MangaDiT: Reference-Guided Line Art Colorization with Hierarchical Attention in Diffusion Transformers

Qianru Qiu[1], **Jiafeng Mao**[1], **Kento Masui**[1], **Xueting Wang**[1]

[1]CyberAgent

qiu_qianru@cyberagent.co.jp, jiafeng_mao@cyberagent.co.jp,
masui_kento@cyberagent.co.jp, wang_xueting@cyberagent.co.jp

Abstract

Recent advances in diffusion models have significantly improved the performance of reference-guided line art colorization. However, existing methods still struggle with region-level color consistency, especially when the reference and target images differ in character pose or motion. Instead of relying on external matching annotations between the reference and target, we propose to discover semantic correspondences implicitly through internal attention mechanisms. In this paper, we present MangaDiT, a powerful model for reference-guided line art colorization based on Diffusion Transformers (DiT). Our model takes both line art and reference images as conditional inputs and introduces a hierarchical attention mechanism with a dynamic attention weighting strategy. This mechanism augments the vanilla attention with an additional context-aware path that leverages pooled spatial features, effectively expanding the model's receptive field and enhancing region-level color alignment. Experiments on two benchmark datasets demonstrate that our method significantly outperforms state-of-the-art approaches, achieving superior performance in both qualitative and quantitative evaluations.

Introduction

The rapid advancements in generative models have revolutionized content creation in the anime and manga industries. Among these developments, reference-guided line art colorization, as shown in Figure 1, has attracted substantial attention due to its practical value in creative workflows. This task enables efficient color transfer from a reference image to a line drawing, ensuring consistent coloring across corresponding semantic regions.

However, reference-guided colorization remains challenging. The core issue lies in preserving color consistency, especially in regions with fine-grained details such as hair, clothing accessories, or intricate decorations. The task becomes even more difficult when character poses or motions differ significantly between the reference and target images. Existing methods (Dai et al. 2024; Yan et al. 2025a,b), including diffusion-based generative models, have attempted to solve this issue but often struggle with maintaining region-level color alignment. Other methods (Liu et al. 2025; Meng et al. 2025) aim to mitigate this challenge by incorporating external correspondence models to detect matching regions between the reference and target images. While effective in some cases, these methods rely heavily on the accuracy of external correspondence models. As these models are typically trained on natural images, they often lack a robust understanding of manga-style line art, making it difficult to accurately capture region-level correspondences. Consequently, incorrect region correspondence often leads to inconsistent color transfer.

We present MangaDiT, a powerful model for reference-guided line art colorization that leverages the strengths of Diffusion Transformers (DiT) (Esser et al. 2024). Built upon DiT, our model effectively captures semantic connections between spatially separated regions by modeling long-range dependencies through transformer-based attention blocks. Unlike prior methods that rely on external correspondence annotations, MangaDiT learns to discover semantically aligned regions implicitly through its internal attention mechanisms. To further enhance region-level color consistency, we introduce a hierarchical attention mechanism with dynamic attention weighting. This design augments standard spatial attention with pooled contextual information, expanding the model's receptive field and enabling more reliable color propagation across semantically similar regions.

MangaDiT's performance is evaluated on two benchmark datasets, one collected from animation videos with small character motions, and another newly created by us, specifically designed to test challenging cases with large character motions and significant pose variations. Experimental results demonstrate that our model outperforms existing state-of-the-art methods in both qualitative and quantitative assessments. We will release the code and benchmark dataset publicly upon acceptance of the paper.

In summary, our main contributions include:

- Introduction of MangaDiT, a powerful DiT-based model for reference-guided line art colorization, trained via a lightweight LoRA-based fine-tuning strategy.
- Design of a hierarchical attention mechanism and a dynamic attention weighting strategy that enhance spatial attention with pooled contextual features, improving region-level color consistency.
- Demonstration of superior performance through extensive experiments on two benchmark datasets, including challenging scenarios with large character motion.

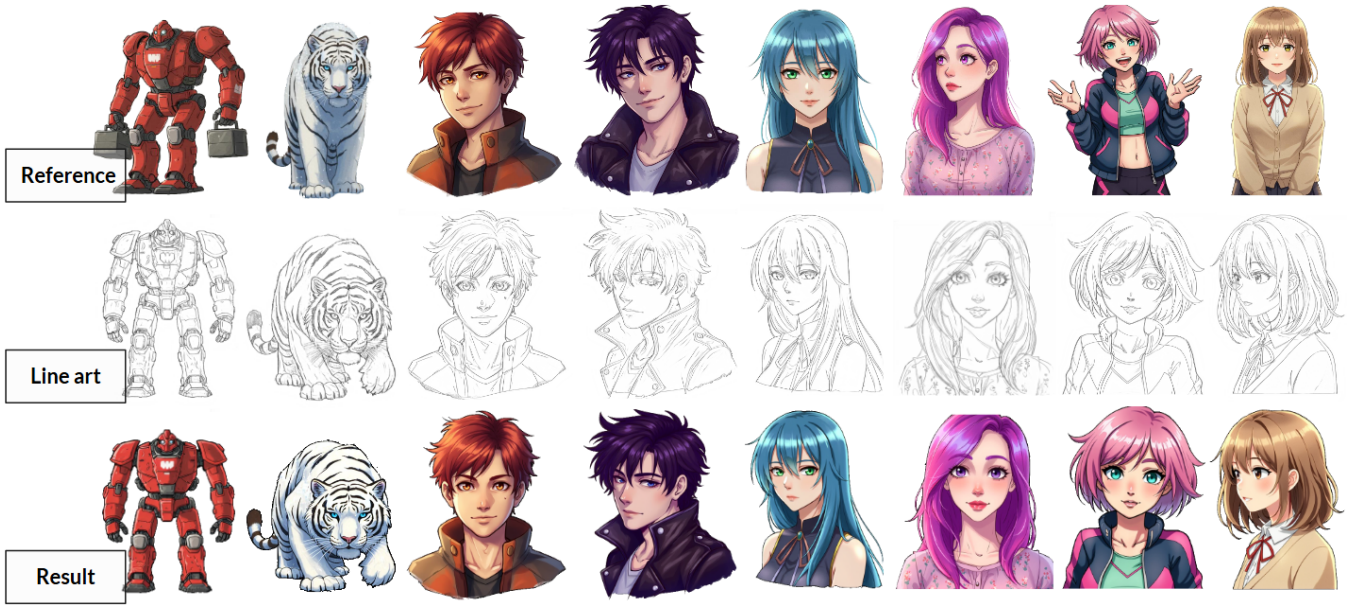

Figure 1: Reference-guided line art colorization results by MangaDiT. The samples are produced by image generation model.

Related Work

Reference-based Line Art Colorization

Line art colorization aims to fill the blank regions of a line drawing with appropriate colors while preserving structural details and stylistic coherence. This task plays an important role in manga and anime production workflows. Various user-guided colorization methods have been explored, including text prompts (Kim et al. 2019), scribbles (Ci et al. 2018; Carrillo et al. 2023), and reference images (Yan et al. 2023; Cao et al. 2024; Cao, Tian, and Mok 2023). Among them, reference-guided methods have become increasingly popular due to their intuitive and comprehensive nature of guidance. Recently, several works like BasicPBC (Dai et al. 2024) and ColorizeDiffusion (Yan et al. 2025a,b) have employed diffusion-based architectures for reference-guided colorization, leading to improved visual quality. However, these methods still struggle to achieve consistent region-level color transfer, particularly in the presence of character pose variations. To address this, some methods like MangaNinja (Liu et al. 2025) and AniDoc (Meng et al. 2025) attempt to improve alignment by integrating external point-to-point correspondence modules, such as LightGlue (Lindenberger, Sarlin, and Pollefeys 2023), which detect matched keypoints between reference and target images. Nevertheless, these correspondence modules are typically pre-trained on natural image datasets and often perform suboptimally on manga-style line drawings due to domain gaps.

In contrast, our method avoids relying on such unreliable correspondences. Instead, we directly model region-level alignment through the internal attention mechanism of the network, enabling more robust and consistent colorization results.

Conditional Diffusion Models

Diffusion models have become a cornerstone of modern generative modeling in image synthesis, offering strong flexibility to incorporate various conditional inputs during the denoising process. Early conditional diffusion models, such as Stable Diffusion (Rombach et al. 2022), introduce a latent-space formulation where images are generated in a compressed latent space. These models typically use a U-Net backbone as the denoising network, with cross-attention layers enabling conditioning on text prompts. To further enhance controllability, ControlNet (Zhang, Rao, and Agrawala 2023) is proposed to inject structural conditions by training parallel control branches within the U-Net. Recently, Diffusion Transformers (DiT) (Esser et al. 2024) replace the U-Net backbone with pure transformer blocks, achieving superior performance in modeling long-range semantic dependencies. Building on this, OminiControl (Tan et al. 2025) explores how to integrate diverse conditional signals into DiT-based architectures. It systematically compared two paradigms: ControlNet-style structural branching and parameter-efficient fine-tuning methods such as LoRA (Hu et al. 2022). Their results show that in transformer-based diffusion models, LoRA-based tuning achieves comparable controllability with significantly lower computational overhead.

Inspired by these insights, we build upon the DiT models and adopts a LoRA-based fine-tuning strategy to enable efficient training for reference-guided line art colorization.

Approach

Preliminary

Diffusion models operate in two stages: a forward noising process and a reverse denoising process. In the forward process, noise is gradually added to a target image x_0 over a

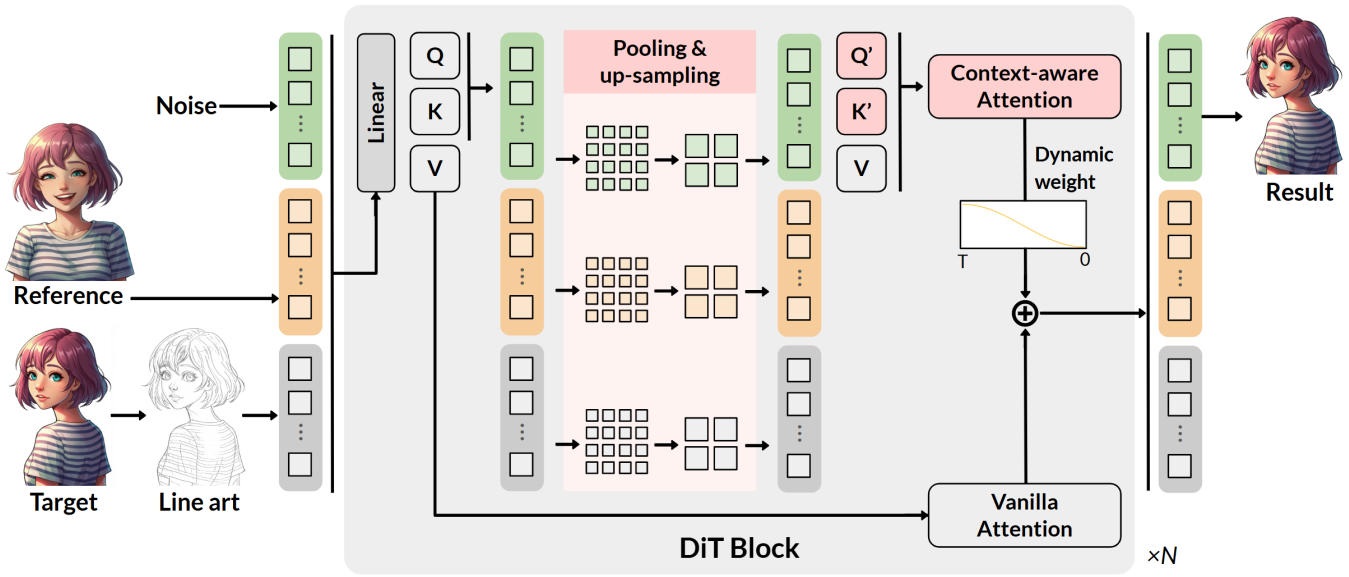

Figure 2: Overview of reference-guided line art colorization with hierarchical attention in DiT block. The empty text condition is omitted here. The modified query and key vectors (Q' and K') for context-aware attention are processed through the same pooling and upsampling procedure for each image features, which preserving the original token structure. The resulting context-aware attention is added to the vanilla attention with a dynamic weight scheduled over diffusion timesteps.

series of time steps $t \in [1, T]$, where T denotes the total number of steps. The denoising process then aims to progressively remove the noise from a noisy image x_T to generate a clean image output. The DiT model, used in architectures such as FLUX.1 (Labs 2025) and Stable Diffusion 3 (Esser et al. 2024), employs transformer blocks as the denoising network to iteratively refine noisy image tokens. During training, the target image and the text prompt are first encoded into latent tokens using frozen encoders. The noisy image tokens X are initialized from Gaussian noise in the latent space. The DiT model takes the noisy image tokens X and the text condition tokens C_T as input, and learns to denoise the image step by step through a reverse diffusion process.

The spatial representation of the image, with dimensions $N \times N$, is flattened into a sequence of N^2 tokens, which are embedded and processed by multiple transformer blocks. Each DiT block consists of layer normalization followed by multi-modal attention to capture spatial information. The multi-modal attention mechanism projects the position-encoded tokens into query Q, key K, and value V representations, allowing attention computation across all tokens of multiple conditional inputs.

Due to its modular design, the DiT architecture is highly flexible and can be extended to incorporate various types of input conditions.

Conditioning Integration Strategy

We extend the DiT architecture to support multiple visual conditions for reference-guided line art colorization. As shown in Figure 2, our model takes three types of visual inputs: the noisy latent representation for the target image, its target line art, and a colored reference image. To integrate these conditions, we adopt a unified token sequence design. Specifically, the latent tokens from the noisy image X, the text prompt condition C_T, the line art condition C_L, and the reference image condition C_R are concatenated into a single sequence $[X, C_T, C_L, C_R]$. This unified token sequence is then passed through the DiT blocks using standard multi-head self-attention, enabling information exchange across all tokens without requiring architectural modifications. The vanilla attention mechanism is computed as:

$$\mathrm{A}_{\text{vanilla}}([X, C_T, C_L, C_R]) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

To efficiently adapt the pre-trained DiT model to our colorization task, we follow a parameter-efficient fine-tuning strategy by applying LoRA to the attention layers. This allows the model to learn task-specific conditioning behaviors without modifying the full set of transformer weights. Based on this setup, our model learns to discover semantic correspondences between structural features in the line art and appearance cues in the reference image.

Hierarchical Attention Mechanism

While the DiT-based baseline can transfer global color information from the reference image to the target line art, we observe that it often fails to maintain local color consistency in semantically corresponding regions, especially in detailed areas such as clothing or accessories. We attribute this limitation to the restricted receptive field of standard self-attention, which focuses on token-wise interactions and lacks explicit access to broader spatial context. To address this issue, we introduce a hierarchical attention mech-

anism that augments vanilla spatial attention with an additional context-aware attention path derived from pooled feature representations. This design enables the model to integrate coarse-level contextual cues, improving its ability to propagate colors consistently across semantically related regions. Specifically, we first extract token sequences for the noisy image, the line art image, and the reference image, then reshape each into a spatial feature map of size $N \times N$. We apply max pooling with a randomly selected kernel size from $[2, 4, 8]$ to capture coarse semantics at multiple scales. These pooled features are then upsampled back to the original resolution via nearest-neighbor interpolation. These upsampled features are projected into a separate set of query and key representations, Q' and K', and used to compute the context-aware attention:

$$\mathrm{A}_{\mathrm{context}}([X, C_T, C_L, C_R]) = \mathrm{softmax}\left(\frac{Q' K'^{\top}}{\sqrt{d_k}}\right) V \quad (2)$$

The final hierarchical attention $\mathrm{A}_{\mathrm{hier}}$ is obtained by blending the vanilla and context-aware attention with a dynamic weight λ:

$$\mathrm{A}_{\mathrm{hier}} = \mathrm{A}_{\mathrm{vanilla}} + \lambda \mathrm{A}_{\mathrm{context}} \quad (3)$$

This hierarchical attention is applied throughout the denoising process, with its influence modulated over time using a timestep-dependent weighting strategy.

Dynamic Weight with Cosine Scheduling

Although the hierarchical attention mechanism enhances region-level color consistency by introducing coarse contextual features, its contribution should not remain constant throughout the denoising process. Prior works (Choi et al. 2022; Lin et al. 2024; Cho et al. 2024) have shown that using timestep-dependent control weights in diffusion models leads to more stable and effective generation. In the denoising process, earlier timesteps (large t) correspond to high-noise states, where the model focuses on coarse structure generation. Later timesteps (small t) focus on refining fine-grained details. Therefore, the importance of coarse context should be emphasized more in early stages and gradually reduced as the model progresses toward detailed refinement. To reflect this intuition, we introduce a dynamic attention weighting strategy that adjusts the strength of the hierarchical attention over time. We define a timestep-dependent weight $\lambda(t)$ using a cosine schedule:

$$\lambda(t) = \lambda_{base} \times 0.5 \times (1 - \cos(\frac{\pi t}{T})) \quad (4)$$

Here, λ_{base} is the maximum blending weight (set to 0.1 in our experiments). This scheduling ensures that the influence of coarse-level attention diminishes at later timesteps, allowing the model to prioritize fine-grained detail preservation. We further investigate the effectiveness of the dynamic weighting strategy, and compare different weight scheduling in ablation study.

Experiments

Datasets

For the fine-tuning phase, we utilize the sakuga-42m dataset, which has been used in prior works (Liu et al. 2025; Meng et al. 2025). We extract keyframes from this dataset and construct training pairs by initially setting the frame interval between the reference and target frames to 18. If the number of matching keypoints, determined by LightGlue (Lindenberger, Sarlin, and Pollefeys 2023), is fewer than 25, we iteratively reduce the interval until a pair with sufficient correspondence is found. The line art images are estimated using an off-the-shell LineartAnimeDetector model (Zhang, Rao, and Agrawala 2023) . The final dataset consists of reference, target, and line art images. No text prompts are used in this work.

For the evaluation phase, since related works (Yan et al. 2025a,b; Liu et al. 2025) do not publicly release their evaluation datasets, we prepare two new benchmark datasets to comprehensively assess model performance. The first evaluation set is selected from the public ATD-12K (Siyao et al. 2021), where each sample contains three keyframe images. We select two of them as the reference-target pair. Line art images are generated using the same method as the training set. We select 200 triplets of reference, target, and line art images to form the ATD-test200 dataset. These samples include both foreground characters and background elements, with relatively small pose differences between frames. In addition, we manually segment each image to extract the foreground subjects, producing a variant we denote as ATD-test200-fg. To evaluate performance under more challenging motion variations, we construct a synthetic dataset using Unity and VRoid Studio. We manually create 20 distinct 3D characters and render each character in multiple poses, resulting in 200 reference-target image pairs. For each pair, we generate the corresponding line art images and remove background elements to isolate the character, forming the Unity-test200 dataset. More details about this dataset are provided in the supplementary materials. We plan to release this dataset publicly upon acceptance of the paper.

Due to copyright constraints in ATD-test200, we create several reference–target pairs generated by a text-to-image model, used solely in figures for visualization. Further details are available in the supplementary materials.

Experimental Setup

Implementation Details. Our model is built upon FLUX.1-dev (Labs 2025), a latent rectified flow transformer model. To adapt it for reference-guided line art colorization, we apply LoRA to the attention layers, using a default rank of 4. All experiments are conducted on a single NVIDIA A100-80GB GPU for 50,000 iterations. Training takes about 36 hours with a batch size of 1 and gradient accumulation over 8 steps. For evaluation, we compare both the final checkpoint at step 50,000 and the Exponential Moving Average (EMA) version selected based on evaluation performance, and report results from the better-performing model.

Compared methods. We compare our method with four recent state-of-the-art approaches for reference-guided line art colorization: BasicPBC (Dai et al. 2024), ColorizeDiffusion v1.0 (Yan et al. 2025b) (denoted as ColDiff1.0), ColorizeDiffusion v1.5 (Yan et al. 2025a)(denoted as ColDiff1.5), and MangaNinja (Liu et al. 2025). All methods are evaluated us-

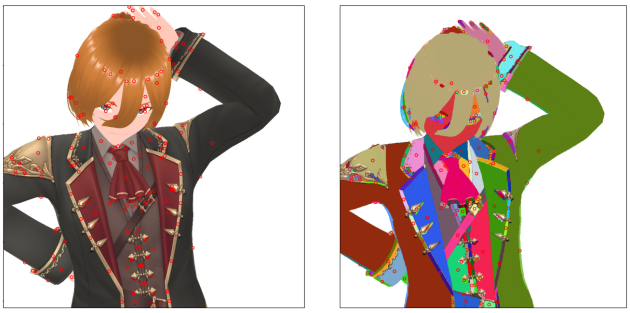

Figure 3: A sample of segmented color regions. The left image is the ground truth image and the right image is the segmented output with color regions. The segmented regions are denoted with different colors and the red point is the represented pixel in the color region.

Table 1: Quantitative results on ATD-test200 dataset

Method	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	MSE_{CR} ↓
BasicPBC	0.755	12.7	0.227	0.591	0.119
ColDiff1.0	0.900	13.0	0.293	0.302	0.095
ColDiff1.5	0.722	8.1	0.137	0.3800	0.241
MangaNinja	0.912	13.9	0.511	0.250	0.103
$MangaDiT_{(Ours)}$	**0.965**	**27.8**	**0.944**	**0.059**	**0.004**

ing the same line art and reference image pairs with a fixed noise seed to ensure fair comparison.

Evaluation Metrics. To evaluate the quality of the generated images, we adopt the following metrics: CLIP (Shen et al. 2022) semantic image similarities (CLIP), Peak Signal-to-Noise Ratio (PSNR) (Wang et al. 2004), Structural Similarity Index Measure (SSIM) (Huynh-Thu and Ghanbari 2008), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). In addition to these standard metrics, we introduce a region-based point-wise color error metric, denoted as Mean Squared Error of Color Regions (MSE_{CR}), to more accurately assess color consistency. This metric emphasizes local color fidelity. To compute MSE_{CR}, we first segment the ground truth image into fine-grained color regions based on CIELab color differences. As illustrated in Figure 3, the segmentation separates even semantically similar regions if they are divided by line boundaries, ensuring precise region-level evaluation. Each segmented region is assigned a representative pixel, and the squared color difference between the corresponding pixels in the generated and ground truth images is calculated.

Results

Qualitative and Quantitative Results

We evaluate our method on three variants of two datasets, ATD-test200, ATD-test200-fg, and Unity-test200, to assess both colorization accuracy and robustness under varying degrees of character motion between the reference and target images. The quantitative results of these datasets are separately presented in Table 1, Table 2, and Table 3. Across all benchmarks and evaluation metrics, our method consistently outperforms existing approaches, demonstrating superior performance in scenarios involving both minor and significant pose or motion variations.

Table 2: Quantitative results on ATD-test200-fg dataset

Method	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	MSE_{CR} ↓
BasicPBC	0.825	5.86	0.218	0.560	0.329
ColDiff1.0	0.870	12.9	0.329	0.246	0.125
ColDiff1.5	0.837	9.1	0.188	0.331	0.169
MangaNinja	0.876	14.1	0.518	0.201	0.080
$MangaDiT_{(Ours)}$	**0.951**	**22.0**	**0.844**	**0.085**	**0.011**

Table 3: Quantitative results on Unity-test200 dataset

Method	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	MSE_{CR} ↓
BasicPBC	0.883	8.7	0.228	0.392	0.217
ColDiff1.0	0.936	13.8	0.403	0.226	0.102
ColDiff1.5	0.837	9.1	0.188	0.331	0.169
MangaNinja	0.896	11.3	0.327	0.277	0.131
$MangaDiT_{(Ours)}$	**0.944**	**17.3**	**0.655**	**0.163**	**0.066**

The qualitative results are shown in Figure 4, showcasing the results of our full model. As illustrated, our method yields higher-quality foreground colorization, particularly in regions with complex structures. Moreover, prior methods often fail to generate coherent background content that aligns well with the reference image. In contrast, our approach is capable of generating backgrounds that closely match the reference image, substantially enhancing visual consistency. This substantially improves the practical utility of our model in real-world applications, where both foreground fidelity and background coherence are crucial.

Ablation Study

Ablation of training strategies. We conduct a series of ablation experiments to examine how different components of our framework contribute to overall colorization performance, with a particular focus on region-level color consistency. We evaluate three model variants that differ in their use of the hierarchical attention mechanism (HierAtt) and dynamic weighting with cosine scheduling (DyW). In the first setting, only the reference image and line art are used as conditional inputs to the attention mechanism, and the hierarchical attention module is disabled. In the second setting, we enable the hierarchical attention mechanism by introducing pooled context-aware attention, but apply a constant attention weight across all diffusion timesteps. In the third setting, we integrate context-aware attention with dynamic attention weighting using a cosine schedule, allowing the influence of hierarchical attention to gradually decrease as the diffusion process progresses. To better isolate the impact of these strategies, we conduct our analysis on

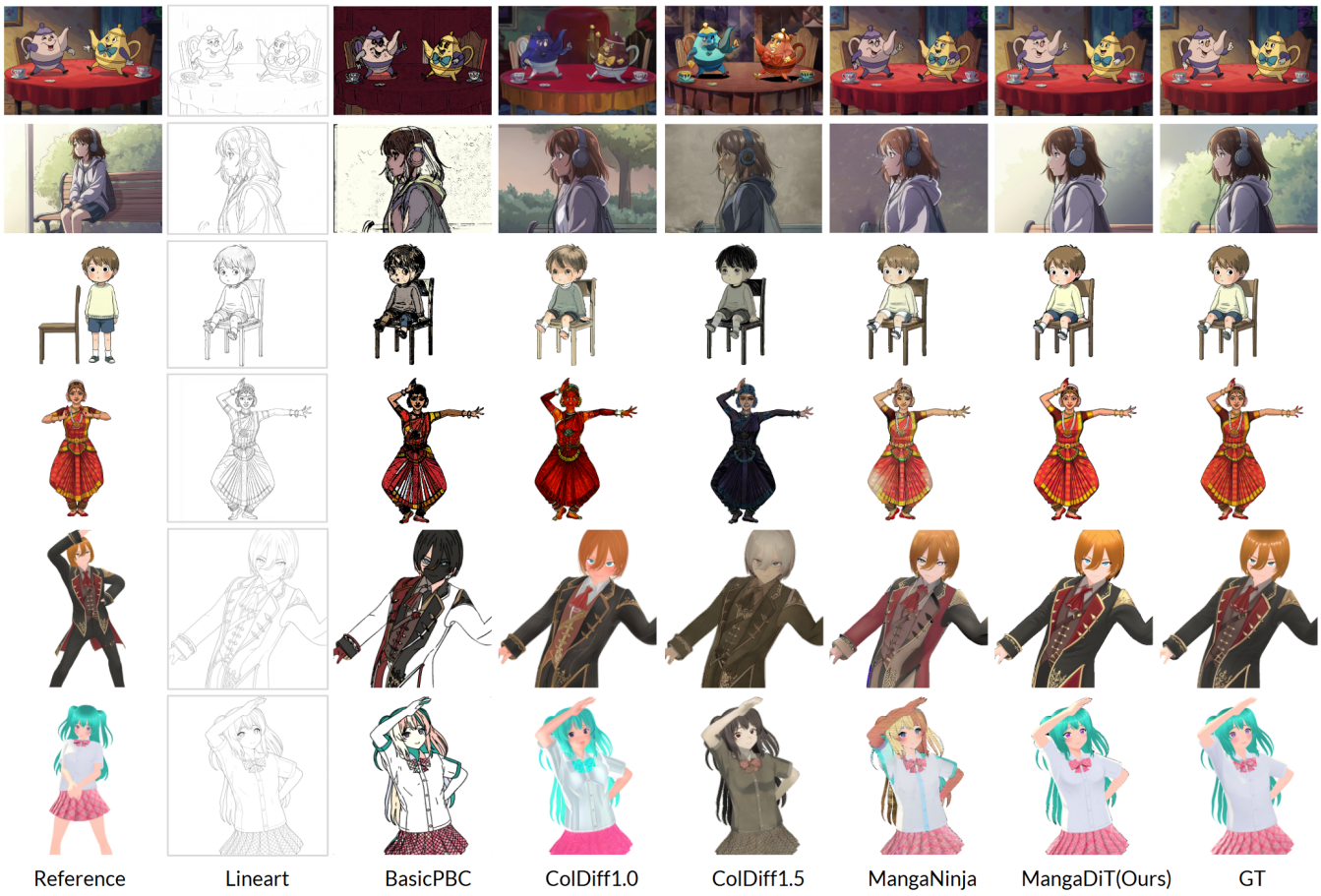

Figure 4: Qualitative comparison of line art colorization results across different methods. The first four rows show samples generated by a text-to-image model, where the top two include backgrounds and the middle two are foreground-only. The final two rows display samples from the Unity-test200 dataset.

Table 4: Ablation results of different training strategies on Unity-test200 dataset

HierAtt	DyW	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	MSE_{CR} ↓
✗	✗	0.939	15.8	0.593	0.181	0.083
✓	✗	0.938	16.6	0.621	0.176	0.072
✓	✓	**0.944**	**17.3**	**0.655**	**0.163**	**0.066**

Table 5: Ablation results of different weight scheduling on Unity-test200 dataset

Schedule	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	MSE_{CR} ↓
sin	0.936	15.7	0.585	0.185	0.084
cosInv	0.939	16.8	0.629	0.169	0.071
cos (ours)	**0.944**	**17.3**	**0.655**	**0.163**	**0.066**

the Unity-test200 benchmark, where the reference and target images exhibit significant pose and motion differences. The quantitative results are reported in Table 4, and visual comparisons are shown in Figure 5. We observe that, enabling hierarchical attention yields noticeable improvements, and further applying dynamic attention weighting leads to enhanced region-level color consistency in the generated results.

Ablation of weighting strategies. We further compare different weight scheduling methods for the dynamic attention weight to analyze how the modulation of hierarchical attention over timesteps affects performance. Specifically, we evaluate three weighting strategies: (1) sinusoidal schedule (sin), where the weight is low at both the beginning and end of the diffusion process and peaks in the middle, computed as $\lambda_{base} \times \sin(\frac{\pi t}{T})$; (2) inverse cosine schedule (cosInv), where the weight is high at early timesteps and gradually decreases, computed as $\lambda_{base} \times 0.5 \times (1 + \cos(\frac{\pi t}{T}))$; (3) our proposed cosine schedule (cos), where the weight is high at large timesteps and decreases over time. As shown in Table 5, the cosine schedule achieves the best performance across all evaluation metrics. This confirms our hy-

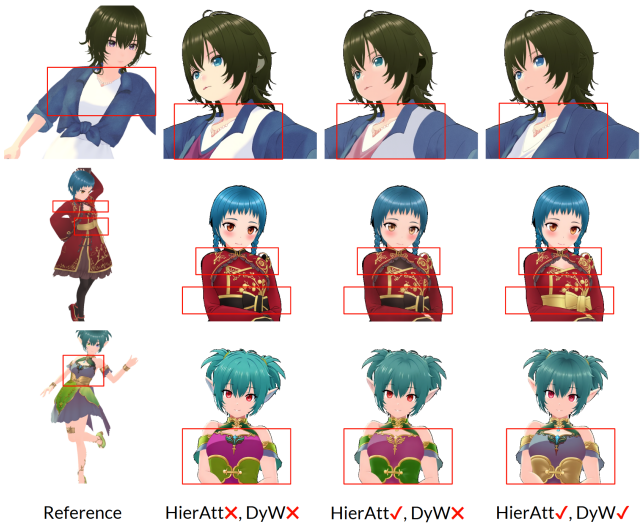

Figure 5: Qualitative comparison of our models with different training strategies.

pothesis that applying stronger hierarchical guidance in the early stages of denoising, when the model primarily focuses on global structure, helps establish better region-level correspondence. As the denoising progresses toward fine detail refinement, reducing the influence of coarse context allows the model to focus more on localized appearance features and achieve higher image quality. Additional ablation studies on attention integration variants and base weight settings are provided in the supplementary materials.

Colorization with References of Different Characters

Although our method is primarily designed for reference-guided colorization in which the reference and target line art depict the same character, we observe that the model can still generate reasonable results even when the reference image shows a different character. In such cases, the model transfers general color patterns, such as eye, hair, and clothing colors, while adapting them to fit the structural characteristics of the target line art. As shown in Figure 6, the model demonstrates robustness in handling appearance discrepancies and produces coherent, visually plausible colorization results.

Limitation

While our method achieves strong performance in reference-guided line art colorization, it is important to acknowledge the limitation arising from the absence of line structures. The model may struggle in regions where the line art does not clearly convey the underlying semantics. As shown in Figure 7, when the line drawing of a sleeve is incomplete, the model may fail to align this region with the corresponding area in the reference image, resulting in color mismatches. In addition, when the reference image contains very small regions, the model may produce inaccurate colors in the corresponding semantic areas of the target image. These ambiguities can cause errors in region matching and colorization. To mitigate such issues, we recommend providing detailed and complete line art, as well as reference images with sufficient and clearly visual information.

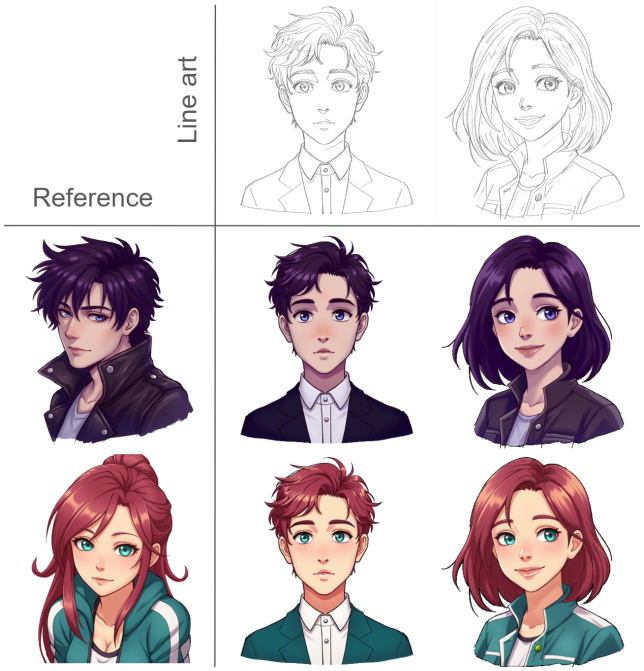

Figure 6: Colorization results with references of different characters.

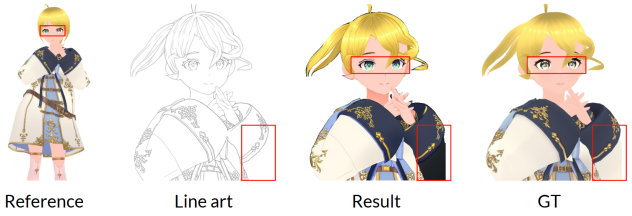

Figure 7: Error case with region limited line art.

Conclusion

In this work, we proposed a novel approach for reference-guided line art colorization using the Diffusion Transformer (DiT) architecture. By introducing hierarchical attention and dynamic attention weighting, our model effectively improves region-level color consistency, particularly under significant pose or motion variations between reference and target images. Experiments on two benchmark datasets show that our method outperforms previous state-of-the-art approaches in both quantitative and qualitative evaluations. Overall, the proposed method shows strong potential for improving the quality and efficiency of reference-guided colorization workflows, making it a practical and effective tool for digital artists and animators.

References

Cao, Y.; Meng, X.; Mok, P.; Lee, T.-Y.; Liu, X.; and Li, P. 2024. AnimeDiffusion: Anime diffusion colorization. *IEEE Transactions on Visualization and Computer Graphics*, 30(10): 6956–6969.

Cao, Y.; Tian, H.; and Mok, P. 2023. Attention-aware anime line drawing colorization. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1637–1642. IEEE.

Carrillo, H.; Clément, M.; Bugeau, A.; and Simo-Serra, E. 2023. Diffusart: Enhancing line art colorization with conditional diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3486–3490.

Cho, W.; Ravi, H.; Harikumar, M.; Khuc, V.; Singh, K. K.; Lu, J.; Inouye, D.; and Kale, A. 2024. Enhanced controllability of diffusion models via feature disentanglement and realism-enhanced sampling methods. In *European Conference on Computer Vision*, 285–301. Springer.

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11472–11481.

Ci, Y.; Ma, X.; Wang, Z.; Li, H.; and Luo, Z. 2018. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia*, 1536–1544.

Dai, Y.; Zhou, S.; Li, Q.; Li, C.; and Loy, C. C. 2024. Learning inclusion matching for animation paint bucket colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25544–25553.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, 12606–12633.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Kim, H.; Jhoo, H. Y.; Park, E.; and Yoo, S. 2019. Tag2pix: Line art colorization using text tag with secat and changing loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9056–9065.

Labs, B. F. 2025. Flux: Official inference repository for flux.1 models. https://github.com/black-forest-labs/flux. Accessed: 2025-07-07.

Lin, Q.; Sun, X.; Gao, Y.; Zhong, Y.; Li, D.; Zhao, Z.; and Wang, H. 2024. TASR: Timestep-Aware Diffusion Model for Image Super-Resolution. *arXiv preprint arXiv:2412.03355*.

Lindenberger, P.; Sarlin, P.-E.; and Pollefeys, M. 2023. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, 17627–17638.

Liu, Z.; Cheng, K. L.; Chen, X.; Xiao, J.; Ouyang, H.; Zhu, K.; Liu, Y.; Shen, Y.; Chen, Q.; and Luo, P. 2025. MangaNinja: Line Art Colorization with Precise Reference Following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Meng, Y.; Ouyang, H.; Wang, H.; Wang, Q.; Wang, W.; Cheng, K. L.; Liu, Z.; Shen, Y.; and Qu, H. 2025. Anidoc: Animation creation made easier. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18187–18197.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, S.; Li, C.; Hu, X.; Xie, Y.; Yang, J.; Zhang, P.; Gan, Z.; Wang, L.; Yuan, L.; Liu, C.; et al. 2022. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35: 15558–15573.

Siyao, L.; Zhao, S.; Yu, W.; Sun, W.; Metaxas, D.; Loy, C. C.; and Liu, Z. 2021. Deep animation video interpolation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6587–6595.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2025. OminiControl: Minimal and Universal Control for Diffusion Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Yan, D.; Ito, R.; Moriai, R.; and Saito, S. 2023. Two-Step Training: Adjustable Sketch Colourization via Reference Image and Text Tag. In *Computer Graphics Forum*, volume 42, e14791. Wiley Online Library.

Yan, D.; Wang, X.; Li, Z.; Saito, S.; Iwasawa, Y.; Matsuo, Y.; and Guo, J. 2025a. Image Referenced Sketch Colorization Based on Animation Creation Workflow. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23391–23400.

Yan, D.; Yuan, L.; Wu, E.; Nishioka, Y.; Fujishiro, I.; and Saito, S. 2025b. ColorizeDiffusion: Improving Reference-Based Sketch Colorization with Latent Diffusion Model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5092–5102. IEEE.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.