Appendix

Implementation of Multi-line Text Editing. As elaborated in the main text, we curated a dataset of 300,000 images containing continuous text segments for training. During the training process, the newline character ("\n") was treated as an independent and special token within the prompt. Each line of text content was concatenated using "\n", ensuring that the positions of text segments in the glyph images approximately matched their corresponding locations. Guided by the textual information and the injected glyph image data, as shown in Fig. 1, our method demonstrates a robust capability for precise multi-line text editing.

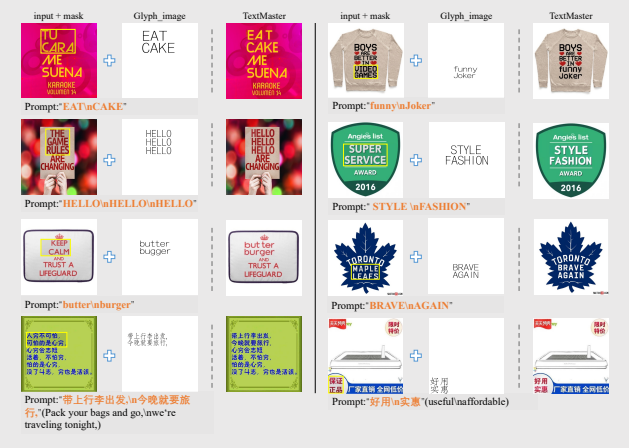

Figure 1. The results of multi-line text editing using **TextMaster**, with test images sourced from the Laion and WuKong evaluation datasets.

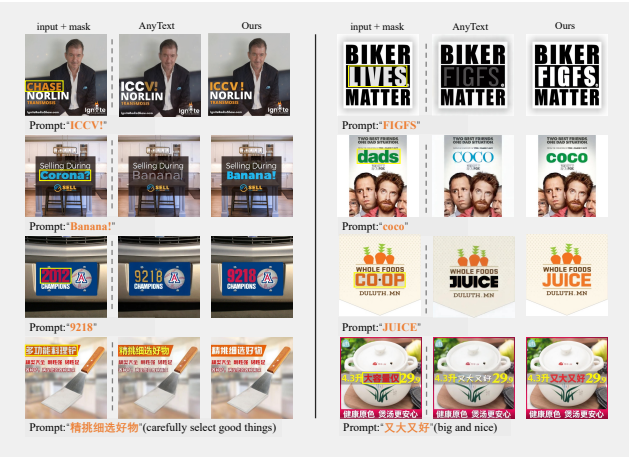

Figure 2. Comparison of results between **TextMaster** and the state-of-the-art AnyText method. The AnyText inference model utilizes the official open-source version, with test images sourced from the AnyText benchmark dataset.

Detailed Description of Style Retention Capability. The re-edited text should maintain the same style as the original, including font style, font color, and font size. However, current text editing methods typically rely on the style of surrounding text, the overall style of the image, or the model's inherent memory to generate style information. As shown in Fig. 2, TextMaster seamlessly integrates the original text style into the newly generated text, whereas state-of-the-art methods can only produce styles based on external conditions, often resulting in random style generation. Additionally, these qualitative results demonstrate that TextMaster excels in layout and typesetting capabilities, further highlighting its superiority. More visual comparison results with the full method are presented in Fig. 4.

General Text Rendering Capability. As shown in the Fig. 3, although our method is not tailored for synthesis scenarios, it can be easily extended to general text rendering. Our method is capable of generating harmonious text rendering effects based on the surrounding background and reference text. Compared to text-to-image-based text rendering approaches, our model offers greater controllability over the placement of the text, preserves the original elements of the image, and maintains high-resolution outputs.

Figure 3. General text rendering visualization.

Limitations and Future Work

TextMaster currently lacks the capability to simultaneously edit non-continuous multi-line text, which is a challenge we aim to address in future work. Furthermore, beyond simply ensuring that the style of the target text matches that of the original, our future work will focus on enabling the controlled integration of arbitrary styles into the target text.

Figure 4. Visual comparison results.