

AniDoc: Animation Creation Made Easier

Supplementary Material

Appendix

A. Reference with Different Background

When using images of the same character with different backgrounds as references, our model can transfer the style from the reference images to generate new backgrounds with diverse styles. The original character remains consistent in their core features, such as expressions and clothing, while the integration of varied backgrounds enriches the overall visual effect, as shown in Fig. S1.



Figure S1. Illustration of reference with different backgrounds.

B. Multiple Characters

Although our work primarily focuses on a single reference image and does not include specific training or processing for handling multiple references, we observe that our model can automatically distinguish between multiple characters in a reference image based on their unique features. It applies the correct coloring to each character, even when there are significant differences in poses, angles, or relative positions between the reference and the line art, as demonstrated in Fig. S2.



Figure S2. Illustration of the multiple characters situation. When the reference image contains multiple characters, our method can correctly infer the correspondence and apply colorization to each character accordingly.

C. Different Line Art Extraction Methods

To evaluate the generalization capability of our method under different line art conditions, we test its performance

using various line art extraction methods. Besides the default line art extraction method [2] used in our paper, we also apply three additional methods: Anime Lineart [3], HED [8] and PiDiNet [6]. Among these, Anime Lineart is a line art extraction method specifically trained on anime datasets. HED, as an edge detection method, produces relatively thick line art, whereas PiDiNet creates simplistic line art that is closer to hand-drawn style.

After extraction, we apply the same binarization process to the line art as described in the main text. Our method successfully colors the line art under different conditions while maintaining consistency with the reference. Due to the varying characteristics of the extracted line art, our method generates different results accordingly.

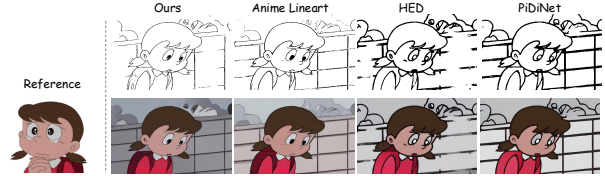


Figure S3. Impact of different line art extraction methods.

D. Motivation for Correspondence Matching



Figure S4. In the early training stage (10k step), the video generation model produces static videos that closely resemble the given reference design.

As an image-to-video model, SVD (Stable Video Diffusion) [1] inherently possesses the ability to extract information from an input image to generate a video. However, during training, we observe that the strong prior in SVD restricts the first frame to be the same with the input reference image, as shown in Fig. S4.

In our formulation, the input image is not the first frame of the video but rather a reference character design from a different viewpoint. The model needs to query colors from this reference image, while the structure information should

align with the given sketch list. This conflicting prior makes training the model significantly more challenging.

To better establish the relationship between the reference character design and the sketch, reduce the learning difficulty for the model, and improve the fine-grained details, we propose a Correspondence Matching Module. This module explicitly injects the matching relationships between the reference image and the sketch, enabling the model to better query and color the correct areas.

E. Illustration of DIFT Matching



Figure S5. Semantic feature can effectively find matching keypoints between reference color image and binarized sketch.

During training, we apply low-level techniques Light-Glue [4] with SIFT descriptor [5] for keypoint selection and matching between the reference image and the training video frames due to its fast speed. During inference, we lack access to the ground truth color image. Techniques that rely on low-level image features, such as SIFT descriptors, are ineffective at accurately matching keypoints between sketches and color reference images due to the significant domain gap between them. Therefore, we use the semantic level keypoint matching method DIFT [7] to establish the correspondence between the color reference image and the sketches, as shown in Fig. S5.

F. Limitation

Although our method can colorize multiple clips containing the same character based on a single character design sheet while maintaining good character consistency, it still has certain limitations.

First, when a line art clip contains objects that are not present in the reference, the model struggles to determine the appropriate colors for these objects, as shown in the 1st row of Fig. S6. It can only infer colors based on the color information available in the reference, leading to inaccuracies in the colorization.

Second, when the clothing of a character in the line art clip differs from that in the reference image (even though it is the same character), our model can only infer reasonable

colors based on the color patterns of the character’s clothing in the reference image. However, our method cannot guarantee accuracy in this situation, as shown in the 2nd row of Fig. S6.



Figure S6. Limitations: in the 1st row, the cartoon bear highlighted within the red square is not present in the reference image. Consequently, the model can only infer the bear’s color as purple, based on the main color of the reference character, which deviates from its actual appearance. In the 2nd row, the character’s clothing in the line art clip is different from the reference. Therefore, our model can only infer the color of the dress and scarf based on the dominant color patterns identified in the reference image.

References

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 1
- [2] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022. 1
- [3] Lei Chen and Contributors. Animegan: A fast and simple image animation method, 2024. 1
- [4] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 2
- [5] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2
- [6] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021. 1
- [7] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2
- [8] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 1