Supplementary Materials

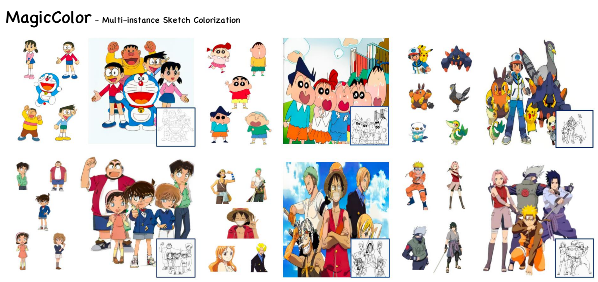

Figure 1. **Applications of our *MagicColor*.** We can use it to color anime posters.

1. Implementation Details

1.1. Network Architecture

Our ***MagicColor*** is based on Stable Diffusion 1.5 [14] and AnimeAnyone [2], an image-to-video generation model. We recommend that readers consult AnimeAnyone for more detailed information about its architectural backbone. Importantly, our method of utilizing image diffusion priors for animation can theoretically be applied to any image-to-image diffusion models that incorporate a cross-attention image-conditioning mechanism. For our implementation, we employ CLIP [13] as both the Unet and the sketch guider cross-attention encoder, while using DINOv2 as the instance encoder. Additionally, we incorporate a feedforward neural network (FFN) layer as an injection adapter to enhance image condition alignment.

1.2. Hyperparameters

All architecture parameter details, diffusion process details, as well as training hyperparameters are provided in Table 1, which should be mostly self-explanatory. Here we give some additional descriptions for some parameters:

Instance_conditioning_channels: The number of input tensor channels for the instance guider.

FFN_channel: The number of input tensor channels for the image FFN layer.

$\lambda : Thehyperparameterofbalanceedgelossandperceptualloss$

Table 1. Model Configuration

Parameter Name	Parameter Value
attention_head_dim	8
block_out_channels	[320, 640, 1280, 1280]
instance_conditioning_channels	1024
conditioning_out_channels	[16, 32, 96, 256]
cross_attention_dim	768
adapter_channel	1024
in_channels	4
layers_per_block	2
mid_block_scale_factor	1
norm_eps	1×10^{-5}
norm_num_groups	32
λ	0.1

2. Dataset construction details

In order to develop a robust colorization model, we constructed a comprehensive data pipeline that includes various datasets tailored for different training objectives. This pipeline integrates reference color pairs, multi-instance training instances, manually labeled datasets, and real-world examples to enhance the model's performance across diverse scenarios.

(1) Reference Color Pair Dataset: The data format comprises three components: (sketch, reference, color). To create this dataset, we sourced video clips featuring character-centric scenes from the Sakuga dataset [12]. For each selected video clip, we extracted two frames at regular intervals. The first frame serves as the reference image, providing the intended color palette, while the second frame acts as the color image, which will be colorized based on the reference.

(2) Multi-instance Training Dataset: This dataset follows the format of (sketch, masks, color). We utilize advanced tools such as RAM (Recognize Anything Model), Grounding DINO, and SAM (Segment Anything Model) [3, 4, 16] to perform automated labeling on the color images. These tools assist in generating accurate masks that delineate each instance within the color image. By isolating these masks, we can extract individual instances from the color images, which serve

as multi-reference conditions for our training process. This methodology allows the model to learn from multiple instances simultaneously, enhancing its ability to generalize across various scenarios.

(3) Manually Labeled Multi-instance Color Dataset: This dataset was constructed from six iconic animated films, chosen for their rich visual content and character diversity. We meticulously extracted clips from these films and obtained still images from promotional movie posters, which often feature multiple main characters. The characters in these films, being prominent figures, have a wealth of high-quality personal photographs available online. This abundance of reference images enables us to effectively gather and curate a robust dataset for training, ensuring that the model can accurately capture the unique color characteristics of each character.

(4) Real-World Dataset: Acknowledging that user-provided reference images may originate from real-world photographs rather than animated sources, we incorporated the Coconut dataset into our training regimen. The Coconut dataset is particularly valuable as it contains a substantial collection of annotated instances across various contexts. By leveraging this dataset, we can bridge the domain gap between animated content and real-world imagery, allowing the model to perform more effectively in diverse scenarios. This integration ensures that the model is well-equipped to handle a wide range of input styles and contexts.

3. Statistics of the dataset

Figure 2 presents the statistics of the dataset used in this study. The dataset is composed of four distinct components: Sakuga: This dataset consists of 9,0000 video clips that have been meticulously cleaned and processed to ensure high quality and consistency. ATD-12K [15]: This component includes 12,000 entries, contributing a substantial amount of diverse content to the dataset. Manual Collection Dataset: Comprising 1,000 items, this subset is manually curated from classic anime films and posters, providing a rich source of artistic styles and themes. Coconut: With 25,000 entries, this dataset is randomly sampled, offering a broad range of content. Overall, these components collectively form a comprehensive dataset that supports various analyses and research endeavors in the field.

4. User Interface

Based on the previous controllable generation method [1, 5–11], the front-end user interface is designed for image colorization based on sketches. On the left side, there's a section for the sketch image, where users can upload their line drawings. Below it is areas for up to three reference images, allowing users to input images of items like clothing and accessories to guide the colorization. Each image area has upload and other related function icons. On the right side, the "Generated Result" area shows the final colorized image. There's also a field indicating the number of reference images used, and a prominent "Generate Result" button for triggering the colorization process. As shown in the video supplementary materials, this UI is intuitive, enabling users to easily upload source materials and obtain vividly colorized results.

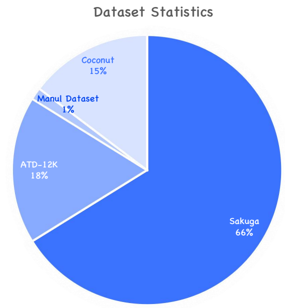

Figure 2. **Statistics of our training the dataset.**

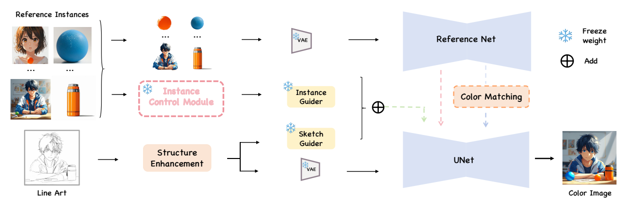

Figure 3. **Inference Pipeline.** The user inputs a line art and one or more reference images, after which our model automatically extracts masks and performs colorization.

References

[1] Daiheng Gao, Shilin Lu, Shaw Walters, Wenbo Zhou, Jiaming Chu, Jie Zhang, Bang Zhang, Mengxi Jia, Jian Zhao, Zhaoxin Fan, et al. Eraseanything: Enabling concept erasure in rectified flow transformers. *arXiv preprint arXiv:2412.20413*, 2024. 2

[2] Li Hu. Animate Anyone: Consistent and Controllable Image-To-Video Synthesis for Character Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao

Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding Dino: Marrying Dino With Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 1

[5] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2

[6] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

[7] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow Your Pose: Pose-Guided Text-To-Video Generation Using Pose-Free Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024.

[8] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*, 2024.

[9] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*, 2024.

[10] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

[11] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2

[12] Zhenglin Pan, Yu Zhu, and Yuxuan Mu. Sakuga-42M Dataset: Scaling Up Cartoon Research. *arXiv preprint arXiv:2405.07425*, 2024. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[15] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep Animation Video Interpolation in the Wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6587–6595, 2021. 2

[16] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize Anything: A Strong Image Tagging Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 1