# Towards High-resolution and Disentangled Reference-based Sketch Colorization

## Supplementary Material

## 1. Plugin module and injection

We use the plugin module to extract low-level features from the background latent and inject them into the denoising backbone through split attention injections, which means we give different KV inputs for foreground and background in self-attention and cross-attention layers when activating the plugin module. The split self-attention injection can be formulated as:

$$y_{\text{sa}} = \begin{cases} \text{Attention}(z_f, z, z) & \text{if } m > ts \text{ (Foreground)} \\ \text{Attention}(z_b, z_{\text{inj}}, z_{\text{inj}}) & \text{if } m \leq ts \text{ (Background)}, \end{cases} \quad (1)$$

where $z, z_f, z_b, z_{\text{inj}}, m, ts$ are the forward features, foreground tokens of forward features, background tokens of forward features, and injection features from the plugin module, the sketch foreground mask, and the user-defined foreground threshold, respectively. Users can adjust $ts$ to adjust the selected foreground regions or manually give a foreground mask. We perform the same split for the cross-attention. The only difference is that we jointly train background LoRAs for the QKV projection layers used in cross-attention layers. And the split cross-attention is organized as:

$$y_{\text{ca}} = \begin{cases} \textbf{Softmax}(\frac{W^q z_f)\cdot(W^k e)}{\sqrt{d}})(W^v e) & \text{if } m > ts \\ \textbf{Softmax}(\frac{(\hat{W}_b^q z_b)\cdot(\hat{W}_b^k e_b)}{\sqrt{d}})(\hat{W}_b^v e_b) & \text{if } m \leq ts. \end{cases} \quad (2)$$

Here, $e, e_b$ are embeddings from reference images and reference background embeddings, respectively, and $\hat{W}_b^*$ denotes the LoRA-injected linear weight, with the LoRA rank set to 16.

Both split attentions are implemented through **masked attention**.

## 2. Failure case and solution

### 2.1. Description

As discussed in the main text, the proposed Gram loss successfully disentangles spatial semantics by ensuring that the structure of the sketch-guided region (defined as the foreground) is derived exclusively from the sketch inputs. However, we observe limitations stemming from **biased training data**. Specifically, when a reference image lacks background details, the network tends to hallucinate random content in the non-sketch regions. These failure cases are visualized in Figure 1.


Figure 1. Sketch images inherently possess rich background spatial semantics, as evidenced by our sketch-only results where no reference images are provided. By incorporating Gram loss, the network is guided to more faithfully follow the spatial semantics of the sketches. Consequently, the model tends to synthesize backgrounds when the reference image offers negligible background spatial semantics, reflecting the fact that the majority of sketch images in our training data contain background details.

### 2.2. Solution and discussion

This background issue can be mitigated by **activating the plugin module**, as it enables low-level feature transfer for the image. In future work, we aim to address the root cause by refining our training data and strategy for robust T2I generation.

Crucially, this challenge is distinct from the spatial entanglement addressed in prior works [2, 3], which stems from foreground embedding leakage and manifests as body overflow and segmentation deterioration, visualized in the w/o Gram loss column in Figure 1. We observe that such entanglement artifacts persist—even with split-attention [2] or our proposed plugin module—at the high resolutions (>1024px) used in this study. This persistence implies a fundamental limitation in the network's ability to differentiate between foreground and background tokens.

## 3. More baseline for cross-content colorization comparison

Nano Banana is also applicable to reference-based sketch colorization. While it struggles with character sketches due to imprecise regional color transfer, it proves effective for cross-content colorization. In Figure 2 and 3, we present additional qualitative results for Nano Banana and IP-Adapter. These serve as extended baselines for the comparisons shown in Figure 9 of the main manuscript.

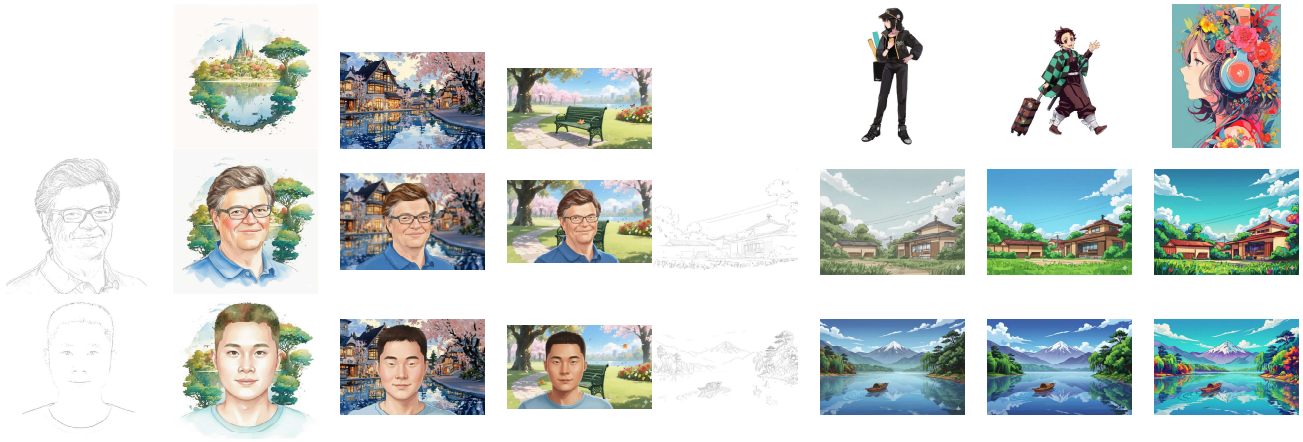

Figure 2. Cross-content results using Nano Banana.

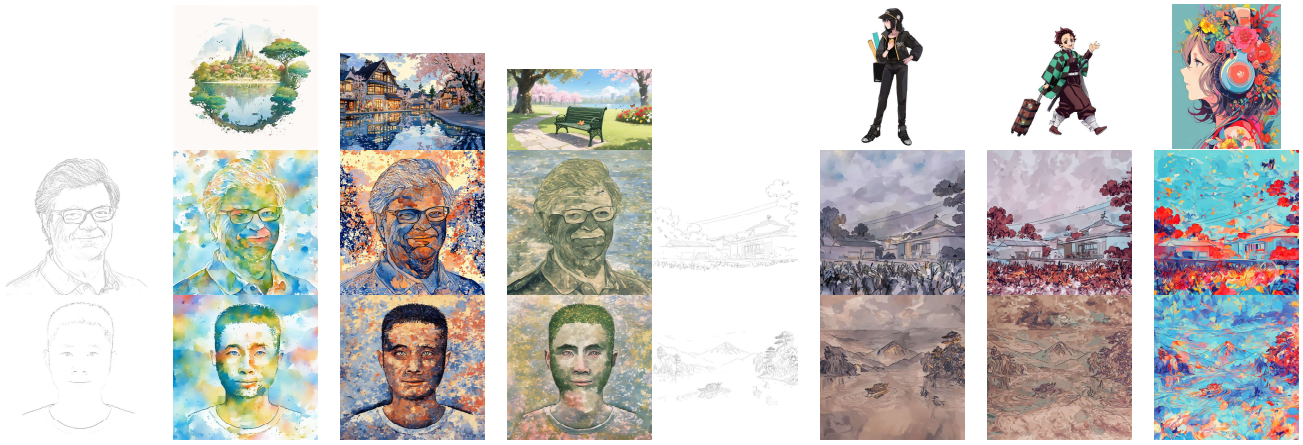

Figure 3. Cross-content results using ControlNet + IP-Adapter [4, 5] with style transfer weights [1].

## 4. User Study

In the user study, we present 25 sets of sketch-reference pairs and results generated by all compared methods to users investigated. We show the user interface in Figure 4, and the images shown in the user study in Figure 5, Figure 6, and Figure 7.

## References

[1] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2

[2] Dingkun Yan, Xinrui Wang, Zhuoru Li, Suguru Saito, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Image referenced sketch colorization based on animation creation workflow. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23391–23400, 2025. 1

[3] Dingkun Yan, Liang Yuan, Erwin Wu, Yuma Nishioka, Issei Fujishiro, and Suguru Saito. Colorizediffusion: Improving reference-based sketch colorization with latent diffusion model. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 5092–5102, 2025. 1

[4] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721, 2023. 2

[5] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2

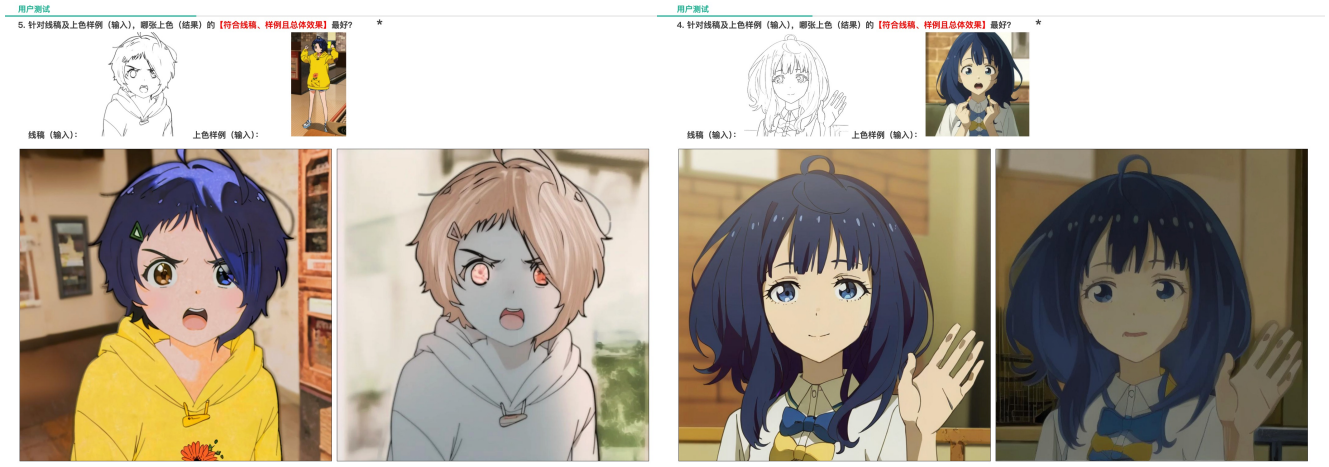


Figure 4. The user interface employed during the perceptual study. Participants were presented with input line art and a reference color sample, then asked to evaluate the generated results based on structural fidelity, style consistency, and overall visual quality.


Figure 5. Images shown during the user study.


Figure 6. Images shown during the user study.


Figure 7. Images shown during the user study.