ArtEditor: Learning Customized Instructional Image Editor from Few-Shot Examples

Supplementary Material

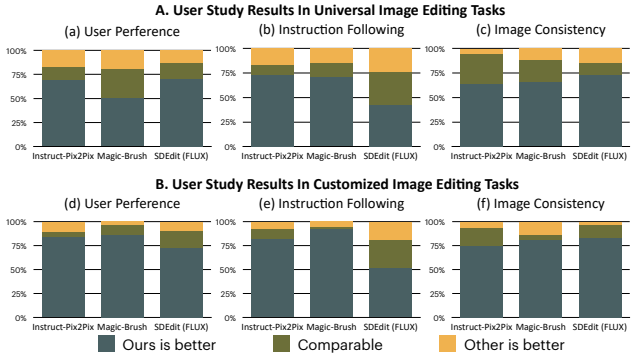

Figure 1. User study results. The scores demonstrate the percentage of users who prefer ours over others under three evaluation metrics. ArtEditor outweighs all other baselines in user study.

A. DoodleArt Benchmark

We demonstrate a brief overview of the proposed DoodleArt benchmark. The training data can be found in Fig. 2, and the evaluation data can be found in Fig. 3

B. Metrics

To comprehensively evaluate the performance of each algorithm, we assess four key aspects: *1. Text Consistency:* We use the CLIP-Score[1] to evaluate the consistency between the generated images and the input text. *2. Editing Performance:* We use the $GPT\ score$ to test how SOTA vllms recognize the editing results. The gpt score is calculated with the following prompt:

From 0 to 100, how much do you rate for EDIT TEXT in terms of the correct and comprehensive description of the change from the first given image to the second given image? Correctness refers to whether the text mentions any change that are not made between two images. Comprehensiveness refers to whether the text misses any change that are made between two images. The second image should have minimum change to reflect the changes made with EDIT TEXT. Be strict about the changes made between two images: 1. If the EDIT TEXT is about stylization or lighting change, then no content should be changed and all the details should be preserved. 2. If the EDIT TEXT is about a local change, then no irrelevant area nor image style should be changed. 3. The first image should not have the attribute described inside the EDIT TEXT, rate low, (¡80) if this happens 4. Be aware to check whether the second image does maintain the important attribute in the left image that is not reflected in the EDIT TEXT. Rate low (¡50) if two images are not related.

3. Image Consistency: For the subject condition, we use CLIP-I[1] to compute the cosine similarity between image embeddings extracted by the CLIP image encoder for both generated and reference images.

C. User Study

We conducted a user study with 30 participants via online questionnaires. We evaluated user preferences in both general and customized image editing scenarios. Participants were presented with ArtEditor's outputs alongside baseline methods, and asked to evaluate which results they preferred based on three criteria: 1) Overall preference, 2) Instruction following, and 3) Consistency between the edited images and the original images. During the study, participants viewed the original unedited images, the edit instructions, and reference images edited by models. They were then asked to decide whether ArtEditor (Option A) or a baseline method (Option B) performed better, or if they were about equally effective. The results of this user study are collected in Fig. 1, where we reported the percentage scores of each criterion, highlighting our method's effectiveness in aligning closely with artistic intentions and maintaining high consistency in edits without introducing unwanted changes.

D. Additional Experiments

D.1. Diverse Style Generalization

D.2. Customized Editing Comparison

E. Limitation and Future Work

One limitation of ArtEditor is its dependence on the collection of dozens of paired datasets (pre-edit and post-edit images) and the need for thousands of training steps using LoRA. This data collection process can be challenging, as paired images are not always readily accessible. In the future, we will attempt to learn doodling strategies from single image pairs using an Encoder structure.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

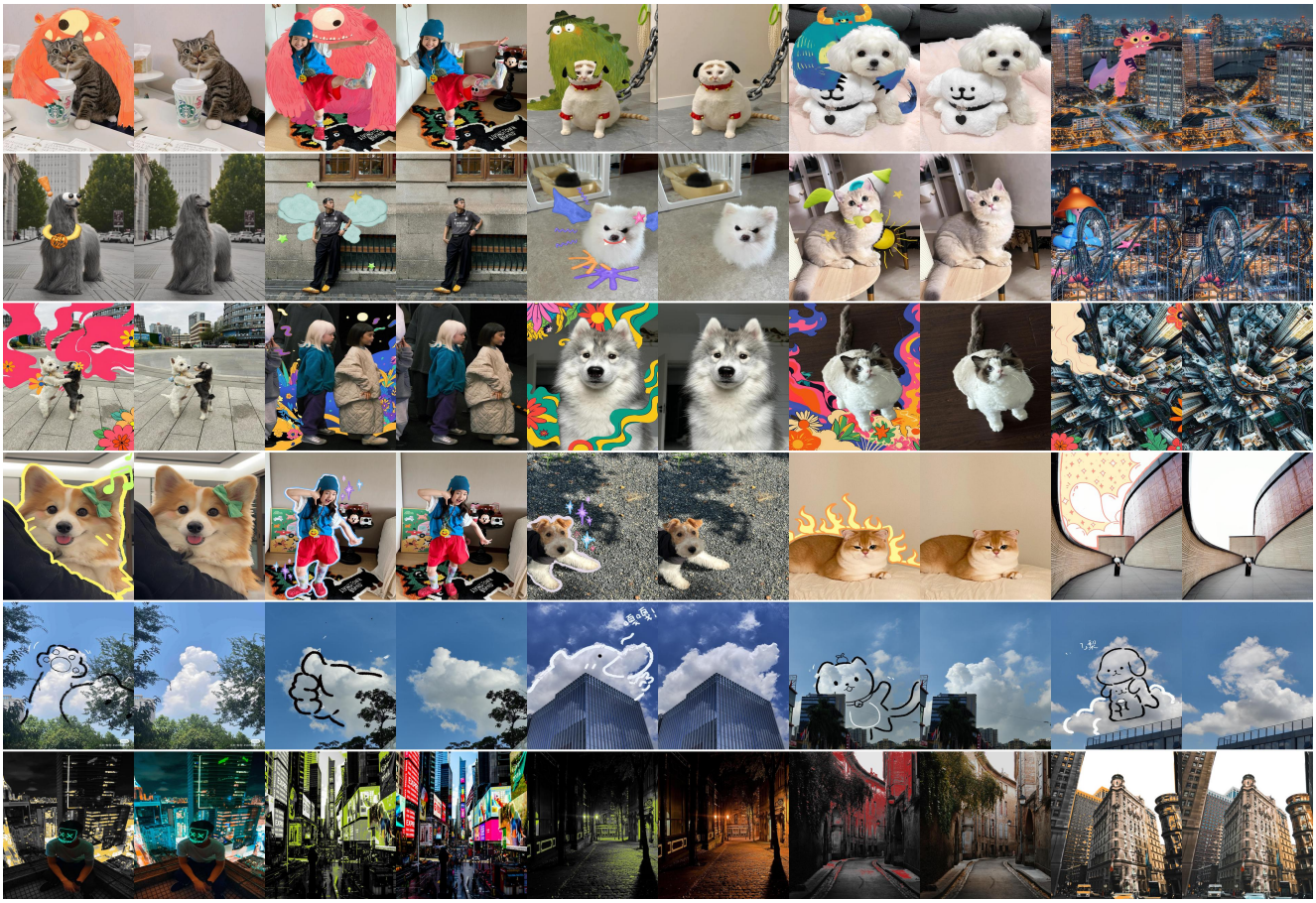

Figure 2. Visualization of samples in the proposed DoodleArt Dataset.

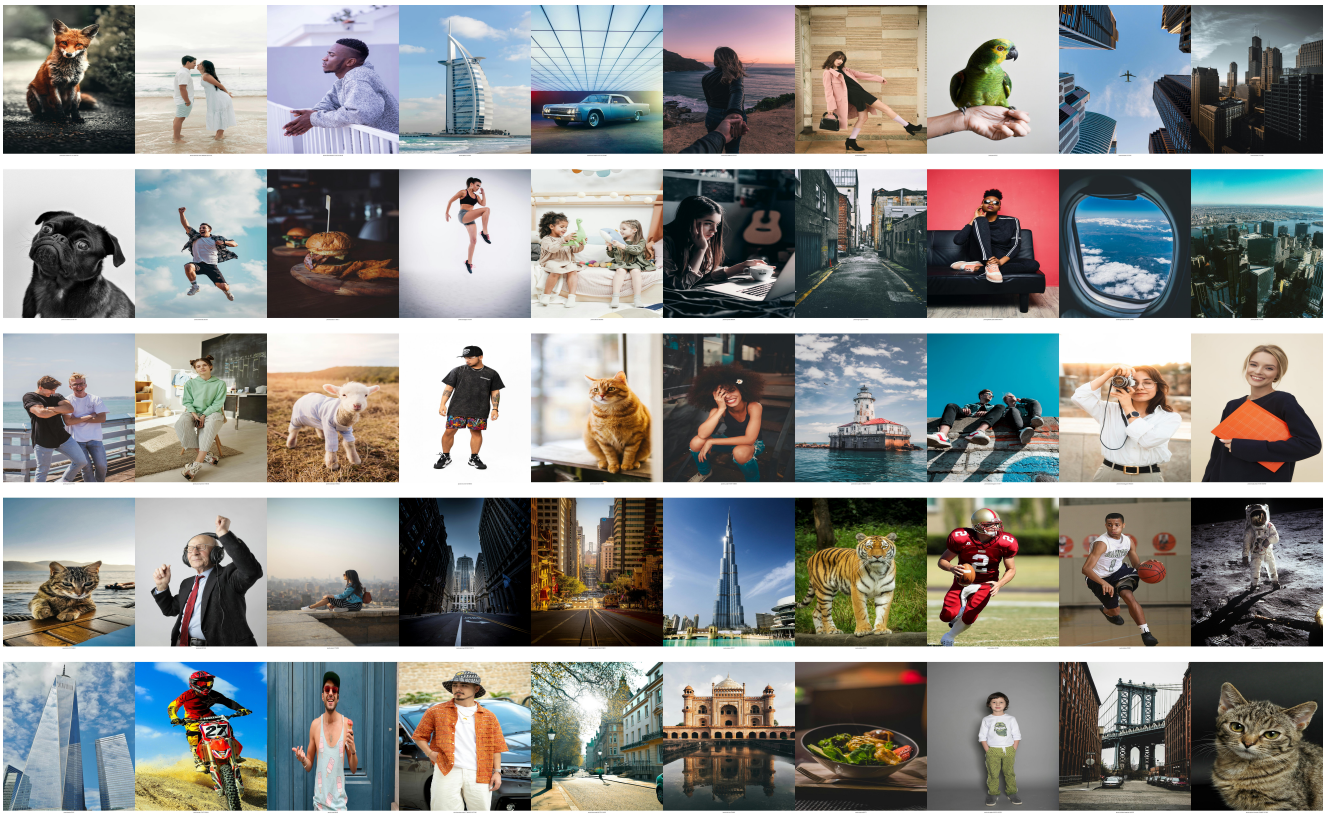

Figure 3. Visualization of samples in evaluation benchmark.

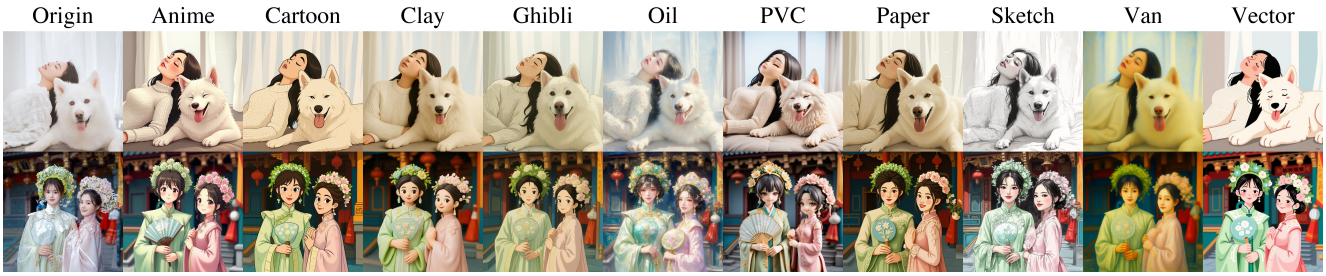

Figure 4. Samples of additional styles used for validation.

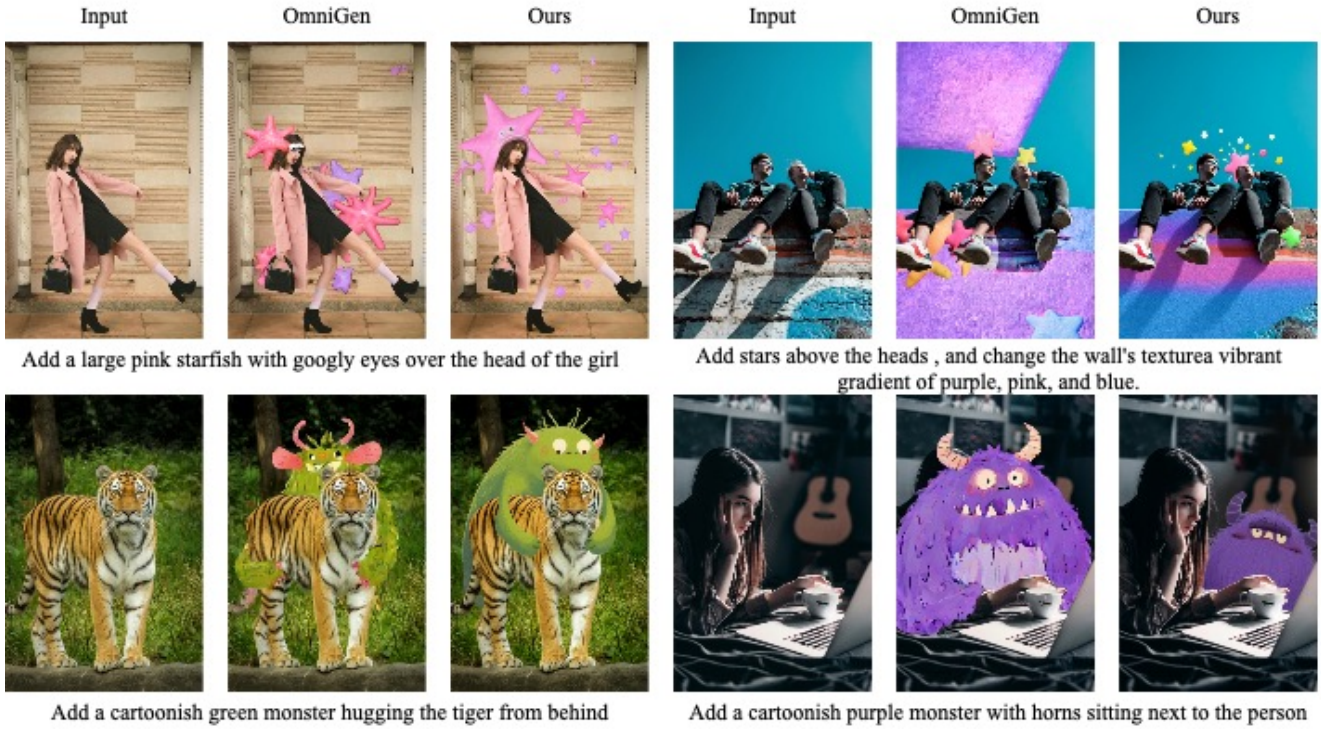

Figure 5. Comparison with OmniGen.