Supplementary Materials: Towards Photorealistic Video Colorization via Gated Color-Guided Image Diffusion Models

Anonymous Authors

1 USER STUDY

To evaluate the subjective quality of our method in the task of video colorization, we conducted a user study involving 30 participants who were not afflicted with color blindness. We randomly selected five videos from both the DAVIS-Test-Dev 2017 test dataset [4] and our curated LDV 3.0 test dataset [5]. The experiment was divided into two groups: automatic video colorization methods and methods utilizing the first frame as a reference image for colorization. The corresponding grayscale videos and the results of the methods under comparison were concatenated to ensure that the video results of all methods could be played simultaneously, with the appearance order randomized. Each participant was required to select the video they deemed to have the best colorization quality (CQ) effect and the highest degree of temporal consistency (TC), with the percentage of votes for each method out of the total votes shown in Table 1.

From the results in Table 1, it can be observed that our method has a clear advantage, with its colorization results highly recognized and positively evaluated by users, achieving a voting rate of over 50% in both automatic and example-based colorization. Users highly praised the quality, realism, and visual perception of the generated color videos. The model's colorization results not only exhibited outstanding performance in color consistency but also demonstrated exceptional visual performance in terms of visual perception.

Table 1: User study for automatic and example-based colorization methods.

Colorization Type	Method	CQ	TC
Automatic	peoldify[1]	21.67%	16.33%
	TCVC[3]	14.33%	20.33%
	VCGAN[8]	7.00%	10.00%
	Ours	**57.00%**	**53.33%**
Example-based	DRemaster[2]	6.67%	5.67%
	DExample[7]	14.67%	17.33%
	BiSTNet[6]	24.67%	26.67%
	Ours	**54.00%**	**50.67%**

2 ABLATION ON NUMBER OF KEYFRAMES

In this section, we focus on the ablation study on the number of keyframes and assess the impact of changes in the number of keyframes on generation performance, as a supplement to the main experiments. This ablation was performed on the LDV3.0 test datasets and the DAVIS-Test-Dev 2017 test datasets. To ensure video length alignment and enhance testing efficiency, we used only the first 30 frames of each video segment. Additionally, to ensure fairness in comparison and eliminate the potential influence of parallel coloration of multiple video frames when $batchsize > 1$, we set $batchsize = 1$ during the single-step coloration process (this means that the actual generation time is shorter, but it is adjusted here for a better comparison). In addition to metrics that can represent the quality of supervised video colorization (FID&CDC), we also calculated the time required for joint colorization with different numbers of key frames, i.e., T_{stage1}, as well as the total time required for our two-stage process, i.e., T_{stage1}+ T_{stage}. The final results are displayed in Table 2 and Figure 1.

# keyframes	T_{stage1}	T_{stage1}+T_{stage2}	CDC ↓	FID ↓
1	4s	91s	0.4421	34.21
2	5s	89s	0.3858	31.65
3	7s	88s	0.3472	29.74
4	10s	88s	0.3321	28.31
5	13s	88s	0.3245	28.43
10	37s	97s	0.3194	27.41
15	72s	117s	0.3168	26.95
20	120s	150s	0.3175	27.15
25	179s	194s	0.3157	27.01
30 (full video frames)	250s	250s	0.3162	26.87

Table 2: Result I of Ablation Study on Keyframe Quantity(partial)

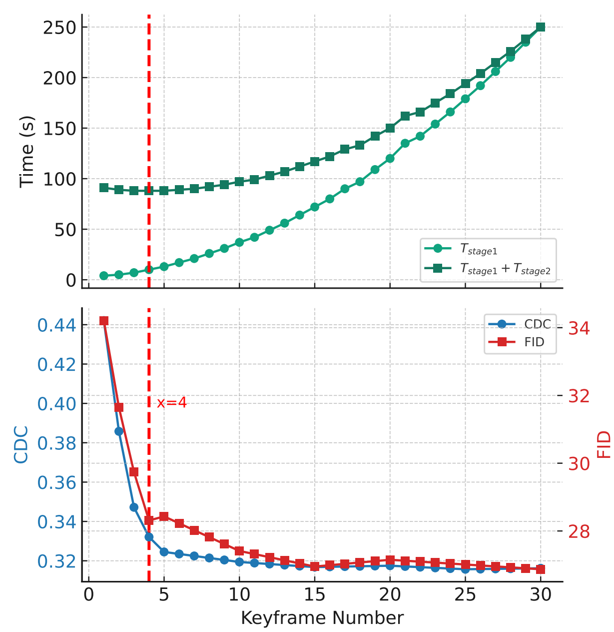

Figure 1: Result II of Ablation Study on Keyframe Quantity

From the results, we observe that as the number of keyframes increases, the time required for joint coloring (*Stage1*) rapidly escalates. This surge is attributed to the increased computational cost of the extended attention mechanism, which approximates to $O(x^2)$, where x representing the number of keyframes. This also implies that joint coloring of a large volume of video frames would incur significantly higher costs, resulting in decreased efficiency. Additionally, it is noteworthy that when the number of keyframes ranges from 1 to 3, there is a brief decline in total time, which can be attributed to the GPU transitioning from an idle to a fully loaded state during *Stage1*. Furthermore, regarding the quality of coloring, it is evident that when the number of keyframes is relatively small, both CDC and FID indices decrease rapidly, indicating that increasing the number of keyframes significantly enhances the quality of coloring for smaller sets of keyframes. Subsequently, there is a gradual and steady improvement in coloring quality.

Based on the above analysis, we selected $x = 4$ as the default number of keyframes in *Stage1* in our experiments, allowing for high-quality coloring results within a shorter duration. Moreover, this approach permits users the flexibility to choose the number of keyframes, balancing between coloring efficiency and quality.

3 ADDITIONAL COLORING EXAMPLES

We provide numerous additional coloring examples, including videos generated using our model for video colorization, along with comparison videos of other automatic and example-based video colorization models. These results offer supplementary information and outcomes to complement the main content of the paper. Please refer to the README file in the supplementary materials, as well as the examples in the "videos" folder.

REFERENCES

[1] Jason Antic. 2019. DeOldify: A Deep Learning based project for colorizing and restoring old images. https://github.com/jantic/DeOldify.

[2] Satoshi Iizuka and Edgar Simo-Serra. 2019. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

[3] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. 2024. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *Computational Visual Media* 10, 2 (2024), 375–395.

[4] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.

[5] Ren Yang, Radu Timofte, Xin Li, Qi Zhang, Lin Zhang, Fanglong Liu, Dongliang He, Fu Li, He Zheng, Weihang Yuan, et al. 2022. Aim 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In *European Conference on Computer Vision*. Springer, 174–202.

[6] Yixin Yang, Jinshan Pan, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, and Jinhui Tang. 2024. Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[7] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. 2019. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8052–8061.

[8] Yuzhi Zhao, Lai-Man Po, Wing Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou. 2022. Vcgan: Video colorization with hybrid generative adversarial network. *IEEE Transactions on Multimedia* (2022).