

LayerTracer: Cognitive-Aligned Layered SVG Synthesis via Diffusion Transformer

Supplementary Material

A. User Study Details

We conducted a user study in the form of an online survey, with a total of 46 participants evaluating 36 questions per questionnaire. The study was divided into two sections:

The first section evaluated text-to-vector generation results, comparing different methods. Each comparison included the corresponding text prompt, allowing participants to select the generated vector graphic that best reflected the original textual description. The evaluation consisted of three examples per method, leading to a total of nine comparisons, with two questions per comparison:

1. Which vector graphics result looks better? 2. Which result better reflects the meaning of the original text?

The second section focused on layer-wise vectorization, comparing different approaches. To ensure a fair evaluation, we provided a structured breakdown of the vectorization process, allowing participants to assess the quality and rationality of the vector graphic creation process. This section also contained three examples per method, resulting in nine comparisons, with two evaluation questions:

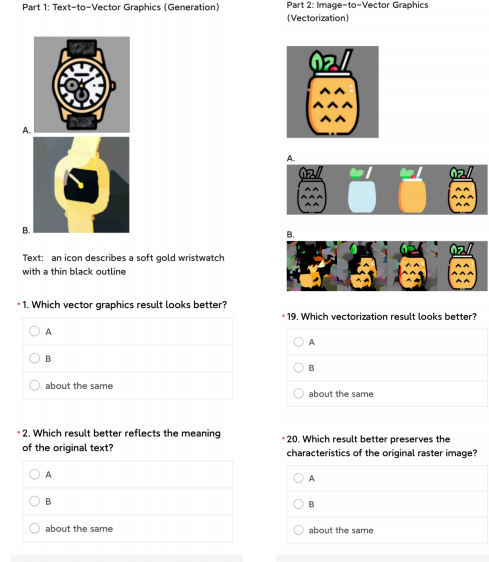
1. Which vectorization result looks better? 2. Which result better preserves the characteristics of the original raster image?

B. Dataset Construction Methods.


The SVG dataset proposed in this work is collected from multiple sources, including the internet, vendor procurement, and designer-created assets. In addition to manually crafted layered SVGs, we also engaged designers to reorganize and arrange layers of other SVGs into a logical sequence that aligns with the design creation process. The final training data was obtained through multiple rounds of human-in-the-loop filtering. The dataset consists of a total of 20K samples across three styles: black outline icons (3K), illustrations (2K), and emojis (15K). For caption annotation, we added different triggers for different datasets and used the Florence-2 model [53] to label the last frame of the SVG sequence.


C. Failure Cases

This section presents some failure cases of LayerTracer. As shown in the figure, when generating pixel-format sequence data, issues such as unwanted repeat, no layering, and inconsistencies between consecutive frames may occur. The problem of undesired repetition can be mitigated during the Layer-wise vectorization stage by comparing adjacent



Part 1: Text-to-Vector Graphics (Generation)

A. 

B. 

Text: an icon describes a soft gold wristwatch with a thin black outline

*1. Which vector graphics result looks better?

☐ A

☐ B

☐ about the same


*2. Which result better reflects the meaning of the original text?


☐ A

☐ B

☐ about the same

Part 2: Image-to-Vector Graphics (Vectorization)

A. 

B. 

*19. Which vectorization result looks better?

☐ A

☐ B

☐ about the same

*20. Which result better preserves the characteristics of the original raster image?

☐ A

☐ B

☐ about the same

Figure 10. Examples of questions in the User Study online questionnaire.

frames to remove duplicates. However, incorrect layering and frame inconsistencies can negatively impact the quality of Layer-wise SVG generation.

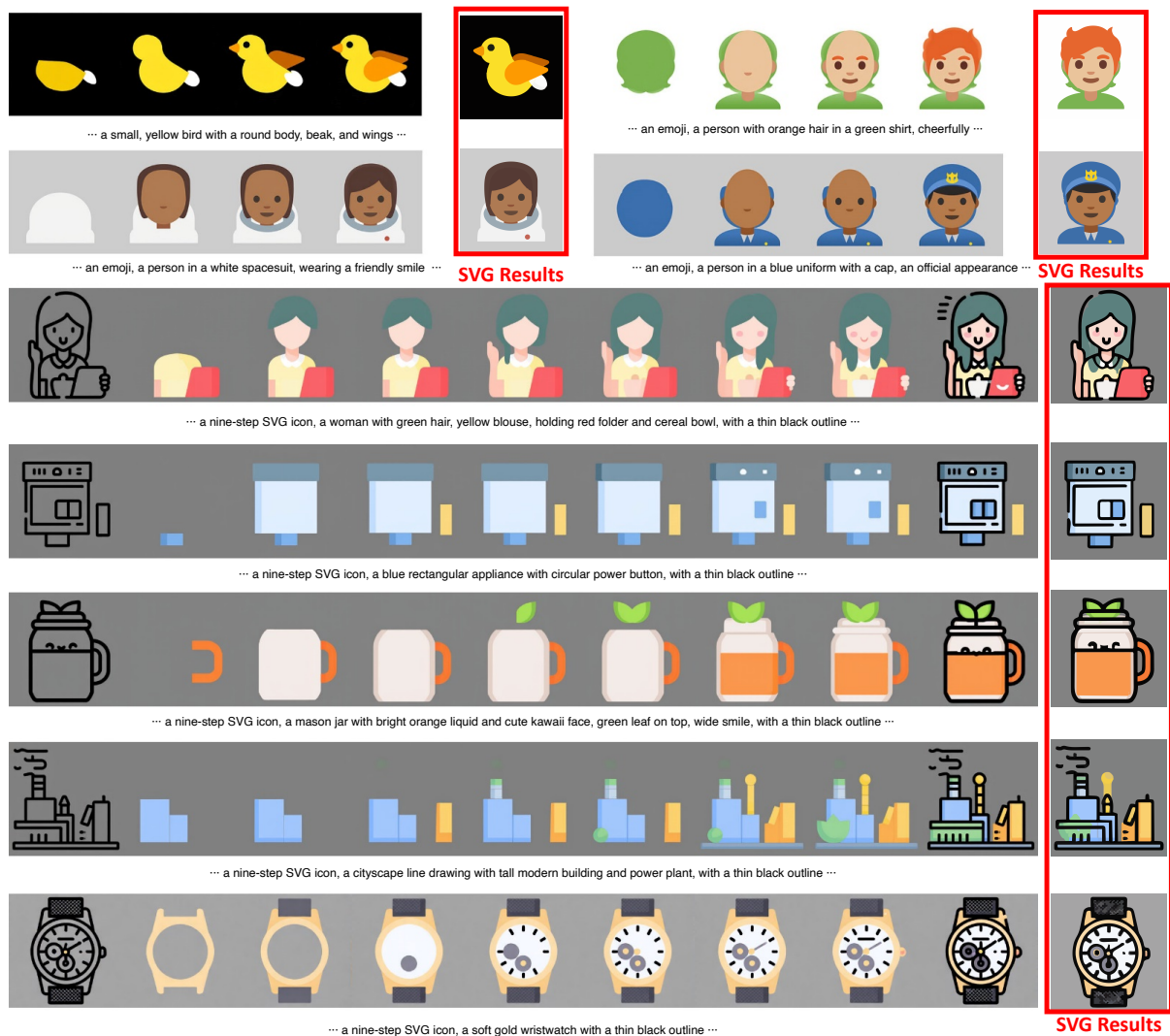


Figure 11. More results of layered vector graphics generation. The red boxes highlight the SVG format results, which can be zoomed in to view the details.

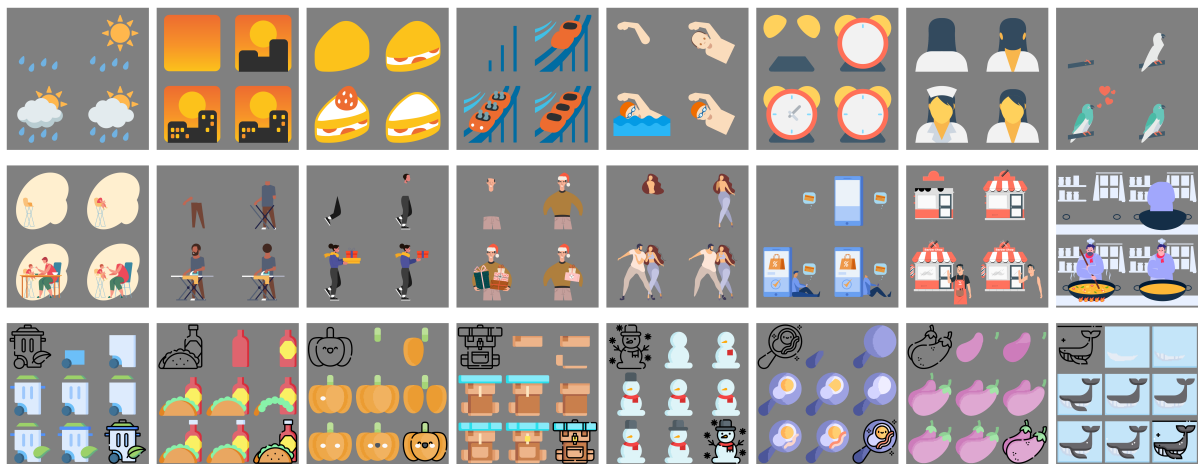


Figure 12. Each row represents examples of SVG datasets for three categories: emojis, illustrations, and black outline icons.

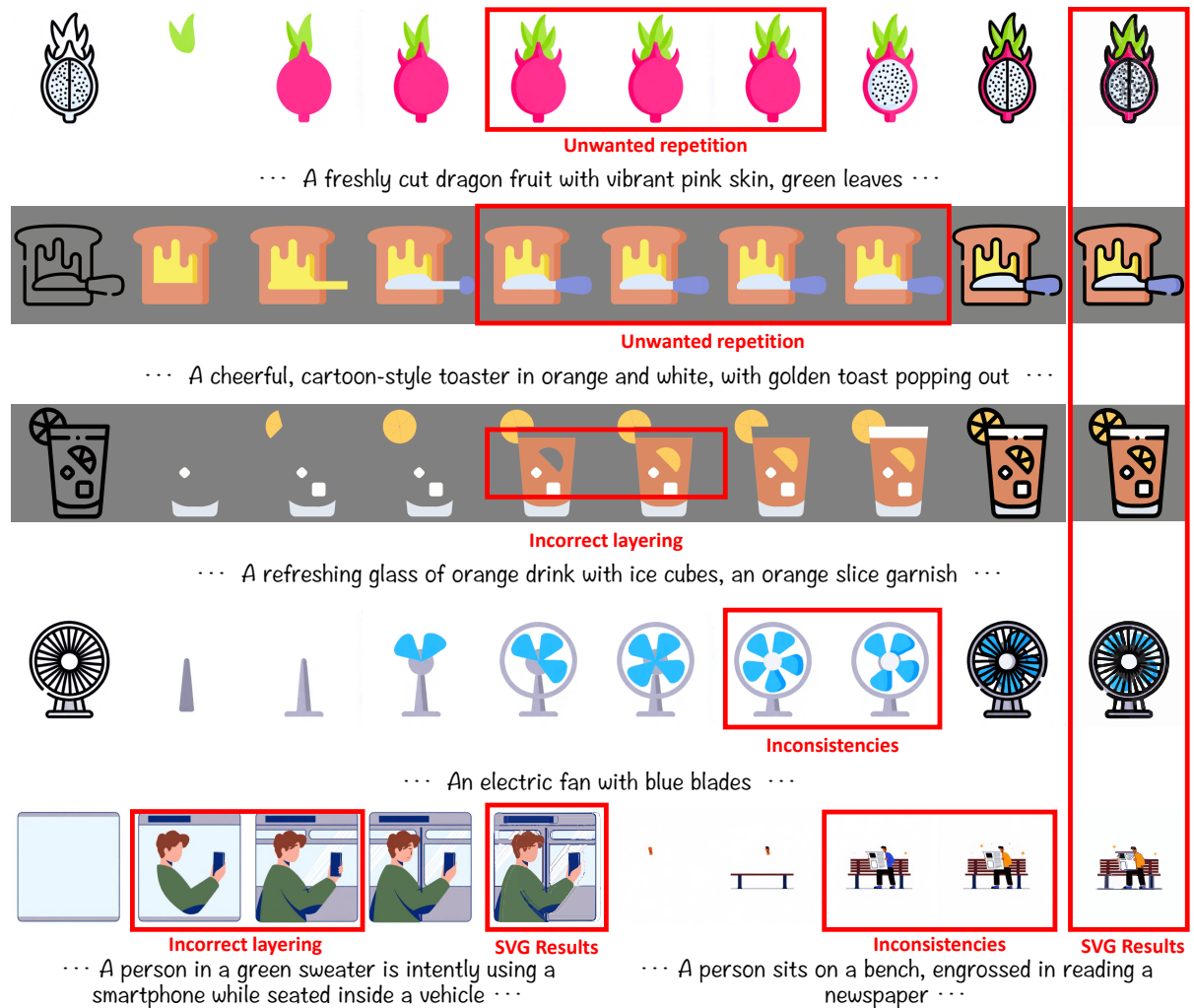


Figure 13. Some failure cases.