Sketch-Guided Anime Hair Editing Using Multimodal Diffusion Transformer (Student Abstract)

Tianyu Zhang[1], **I-Chao Shen**[2], **Haoran Xie**[1, 3*]

[1]Japan Advanced Institute of Science and Technology, 923-1292, Ishikawa, Japan
[2]The University of Tokyo, 113-8654, Tokyo, Japan
[3]Waseda University, 169-8050, Tokyo, Japan
s2320434@jaist.ac.jp, jdilyshen@gmail.com, xie@jaist.ac.jp

Abstract

Anime hair design is crucial but challenging, as it conveys personality and emotion through stylized geometry and layered structure. In this work, we propose a sketch-guided approach for intuitive control of multimodal diffusion transformers (MMDiT) to generate semantically consistent anime hairstyles. We adopt a wisp-level flowline input integrated with a fine-tuned MMDiT to transfer hairstyles while preserving character identity. We believe that this fine-grained sketch control within the MMDiT framework may offer a promising path for structured anime hair editing.

Introduction

In anime character design, hair is not merely a decorative element but a crucial visual component used to convey the personality and emotions of a character. The highly stylized structure, distinct highlights, and complex layering make anime hair editing particularly challenging. With the advancement of deep learning, methods based on Generative Adversarial Networks (GANs) (Luo, Xie, and Miyata 2021) have attempted to edit anime hairstyles using masks, but the generated results remain unsatisfactory. In recent years, diffusion models have demonstrated impressive capabilities in generating high-quality images, and diffusion-based methods for anime hair editing have shown promising results. In particular, Flux Kontext (Batifol et al. 2025), a state-of-the-art multimodal diffusion transformer for text-based editing, provides a natural way to edit images with semantic consistency. However, an intermediary mechanism is required between creators and models to serve the dual functions of intention transmission and outcome regulation.

Sketches serve as an intuitive and accessible form of expression that can be effectively interpreted by models while also aligning closely with the traditional creative practices of anime artists. In previous studies, sketch has often been used for structural editing of human hair (Xiao et al. 2021). Human hair comprises thousands of fine, physically coherent strands with complex global interactions. By contrast, anime hair is highly stylized and abstract. Instead of depicting individual strands, artists group them into a small number of coherent wisps or locks, outlined by clean sketch strokes and shaded with simplified color regions. This design yields an exaggerated yet internally consistent structure that prioritizes character identity and aesthetic appeal over physical fidelity. Consequently, a substantial domain gap arises: human hair exhibits fine-scale geometry, randomness, and physical realism, whereas anime hair emphasizes visual clarity, iconic silhouettes, and wisp-level grouping.

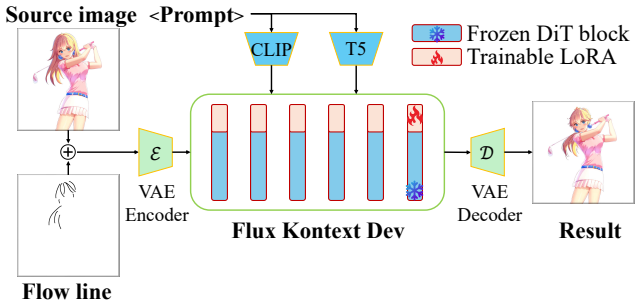

Figure 1: Pipeline of the proposed sketch-guided anime hair editing method.

To address the above issues, we propose an anime hair editing method that transfers hairstyles to the source image under sketch guidance while preserving the character's identity. Inspired by the anime hair creation process, we abstract a flow line structure, where each line represents an anime hair wisp, starting from the hair whorl and ending at the tip. We use this flow line structure as a guiding sketch input to the fine-tuned Multimodal Diffusion Transformer model for hairstyle editing of the source image. This is the first attempt to introduce fine-grained sketch control into the Flux Kontext framework for structured editing of anime hairs.

Method

This work aims to control Flux Kontext for sketch-guided hairstyle generation. First, we concatenate the source image and sketch as input and insert Low-Rank Adaptation (LoRA) modules (Hu et al. 2022) into each layer for fine-tuning. The transformer is trained with a flow-matching loss, enabling the model to follow the sketch guidance while preserving character identity.

We collect images containing only a single character from

*Corresponding author

Danbooru[1], remove their backgrounds, and resize them to 1024×1024 to construct the ground truth of our dataset. We further employ Flux Kontext to generate diverse edited results through text-based editing, and manually select the best outputs as the source images. We then invited three creators to annotate flow lines on the ground truth images as sketch guidance for initial training. In total, we obtained a dataset of 1,173 triplets source image, sketch, ground truth, with 973 for training and 200 for testing.

As shown in the Figure 1, the source image and sketch are height-aligned, horizontally concatenated, and resized to the nearest Kontext-preferred resolution. The sketch is concatenated with the source image and fed into Flux Kontext (Batifol et al. 2025), and LoRA modules are added at each layer to fine-tune the model. We encode the target image with the VAE and fine-tune the transformer using a flow-matching loss (Esser et al. 2024) with uniform weighting in the style of Stable Diffusion (SD) 3. Concretely, the target image x is encoded by the VAE into latents z (using the pipeline's shift/scaling factors). For a randomly sampled scheduler timestep t with corresponding noise level σ, we form

$$z_t = (1-\sigma)z + \sigma\varepsilon, \varepsilon \sim \mathcal{N}(0, I) \quad (1)$$

and we minimize a uniformly weighted mean-squared error. To minimize the influence of prompts on the model, we set the prompt during training to *"<trigger token> follow the hair layout on the right; preserve character identity"*, thereby emphasizing the structural guidance provided by the sketch.

Results

To validate the effectiveness of the proposed method, we conducted both qualitative and quantitative comparisons on the test set. We selected SD-v1.5 (Rombach et al. 2022) as the base model, and adopted its retrained variant, Anything-v4.5 Inpainting, which is specialized for anime and supports region editing via masks. To support inputs compatible with Anything-v4.5 Inpainting, we generated the masks from the segmented ground truth hair region, and used Qwen2-vl (Wang et al. 2024) to generate prompts describing the target hairstyle. All models were run with 28-step inference to generate 1024×1024 resolution results.

As shown in Figure 2 and Table 1, our method edits the source hairstyle in accordance with the sketch-specified structure while preserving character identity. In particular, as highlighted by the orange box in the second row, our proposed flow-line representation captures fine-grained hairstyle design. By contrast, mask-based editing excels at pixel-level object manipulation, but is not fully adequate for fine-grained detail editing. The mask-based Anything-v4.5 Inpainting baseline has limited influence beyond the masked region, leaving large portions of the original hairstyle unchanged (see the red box in the fourth row). Moreover, our approach achieves higher perceptual quality and more consistent style than the compared baseline.

[1]https://danbooru.donmai.us/

Method	FID(↓)	LPIPS(↓)	CLIP Aesthetic (↑)
Anything-v4.5	77.287	0.198	6.011
Ours	**47.295**	**0.116**	**6.466**

Table 1: Quantitative comparison between the Anything-v4.5 Inpainting (Rombach et al. 2022) and our method.

Figure 2: Visualization comparison of our method and stable diffusion (Anything-v4.5 Inpainting), our method achieved superior character consistency and quality.

Conclusion

We proposed the first sketch-guided pipeline of the multimodal diffusion transformer to edit anime hairstyles, in which each line segment of the sketch guides the flow line of a single hair wisp. The experiments demonstrated that the method produces high-quality edits that are consistent with anime stylistic conventions. For future work, we plan to enrich the conditioning from pixel-level sketches to SVG representations that carry explicit directional information, enabling hairstyles that better follow the intended flow. We also intend to attach color to individual lines, allowing users to specify localized color for wisps.

Acknowledgements

This paper is based on results obtained from GENIAC (Generative AI Accelerator Challenge, a project to strengthen Japan's generative AI development capabilities), a project implemented by the Ministry of Economy, Trade and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO). Additionally, this work was supported by the JST BOOST Program (Japan), Grant Number JPMJBY24D6.

References

Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv e-prints*, arXiv–2506.

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 *Forty-first international conference on machine learning*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Luo, S.; Xie, H.; and Miyata, K. 2021. Sketch-based anime hairstyle editing with generative inpainting. In *2021 Nicograph International (NicoInt)*, 7–14. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiao, C.; Yu, D.; Han, X.; Zheng, Y.; and Fu, H. 2021. SketchHairSalon: Deep Sketch-based Hair Image Synthesis. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2021)*, 40(6): 1–16.