# NumeriKontrol: Adding Numeric Control to Diffusion Transformers for Instruction-based Image Editing

Supplementary Material

## A. User Study

To further demonstrate the effectiveness of NumeriKontrol, we conduct a user study comparing five methods: NumeriKontrol, Kontinuous [33], FLUX.1 Kontext [19], Nano Banana, and Seedream 4.0 [40]. We sample 30 examples for each method. As described in Sec. 5.2, Nano Banana and Seedream do not support specific numerical instructions. All instructions are adjusted and tested to ensure consistent evaluation across methods.

Following standardized evaluation protocols, we anonymize all samples by removing method identifiers and randomize their presentation order. The study is conducted through an online questionnaire distributed to participants. Participants view examples from all five methods and rate each sample on a 1-5 Likert scale (allowing one decimal place). The evaluation includes three criteria: (1) *Is the image edited according to the instruction?* (2) *Does the edited result match the numeric information in the instruction?* (3) *How much do you prefer the edited result?* These questions correspond to success rate, numerical alignment, and user preference, respectively.

Table 3. User study with state-of-the-art methods. The best score is **emphasized**.

| Method | Success (↑) | Alignment (↑) | Preference (↑) |
|---|---|---|---|
| FLUX.1 Kontext [19] | 3.64 | 2.70 | 3.57 |
| Kontinuous Kontext [33] | 4.75 | 4.24 | 3.85 |
| Nano Banana | 4.42 | 3.25 | 3.62 |
| Seedream [40] | 4.31 | 3.09 | 3.78 |
| NumeriKontrol(Ours) | **4.80** | **4.57** | **4.08** |

The results of the user study is shown in Tab. 3. Our method scores best in all three criteria.

## B. More Details for Comparison

### B.1. Implementation Details for Comparison

As Kontinuous Kontext [33] is implemented by ourself, the strength projector in our implementation is designed with dimensions of $1536 \rightarrow 256 \rightarrow 128 \rightarrow 6144$. The scale and shift are separated into 3072 dimensions respectively from the output of the projector. The modulation parameter is set to 1.0 in our experiments.

### B.2. Quantitative Comparison

Certain prompts may not function properly with Nano Banana and Seedream. To address this limitation, we develop adapted instruction formats for these methods. For camera-related aspects in the test set, we append *"Work as a DSLR camera"* to the instructions to enable editing. For other numeric instructions, we adopt the following format: *"Edit the image with a strength of $\delta$. The full strength is $\Delta$*, where $\delta$ represents the current editing strength and $\Delta$ denotes the maximum strength. Details of the numeric instructions are provided in Sec. 4.4. All instructions used in quantitative comparisons are standardized using these two approaches to ensure fair evaluation.

As described in Sec.5.3, LLM-based methods successfully edit images for low-level tasks. However, they fail to preserve accurate numeric information from the instructions. Consequently, these two methods perform poorly on low-level tasks. An example from the quantitative results is shown in Fig.8.

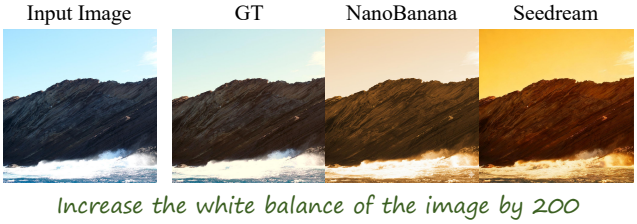


Figure 8. The low-level task of editing the white balance of the image. The white balance is divided into 400 steps, with 200 lower and 200 higher.

## C. Additional Results

Our method supports continuous editing through the online delta calculation described in Sec.4.2. However, errors may accumulate during successive operations. As illustrated in

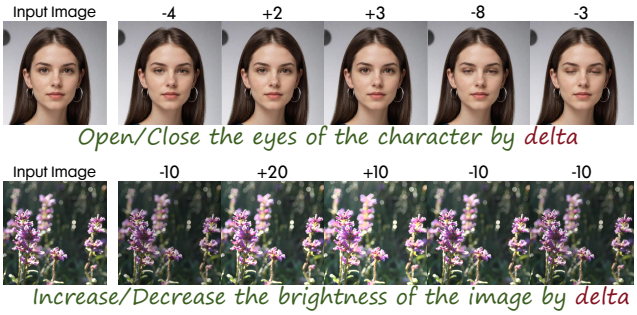


Figure 9. **Visualization of Continuous Editing.**

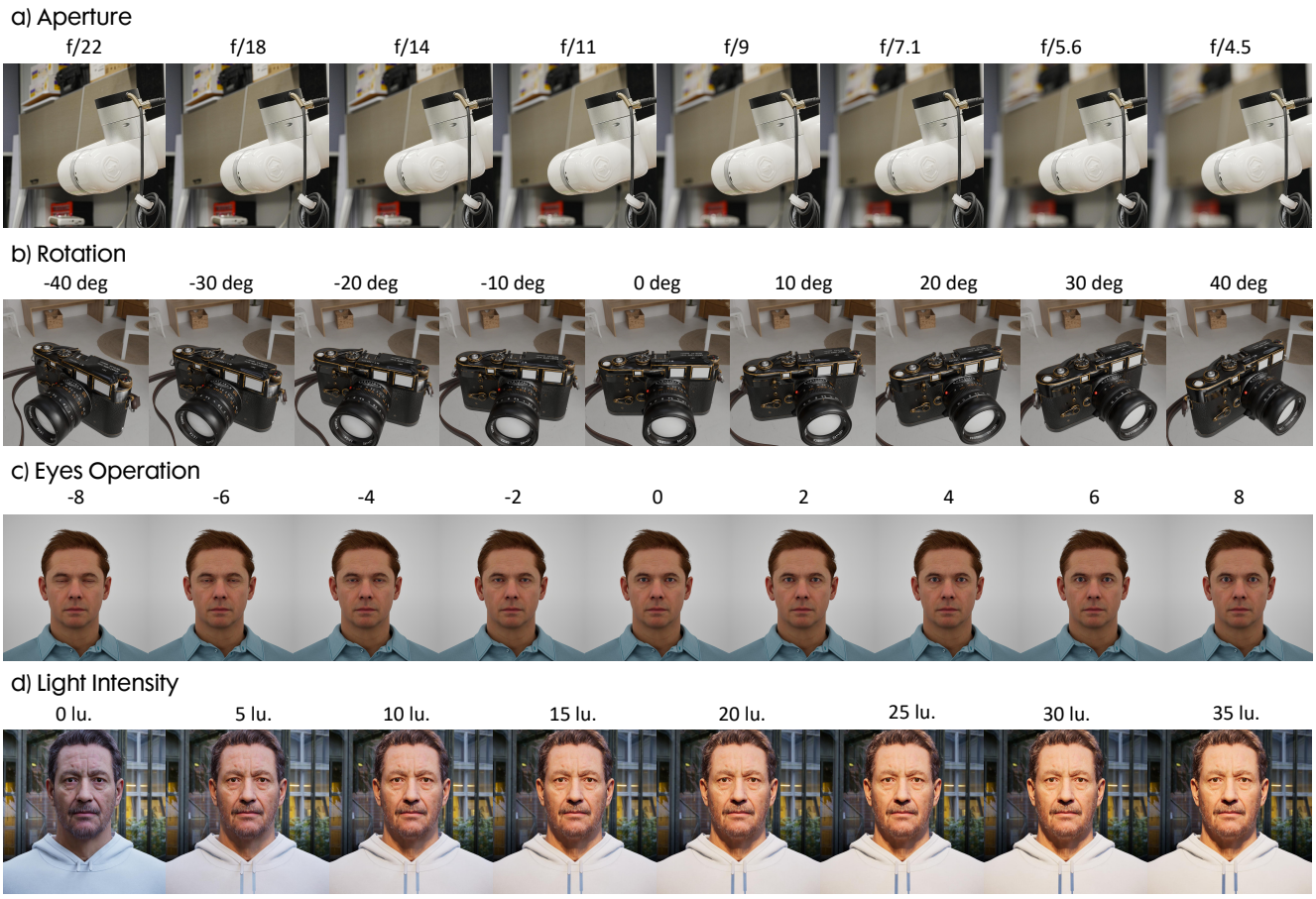


Figure 10. **Visualization of samples in CAT dataset.**

Fig.9, despite the first row's cumulative parameter changes totaling -10, the final state exhibits incomplete closure; conversely, the second row achieves zero net change yet displays reduced brightness. While continuous editing cannot guarantee absolute numerical precision, the overall directional trends remain consistent.

## D. More Details on Synthesized Dataset

We illustrate samples from our Common Attribute Transformation dataset in Fig.10. All samples are annotated with accurate attribute values. For facial expression tasks such as eye opening (Fig.10 (c)), we use a simple gray HDRI for training. This choice does not interfere with the model's generalization ability. However, for lighting-related tasks such as intensity adjustment (Fig.10(d)), we employ diverse HDRI backgrounds. This design reflects the complexity of real-world lighting conditions. Multiple HDRIs approximate these complex light sources and improve generalization performance. For object manipulation scenarios (Fig.10(b)), the object name is inserted into the instruction template. All low-level tasks follow a unified prompt format described in Sec.4.4, as shown in Fig.10 (a). All images in our dataset is taken in $1024\times1024$. During training, the samples are resized to $512\times512$.