SketchDiT has significantly fewer layers than CogvideoX, using cross-layer control instead of full-layer control can reduce computational overhead and improve training efficiency. Additionally, compared to controlling only the latter or former half of CogvideoX, cross-layer control provides more diverse features, thereby improving performance.

B.4. Human Perception Evaluation

We recruit 100 volunteers to evaluate *longAnimation* with previous methods on three aspects: frame colorization consistency (FCC), video coherence (VC), and consistency with the original video (COV), using the scale from 1 to 10. We present the volunteers with 20 long animations generated by different methods using the same reference images.

Figure 2. **Qualitative comparison** with existing methods. *LongAnimation* can achieve long-term consistent animation generation through Dynamic Global-Local Memory (e.g., the girl's clothes). In contrast, previous methods exhibit unstable color changes. Watching videos will offer a more direct and clear understanding.

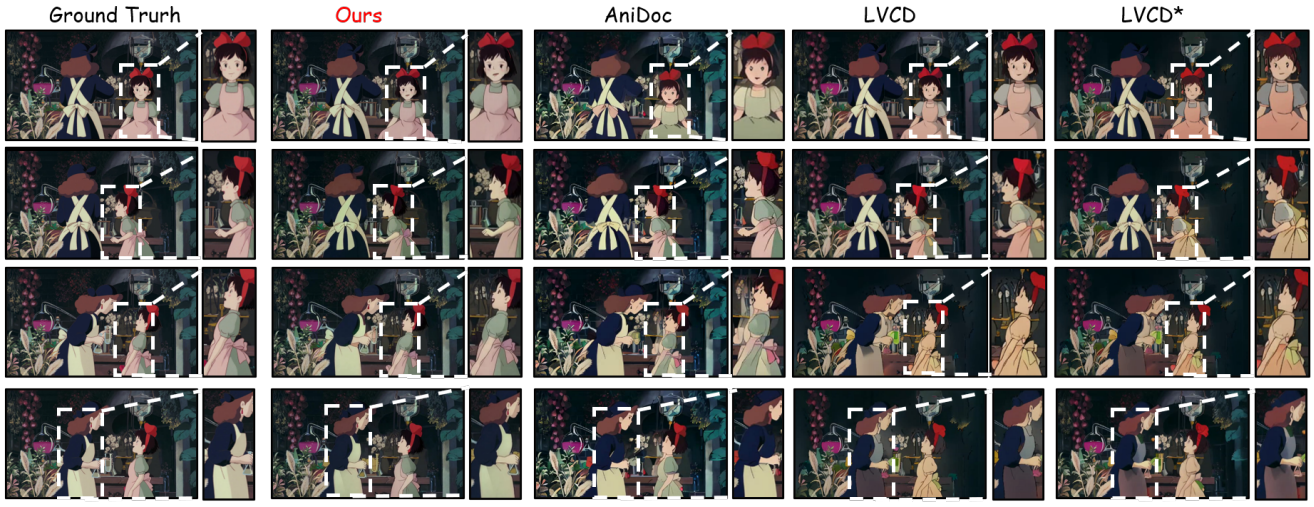

Figure 3. **Qualitative comparison** with existing methods. Watching videos will offer a more direct and clear understanding.

The results are shown in Tab. 3. As indicated in Tab. 3, our method outperforms previous methods in all aspects, achieving superior long-term colorization effects.

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] Zhitong Huang, Mohan Zhang, and Jing Liao. Lvcd: reference-based lineart video colorization with diffusion models. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024. 1, 2

[3] Yihao Meng, Hao Ouyang, Hanlin Wang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Zhiheng Liu, Yujun Shen, and Huamin Qu. Anidoc: Animation creation made easier. *arXiv preprint arXiv:2412.14173*, 2024. 1, 2

[4] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024. 1

[5] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Con-*

ference on Computer Vision, pages 399–417. Springer, 2024. 1

[6] Yuxue Yang, Lue Fan, Zuzen Lin, Feng Wang, and Zhaoxiang Zhang. Layeranimate: Layer-specific control for animation. *arXiv preprint arXiv:2501.08295*, 2025. 1

[7] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1