

7 Appendix of "OmniConsistency: Learning Style-Agnostic Consistency from Paired Stylization Data"

7.1 Implementation Details of the GPT-4o Evaluation

In the GPT-4o evaluation process, we establish specific metrics to assess various aspects of image generation tasks. These metrics are tailored to ensure comprehensive evaluation, capturing both objective scoring and comparative analysis for different types of tasks.

7.1.1 Direct Scoring Evaluation (for Style Transfer and Content Consistency Assessment)

The evaluation involves assessing the quality of the image generated through style transfer, considering both the consistency of the artistic style and the alignment with the original content. The scoring metrics used in this context include:

- **Style Consistency:** This measures how well the generated image reflects the artistic style of the reference images. The rating is provided on a scale from 1 (highly inconsistent) to 5 (extremely consistent).
- **Content Consistency:** This evaluates how closely the generated image mirrors the content of the original image, focusing on key elements such as facial features and overall layout. The scale ranges from 1 (highly inconsistent) to 5 (highly consistent).

For each aspect, the assistant provides a score based on a careful analysis of the image characteristics. The scores are then outputted in JSON format as follows:

```
{
  "style_consistency": {
    "score": 5,
    "reason": "xxx"
  },
  "content_consistency": {
    "score": 4,
    "reason": "xxx"
  }
}
```

7.1.2 Example of Task Prompt and Evaluation

Task Prompt: "Evaluate the style transfer of an image based on the provided reference style images and the original content image."

Images: [Upload images of the original content image, reference style images, and the generated images]

Evaluation: The assistant evaluates the generated image for both Style Consistency and Content Consistency, using the following criteria:

Style Consistency: How well does the generated image reflect the artistic style and overall atmosphere of the reference style images? The rating is given on a scale from 1 (highly inconsistent) to 5 (extremely consistent).

Content Consistency: How closely does the generated image resemble the content of the original image, including key elements like facial features and the overall layout? The rating is given on a scale from 1 (highly inconsistent) to 5 (extremely consistent).

This dual evaluation approach, focusing on both Style Consistency and Content Consistency, ensures a detailed and effective assessment of the quality of style transfer images generated by GPT-4o models.

7.2 User Study

7.2.1 Implementation Details

We conducted a user study through a questionnaire to evaluate the performance of different models in terms of style consistency and content consistency. A total of 30 questionnaires were distributed, each containing 30 questions. In terms of style consistency, we did not directly compare with GPT-4o because it does not support style LoRA injection. Instead, we approximated the desired style effects by carefully adjusting the prompts.

For each question, participants were provided with a reference image and the original image. They were then asked to select the best outputs for style consistency and content consistency from the results generated by

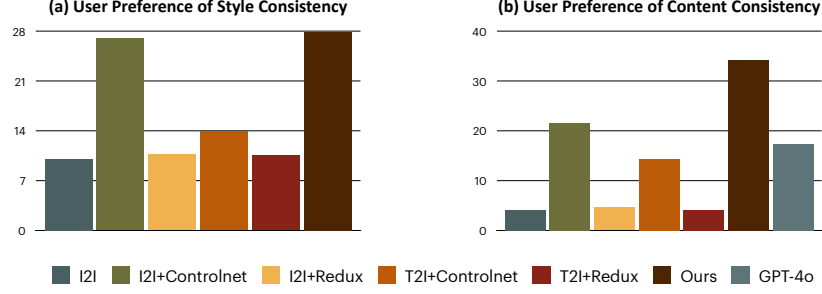


Figure 7: User study: Preference rates for style and content consistency across methods.

(a) Stylization of images containing non-English text (b) Stylization of complex scenes with multiple people

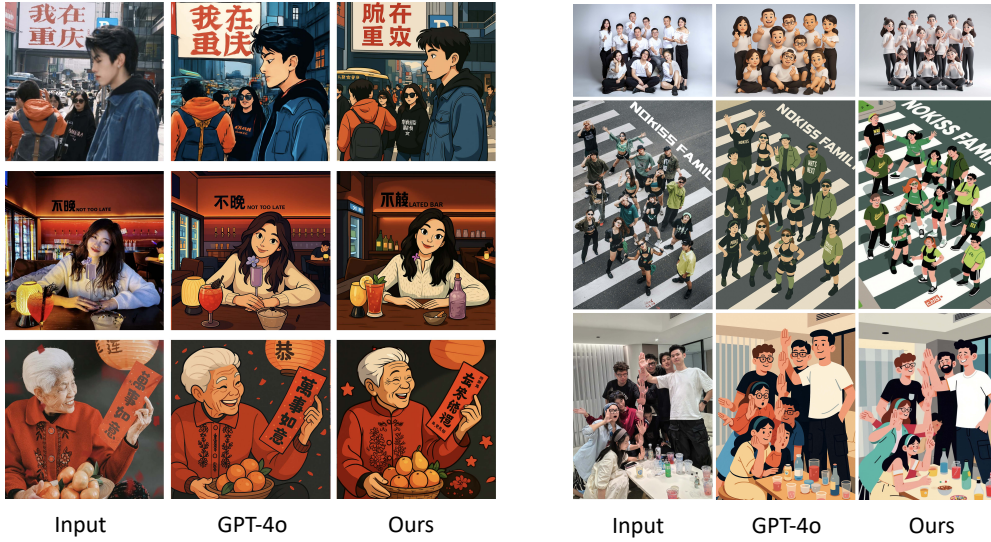


Figure 8: Failure cases.

different models (multiple selections allowed). During the analysis, each selection made for a particular model was counted as one point, and the percentage score for each model was calculated based on the total number of selections. As shown in Fig. 7, our results received higher user preference in terms of both style consistency and content consistency.

7.2.2 Example of User Study

Question: Given the reference image and the original image, select the best outputs in terms of style consistency and content consistency from the provided options.

Style Consistency: How well does the generated image reflect the artistic style and overall atmosphere of the reference style images? Choose the best options from the provided images.

Content Consistency: How closely does the generated image resemble the content of the original image, including key elements such as facial features and overall layout? Choose the best options from the provided images.

7.3 Limitations and Failure Cases

We present several limitations and failure cases in Fig. 8. Specifically, Fig. 8 (a) illustrates stylization results on images containing Chinese text. While GPT-4o largely preserves the shape and legibility of the characters, our method struggles with maintaining the integrity of non-English text, likely due to limitations in the FLUX backbone. Fig. 8 (b) shows stylization outcomes on group photos and complex scenes. Both our method and GPT-4o occasionally exhibit inconsistencies in the number of people depicted, often omitting individuals who occupy smaller portions of the image. Additionally, artifacts may appear in small facial or hand regions.

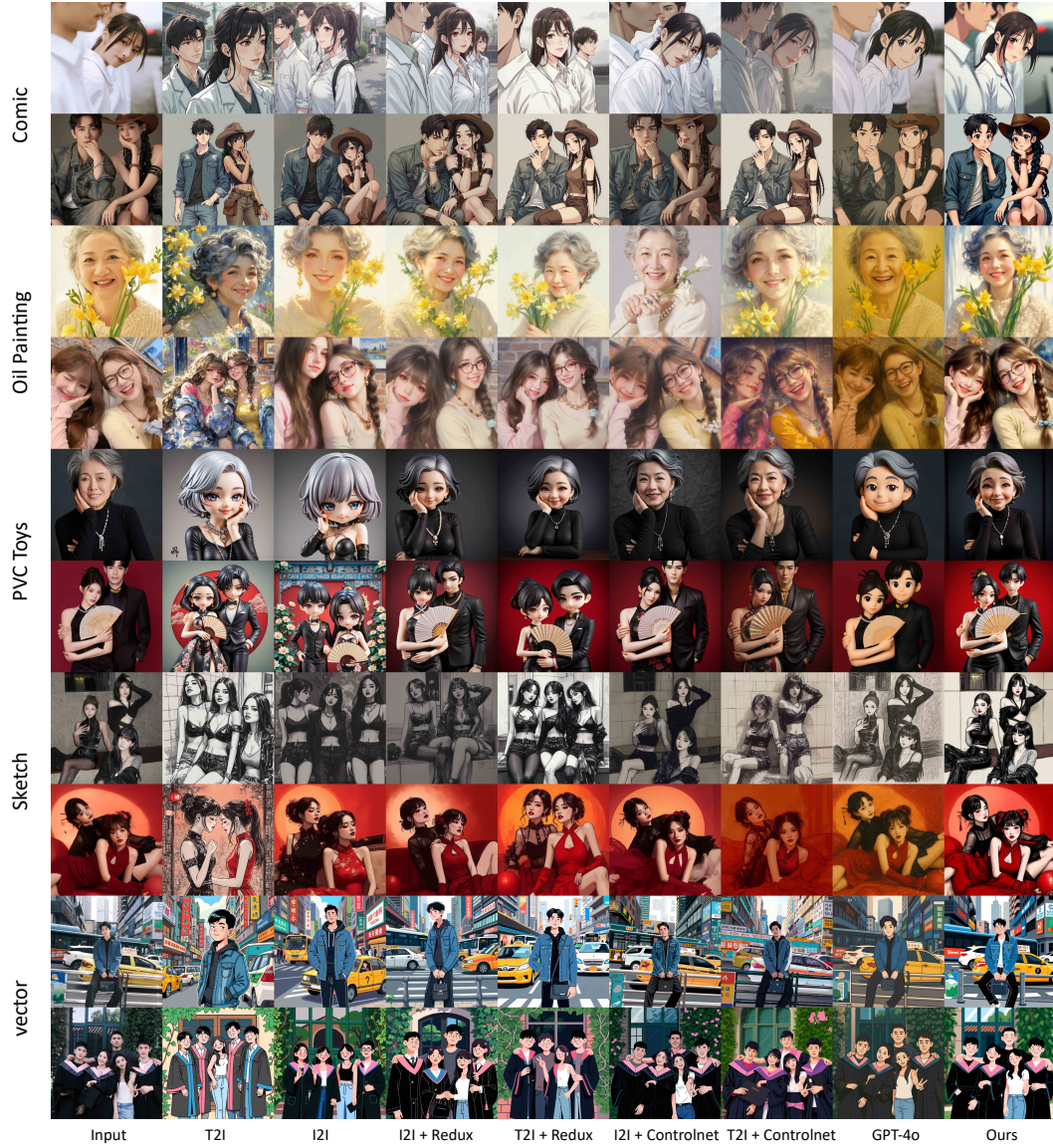


Figure 9: More Comparison results.

7.4 More Results

We present additional experimental results in this section. Fig. 9 shows the comparative results, while Fig. 10 and Fig. 11 demonstrates our method applied to a wider range of styles.



Figure 10: More image stylization results of OmniConsistency.



Figure 11: More image stylization results of OmniConsistency.