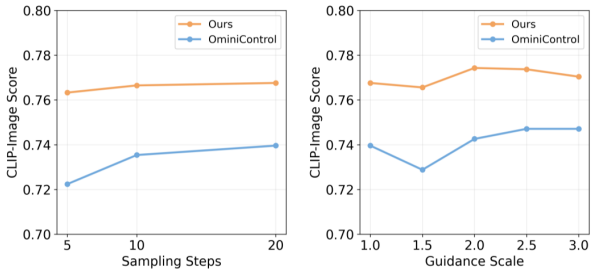
Gated Condition Injection without Multimodal Attention: Towards Controllable Linear-Attention Transformers

Supplementary Material

Figure 1. Gated control on the original OminiControl. Our approach enables the model to capture control signals much earlier during training, even when using softmax attention.



(a) CLIP-Image w.r.t. steps. (b) CLIP-Image w.r.t. CFG.

Figure 2. Robustness to sampling steps and guidance scale. Our model produces better and more stable outputs than OminiControl under both low-step inference and varying guidance scales.

A. Gated control on OminiControl

To further validate the generality of our gated mechanism on the softmax attention, we implement a minimal variant on top of the original OminiControl (based on the FLUX.1-dev) by introducing a single-block gated interaction between image-condition tokens and latent tokens, adding only 0.2M parameters. Even with this lightweight modification, our approach enables the model to grasp control information much earlier during training. Figure 1 presents the test-time outputs at step 400 under strictly identical training configurations. With the addition of a lightweight gated control mechanism, the model achieves significantly faster alignment with the vase specified in the condition input.

B. Visual analysis of ablations

We further provide a more intuitive analysis of the ablation on gated application through visualizations, as shown in Figure 4. Without gating, the model's ability to leverage conditional information is substantially impaired, which may lead to outputs that do not comply with the condition

Figure 3. Further visual comparisons on the subject-driven tasks. Our approach preserves object-specific features with greater fidelity, while simultaneously adapting the environment according to the provided editing prompt in a natural and flexible manner.

or generate unreasonable artifacts, such as unnatural protrusions on a horse's back, and a decrease in overall image quality. Moreover, using input features from different sources to generate the gate scores can affect global aspects of the image, such as style and color. This observation suggests that during the model's forward pass, input features from different positions may exhibit a certain degree of semantic progression, for example, a growing sensitivity to

Method	Identity preservation	Material quality	Color fidelity	Natural appearance	Modification accuracy	Average score
IP-Adapter (SANA)	24.8	30.4	37.2	56.1	44.8	38.7
Ours	**52.9**	**63.5**	**58.4**	**72.4**	**55.6**	**60.6**

Table 1. Quantitative evaluation results (in percentage) across different evaluation criteria. Higher values indicate better performance.

Figure 4. Intuitive analysis of the ablation on gated application through visualizations. Without gating, the model's ability to leverage conditional information is substantially impaired, which may lead to outputs that do not comply with the condition or generate unreasonable artifacts. Moreover, using input features from different sources to generate the gate scores can affect global aspects of the image, such as style and color.

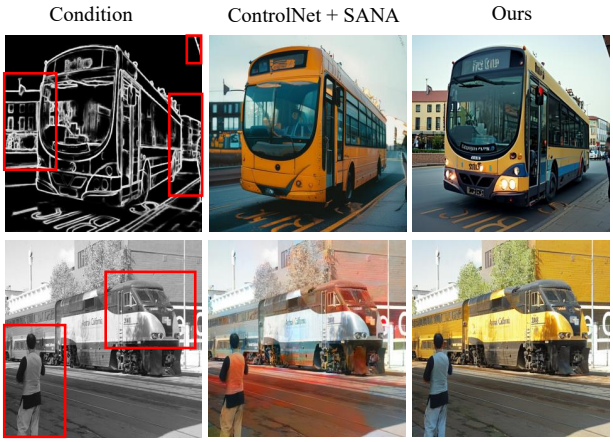


Figure 5. Comparison between our approach and SANA-based ControlNet. Our method more accurately follows the conditions and can generate more natural and realistic colorization.

style and color.

C. Sampling steps and guidance scale

We further evaluate the robustness of our model with respect to sampling steps and guidance scale. The results in Figure 2 indicate that our model produces **better** and more **stable** outputs than OminiControl under both low-step inference and varying guidance scales, further supporting the suitability for low-latency scenarios.

D. Image editing and multi-condition

GateControl can acquire basic editing capability within limited training steps, as shown in Figure 6(a), demonstrating fast adaptation and broad applicability. We further conduct multi-condition experiments (subject + depth), as depicted in Figure 6(b), showing that our model can simultaneously incorporate multiple conditions. However, when multiple conditions are combined, conflicts may arise; for example, satisfying geometric constraints may slightly alter the original subject shape.

E. Comparison on the subject-driven tasks

E.1. Evaluation for subject-driven generation

Evaluation criteria. Following OminiControl, we utilize a five-dimensional evaluation protocol that systematically measures both fidelity to subject characteristics and compliance with user-specified modifications. The assessment dimensions comprise Identity Preservation, Material Quality, Color Fidelity, Natural Appearance, and Modification Accuracy. Evaluations are conducted using the GPT-4o multimodal model.

Results comparison. Results compared with the SANA-based IP-Adapter are presented in Table 1. These results demonstrate that our method substantially surpasses IP-Adapter on subject-driven generation. It achieves markedly stronger preservation of object-specific attributes, provides more flexible and faithful control over user-specified modifications, and produces outputs that exhibit more natural integration with the surrounding context.

(a) Image Editing: Golden sunlight.

(b) Multi-Condition: Sitting on a sofa..

Figure 6. Examples of image editing and multi-condition control. Our model can acquire basic editing capability within limited training steps. Moreover, it is able to simultaneously incorporate multiple conditions.

E.2. Further visual comparisons

Figure 3 illustrates the qualitative differences between our method and the SANA-based IP-Adapter. Our approach preserves object-specific features with greater fidelity, while simultaneously adapting the environment according to the provided editing prompt in a natural and flexible manner, significantly surpassing the IP-Adapter results. Furthermore, our approach is generalizable: the same architecture can handle both subject-driven and spatially-aligned tasks, with only LoRA and gating parameters adjusted.

F. Comparisons on spatially aligned tasks

Figure 5 presents a comparison between our approach and SANA-based ControlNet. Our method more accurately follows the conditions, generating surrounding objects like vehicles and buildings more faithfully, and produces higher-quality outputs for the bus. For the coloring task, our model achieves more natural and realistic colorization.