Agentic Retoucher for Text-To-Image Generation

Supplementary Material

In this supplementary material, Sec. 1 presents additional details of our proposed GenBlemish-27K. Sec. 2 provides the implementation details of the proposed Agentic Retoucher. Sec. 3 reports additional qualitative results to demonstrate the superiority of Agentic Retoucher. Finally, Sec. 4 discusses the limitations of our method.

1. GenBlemish-27K: Further Details

Additional Annotation Details. We present the interface of the annotation tool used in our study. Because precise localization of distorted regions is critical, we decouple region annotation from the subsequent natural-language description stage, allowing annotators to focus exclusively on localization. As illustrated in Fig. 2, the top panel lists the twelve predefined distortion categories, each visualized with a distinct color overlay to facilitate differentiation. Annotators mark points, with each point defining a circular coverage area centered at the point and having a radius equal to 1/20 of the image height; whenever possible, distorted regions should be fully encompassed by these coverage circles. In addition, since we use images from EvalMuse-Structure [1], we overlay the dataset's native distortion labels on the image for annotator reference.

Formatted Annotations. As shown in Fig. 3, we provide further details on the GenBlemish-27K annotation formatting. We have endeavored to provide comprehensive annotations covering all defined distortion types.

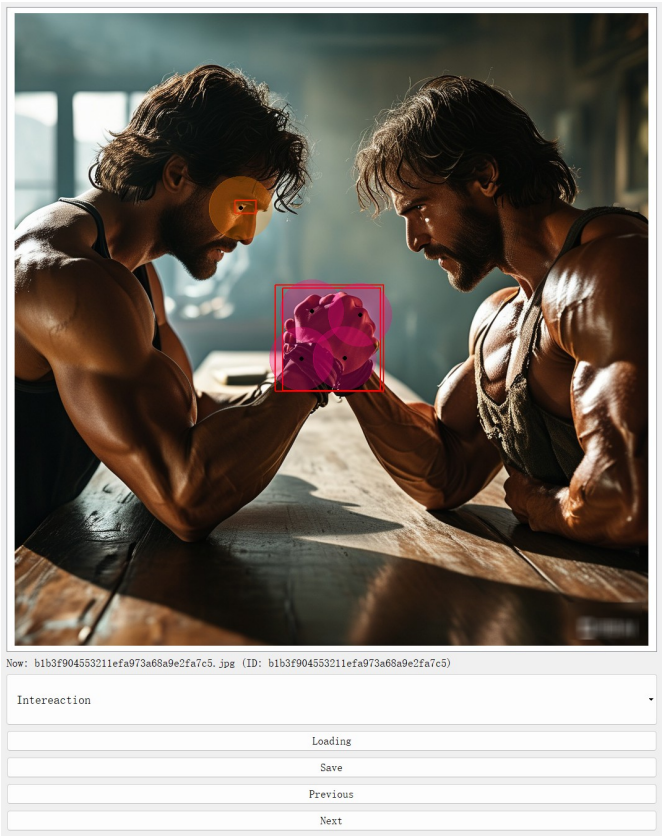

Figure 2. The interface of the annotation tool.

2. Implementation Details

2.1. More Details on Evaluation Metric

Evaluation of the Agentic Retoucher. When evaluating the final outputs of Agentic Retoucher, our retouching is confined to fine-scale distortion regions. Under this setting, we observe that conventional pixel-level metrics, such as FID and aesthetic scores, are primarily suited to assessing global image quality. Furthermore, quality assessment (QA)-based approaches like the VQA score [3], are still designed for whole-image evaluation. As shown in Fig. 4, the second-column image is nearly indistinguishable from its first-column counterpart under subjective inspection, and both exhibit warped distortions in the hand region. Nevertheless, VQAScore and Aesthetic Score rank the second image highest in visual quality, underscoring their failure to assess fine-scale regional distortions. By contrast, RichHF [2] introduces four metrics: plausibility, aesthetics, alignment, and overall score, explicitly designed to assess whether fine-scale structures are plausible, whether the image accords with human aesthetic perception, whether textual content is aligned, and the overall subjective quality. As these objectives closely align with our task setting, we adopt these metrics for quantitative evaluation.

Evaluation of the perception agent. In the main text, we evaluate using five metrics grouped into two categories: distribution-based metrics (CC, SIM, KLD) and location-based metrics (AUC-Judd, NSS). These two perspectives further substantiate the effectiveness of the Context-Aware Perception Agent.

Evaluation of the reasoning agent. For the Human-Alignment Reasoning Agent, we evaluate the quality of generated natural language descriptions using standard NLP metrics, namely ROUGE and METEOR. Beyond lexical overlap, we further extract high-level semantic features with Word2Vec and SimCSE for both the generated descriptions and the human-annotated ground truth, and compute feature similarity to assess alignment.

Details on Human Evaluation. To more effectively assess the post-editing performance of different models on

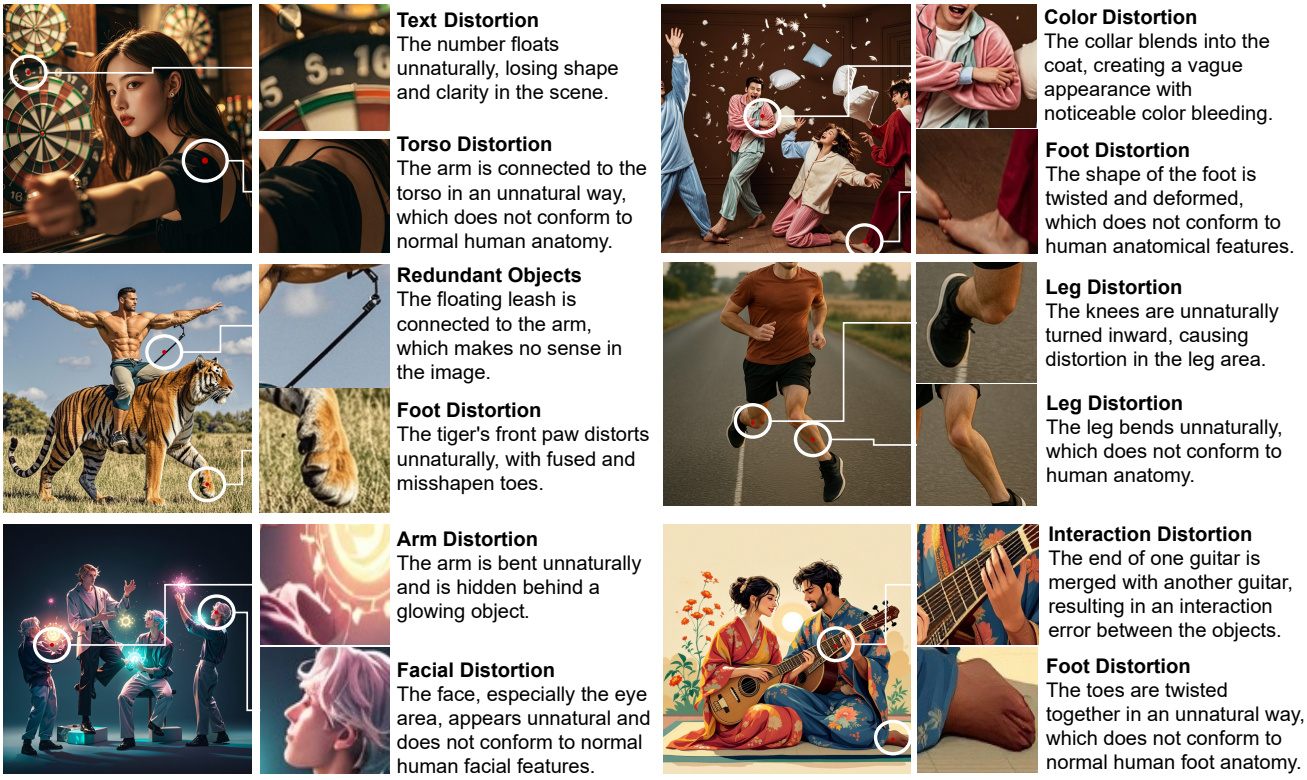

Figure 3. Additional distortion types annotated in the GenBlemish-27K dataset.

Figure 4. Aesthetic and VQAScore exhibit severe degradation on fine-scale regional distortions, failing at image quality assessment.

text-to-image outputs, we recruited five volunteers with research experience in image generation for a human evaluation. We randomly sampled 100 image sets, each comprising the original image, the result refined by Agentic Retoucher, and the corresponding outputs from baseline post-editing methods. The annotators conducted pairwise comparisons between the original and each refined image, selecting one of five options: Image A is significantly/slightly better than Image B; approximately the same; Image B is slightly/significantly better than Image A. The annotators were blind to which method produced the refined images and were instructed to pay particular attention to fine-grained regions.

2.2. More Details on Agentic Retoucher

When evaluating the refined outputs of Agentic Retoucher, we randomly split GenBlemish-27K into training and test sets at a 5:1 ratio. For SynArtifacts-1K, we performed inference on the entire dataset. When employing the two categories of inpainting tools within our Adaptive Action Toolkit, we standardized all prompt inputs except for the outputs provided by the Context-Aware Perception Agent and the Human-Alignment Reasoning Agent. Specifically, for VLM-based methods, we used the prompt *"Please fix the distortions in the provided image."* For our method, we appended the distortion-region description generated by the Human-Alignment Reasoning Agent to this prompt. For mask-based methods (e.g., Flux-fill), which have limited text-handling capability, we used the original image caption as input and supplied the distortion-region mask produced by the Context-Aware Perception Agent as the specific input of our method. We then compared qualitative and quantitative results to demonstrate the effectiveness of Agentic Retoucher.

Context-Aware Perception. In the main text, we analyze the limitations of saliency prediction using both traditional algorithms and deep learning–based methods. Furthermore, to demonstrate the shortcomings of general-purpose VLMs in localizing distorted regions, we designed

the following prompt: *"You are an assistant for detecting AIGC artifacts. You will be given an image and a bounding box, and your task is to determine whether the specified region contains any AIGC-related visual distortions."* We fine-tuned the VLMs referenced in the main text (InternVL-3.5, Qwen2.5-VL, GLM-4.1V). For the RichHF method cited in the main text, due to its task formulation, we did not train it on GenBlemish, in order to assess its zero-shot distortion detection capability.

Human-Alignment Reasoning. We adopt a unified query setting for the model as follows: *"For the provided bounding box, identify the distortion type and provide a description of it. The type classification is as follows: A: text distortion; B: arm distortion; C: hand distortion; D: leg distortion; E: foot distortion; F: torso distortion; G: facial distortion; H: redundant objects; I: object deformation; J: color distortion; K: interaction distortion; L: other distortions."*

Figure 5. More Qualitative visualization of saliency prediction.

3. Additional Qualitative Results

3.1. More results of Agentic Retoucher

To further demonstrate the effectiveness of the proposed Agentic Retoucher, we present additional qualitative results that encompass a broader range of distortion types. As shown in Fig. 7 and Fig. 6, these qualitative findings further substantiate Agentic Retoucher's capability on the text-to-image refinement task.

3.2. More results of Perception and Reasoning

Context-Aware Perception. We provide additional qualitative comparisons on saliency prediction. As shown in Fig 5, our approach consistently demonstrates more accurate and context-aware localization. Specifically, our method shows clear advantages in localizing textual distortions while effectively mitigating RichHF's tendency to overemphasize facial regions. Compared with deep learning–based baselines, coupling textual input evidently endows saliency prediction with stronger context awareness.

Human-Alignment Reasoning. As illustrated in Fig. 8, the left panel provides further examples to demonstrate that existing state-of-the-art vision-language models continue to struggle with reliably identifying regions of distortion, frequently misclassifying clearly abnormal areas as normal. The right panel presents a comparative evaluation between our proposed Human-Alignment Reasoning Agent and the Qwen-2.5VL model on text-to-image content assessment. The zero-shot Qwen model fails to accurately identify distortion types and often provides incorrect descriptions of image distortion, whereas our method achieves more precise and reliable results.

4. Limitations

As Agentic Retoucher is a comprehensive system, error accumulation across its various components is inevitable. We mitigate this issue as much as possible through iterative refinement. Furthermore, in existing text-to-image outputs, distortions predominantly involve human subjects, with hand regions being the most affected. This inevitably results in data imbalance, which in turn leads to models disproportionately focusing on hand or facial areas. Our training strategies and model design aim to alleviate these issues. Additionally, the adaptive action agent is constrained by the capabilities of the tools it employs: for example, VLM-based methods may occasionally fail to fully preserve non-distorted regions, while mask-based approaches may be limited by input text length. In future work, we will continue to address these limitations.

References

[1] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation, 2024. 1

[2] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, and Nicholas et al. Carolan. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF CVPR*, pages 19401–19411, 2024. 1

[3] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IX*, page 366–384, Berlin, Heidelberg, 2024. Springer-Verlag. 1

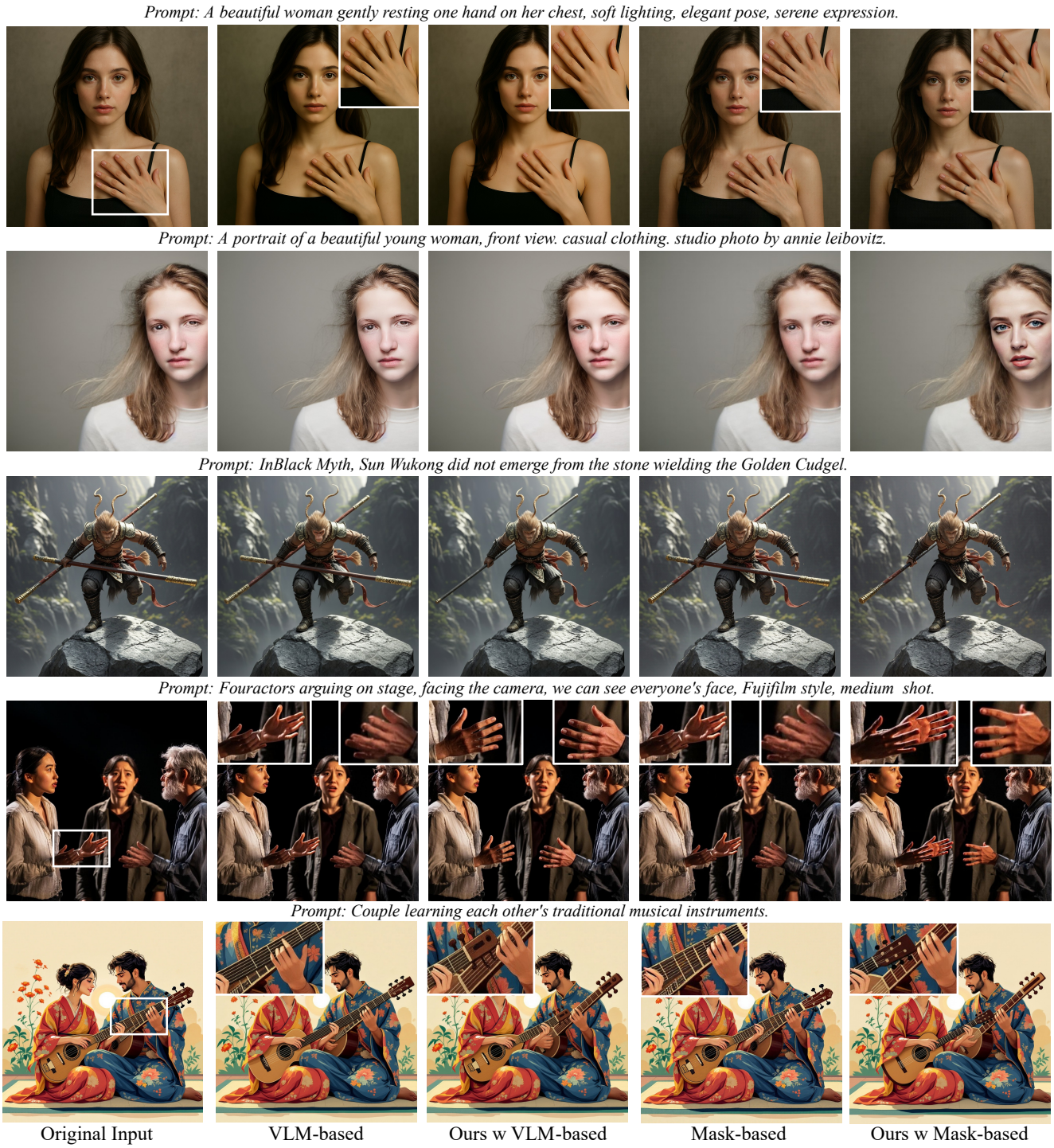

Figure 6. More **Qualitative comparison** of retouching results across diverse prompts. White bounding boxes indicate zoomed-in fine-grained regions.

Figure 7. **Qualitative comparison** of retouching results across diverse prompts. White bounding boxes indicate zoomed-in fine-grained regions. To facilitate the presentation of additional results, we provide side-by-side comparisons of the original images, the baseline refined outputs, and the images refined by our Agentic Retoucher.

Question: Does this image contain any regions exhibiting visual distortions? If so, please briefly indicate their approximate locations.

Doubao-seed-1.6: Upon observation, the image exhibits no apparent distortions. The proportions of the depicted subjects, furniture, and other elements remain consistent, and the overall visual presentation aligns with conventional realism.

Question: Does this image contain any regions exhibiting visual distortions? If so, please briefly indicate their approximate locations.

Claude: The image appears to have no obvious distortion. The facial features of the individuals, their body proportions, and the background environment all seem normal, with no detectable signs of digital manipulation or visual anomalies.

Ours: Hand Distortion.
The finger appears unnaturally elongated and fused with the spoon, distorting its anatomical structure.

Qwen-2.5VL: Arm Distortion.
This classification suggests that the arm might be in an unusual position or shape, but without more context or a clearer view, it is difficult to determine the exact nature of the distortion.

Ours: Foot Distortion.
The foot distorts unnaturally, merging into the ground with no clear separation.

Qwen-2.5VL: Leg Distortion.
This classification suggests that there might be an issue with the leg, such as a missing leg, an extra leg, or a distorted leg shape, which is a common type of distortion in digital images.

Figure 8. **Left:** More evidence of existing VLMs hallucinate and fail to localize distortions in AIGC-images. **Right:** Additional qualitative visual results are presented to further demonstrate the capabilities of the Human-Alignment Reasoning agent.