The Consistency Critic: Correcting Inconsistencies in Generated Images via Reference-Guided Attentive Alignment

Supplementary Materials

The supplementary material is structured as follows:

1. We provide additional details on the construction of our dataset in Sec. 1.
2. We compared additional existing editing methods on the restoration task in Sec. 2.
3. We provide representative cases from our dataset and compare them with those from other existing datasets in Sec. 3.
4. We present the correction results of our proposed ImageCritic across multiple languages on images generated by different models in Sec. 4.
5. We present an example of using the agent chain and the prompts used by each of our agent components in Sec. 5.

1. Data Curation details

As illustrated in Fig. 1, our data construction pipeline proceeds sequentially as follows.

Step 1: Reference-target pairs generation

We begin by collecting a large-scale product dataset through web crawling and downloading and employ state-of-the-art generative models to produce synthetic image variants for each product sample.

Step 2: Generation quality filtering

We then used Qwen3-vl [2] to perform quality filtering with prompt: *"Determine whether the text details in the input image are both strictly clearly visible and fully readable. If any part of the text is blurred, low-resolution, or difficult to recognize, answer 'No'. Otherwise, answer 'Yes'. Respond with only 'Yes' or 'No', followed by a brief reason."*.

Step 3: Semantic tags generation

Next, we employed Qwen [2] to assign semantic tags to the generated images in order to extract object categories for training purposes with the prompt *"Given the image of an object, return only the most general category of the object using exactly 1 to 3 words. Strictly avoid any additional details or descriptions."*.

Step 4: Image Grounding

To ensure that the model correctly interprets semantics in complex scenes, we utilized the Qwen [2] for grounding. The prompting strategy was as follows: *"Given image1 as a reference, detect the same object in image2 and output the bounding box coordinates strictly in the format [x1, y1, x2, y2], where x1,y1 are the top-left integers and x2,y2 are the bottom-right integers. You must return only the coordinates in that exact format with no extra text."* where image1 and image2 referred to the input image and the generated image, respectively.

Step 5: Consistency Verification

Based on the bounding boxes predicted by Qwen [2], we adopted SAM [8] to detect and extract precise object masks. To further guarantee the accuracy of these masks and enhance generative consistency, we invoked the Qwen again to compare the full-object masks against the product references, using the following prompt: *"Please analyze if the extracted region in Image 1 corresponds to the product in Image 2. Ignore the background and focus only on the main object. 1. Are the objects in both images the same product? If Image 2 contains only a small portion of a local region from Image 1, consider it as No. (Yes/No). 2. Explain based on visual features such as shape, color, texture, and context. If the object in Image 2 contains the mask region, describe the match."* where image1 and image2 referred to the input image and the generated image, respectively.

Step 6: Image degradation

From each verified full-object mask, we sampled 20%70% of the region to obtain a random local rectangular mask used as the inpainting region for the Flux-Fill [6] model. For Flux-Fill, we designed three types of prompts—*"English words"*, *"Chinese characters"*, *"logos"*—as well as an empty-prompt setting to accommodate different scenarios.

Step 7: Degradation quality filtering

After generation, we again used a Qwen [2] to filter out severely degraded results produced by Flux-Fill, with the

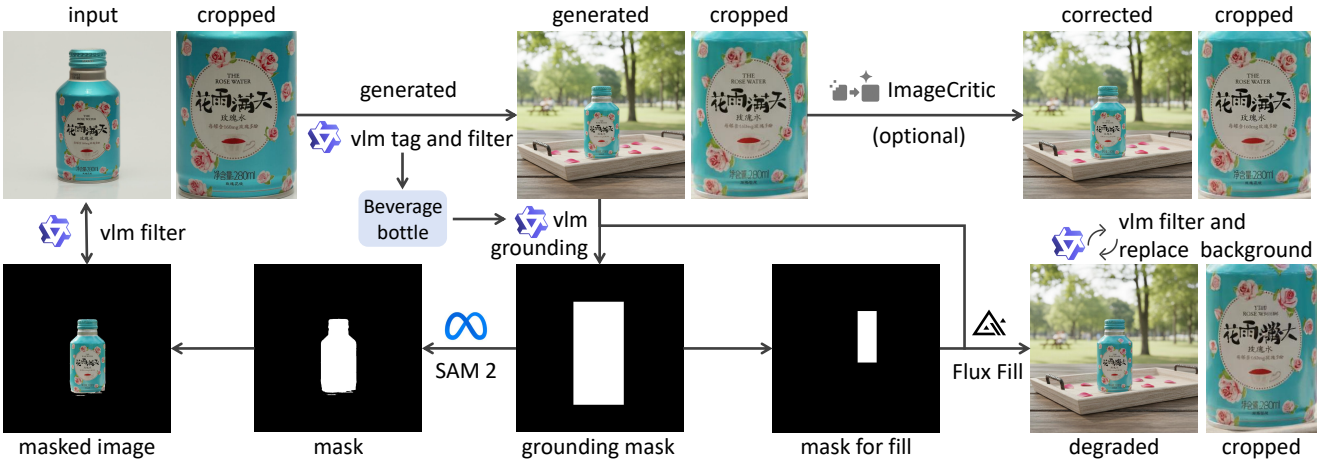

Figure 1. **Data curation details.** We present a comprehensive account of our dataset construction methodology, encompassing data generation, data annotation, and data augmentation procedures.

following prompt: *"Is there is obvious distorted text or mismatched elements in the image. Answer with 'Yes' or 'No', followed by a brief explanation of the reason"*.

Step 8: Final Image Composition

Let M denote a binary full-object mask generated by SAM [8], D the degraded image, and G the generated image. We construct the final degraded sample by combining the degraded content inside the object region with the generated content outside it:

$$I_{\text{final}} = M \cdot D + (1 - M) \cdot G. \tag{1}$$

This procedure effectively suppresses artifacts that may arise in background areas during the filling process, ensuring that degradation occurs strictly within the object region without affecting the original background environment.

Step 9: Iterative Data Enhancement Strategy

After following the training procedure described above, our model is able to perform local region restoration. To further improve the consistency and quality of the generated outputs, we apply additional enhancement to the generated image with Chinese characters. Specifically, since we obtained grounding masks for objects during data construction, we crop the corresponding regions and feed themtogether with the original input imageinto our method for consistency correction. The corrected images are then used as new targets to update the dataset, thereby further improving overall data quality.

2. Image Correction Comparison

We further compared additional multimodal large language models, such as nano-banana [4] and GPT-Image 1 [1], as well as editing models like Omnigen2 [12] and Qwen-image [11], evaluating their capabilities on the task of consistency correction in generated images.

To help the models better understand the task, we designed the following prompt: *"Use the product in the left first reference image as a reference to refine, replace, and enhance the product in the right second to-be-refined image, matching their texture, details, color, logo, and texts, while preserving everything else in the right second to-be-refined image untouched."* As shown in the Fig. 2, existing methods fail to perform consistent correction on generated images while maintaining background fidelity. In contrast, our method substantially improves image consistency while preserving the original background.

3. Dataset Comparison

As shown in the figure, we present several pairs from our dataset and compare our referencetarget pairs with those from other datasets. From the visualizations, it is evident that our paired data maintain high global consistency and rich local details, even under diverse languages, scenes, generation sources, and viewpoints. In contrast, datasets such as Subjects-200k [10] and UNO-1M [13] do not strictly preserve local consistency and often exhibit region-level blurring or mismatches.

4. Additional Visual Results

To further demonstrate the robustness of our method across different languages and scenarios, we first used GPT-Image 1 to generate images with diverse categories in multiple languages, and then performed customized generation using the open-source models XVerse [3], Dreamo [7], MOSAIC [9], OmniGen2 [12], UNO [13], and Qwen-Image [11], as well as the closed-source mod-

els NanoBanana [4] and GPT-Image [1]. We then applied our method for consistency correction and used an OCR model [5] to detect the text in the generated images. As shown in Figures Fig. 4, Fig. 5, and Fig. 6, our method delivers effective consistency restoration, and the OCR results further underscore its robustness across diverse linguistic and contextual settings.

5. Agent Chain Details

As shown in Figure 7, the system progressively corrects the generated image through the coordinated operation of multiple specialized agents. After the user provides a reference image and a generated image, the inconsistency detector, reference finder, and ImageCritic agents sequentially perform comparison, cropping, and correction until the user is satisfied with the result. During this process, the user may choose whether to accept each proposed patch or provide a new target region or description to guide further correction by the agent chain. This iterative loop continues until the generated image aligns with the users intent. The prompts for the Inconsistency Detector, Reference Finder, TagGrounder, and Coordinator are configured in List. 1, List. 2, List. 3, and List. 4, respectively.

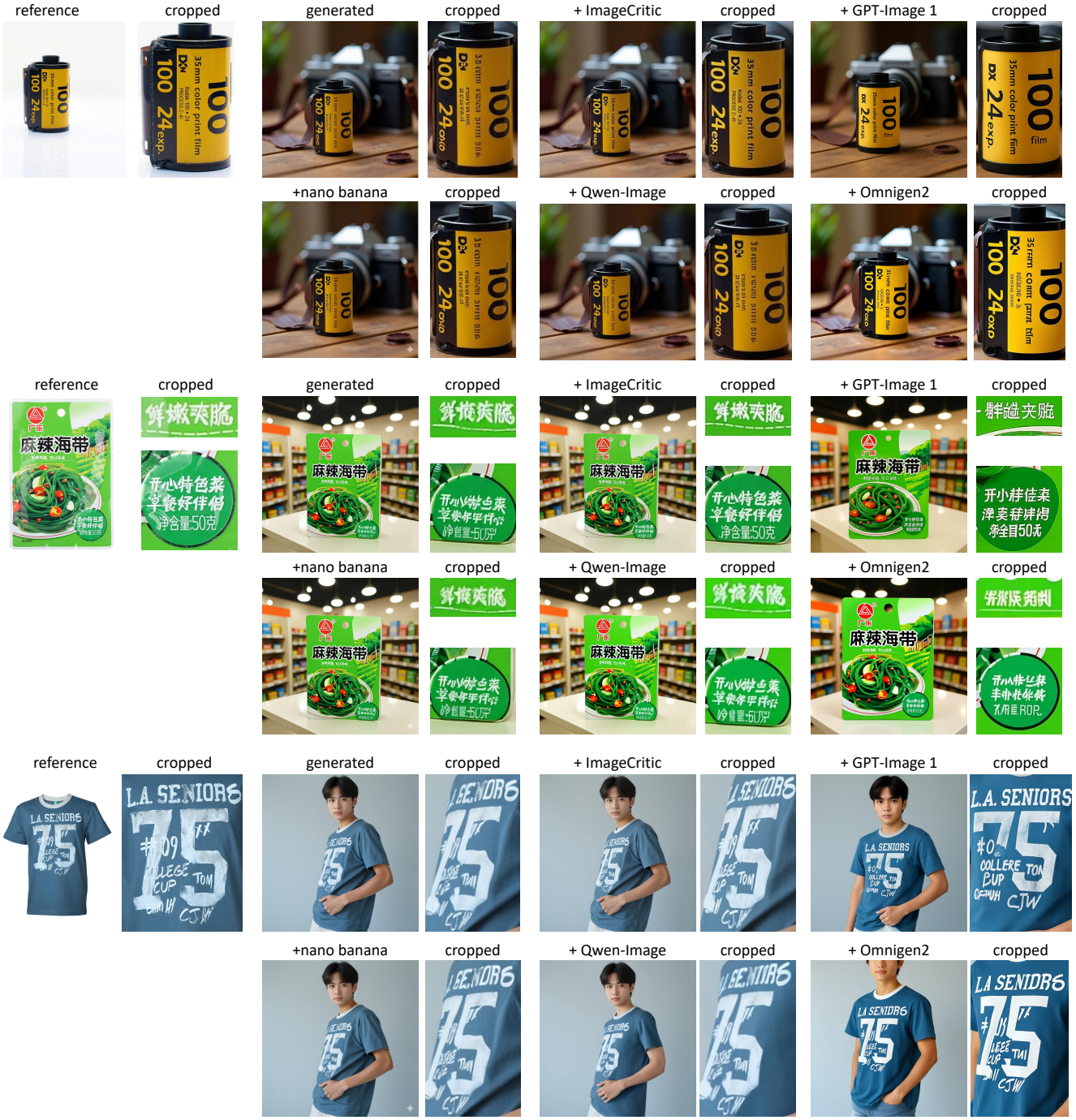

Figure 2. **Comparisons with multimodal and editing model**. We evaluated existing multimodal and editing methods. As shown, current models exhibit issues in preserving global coherence or performing localized corrections, which affects their practical applicability. In contrast, our approach maintains local consistency in the generated content while ensuring overall background coherence.

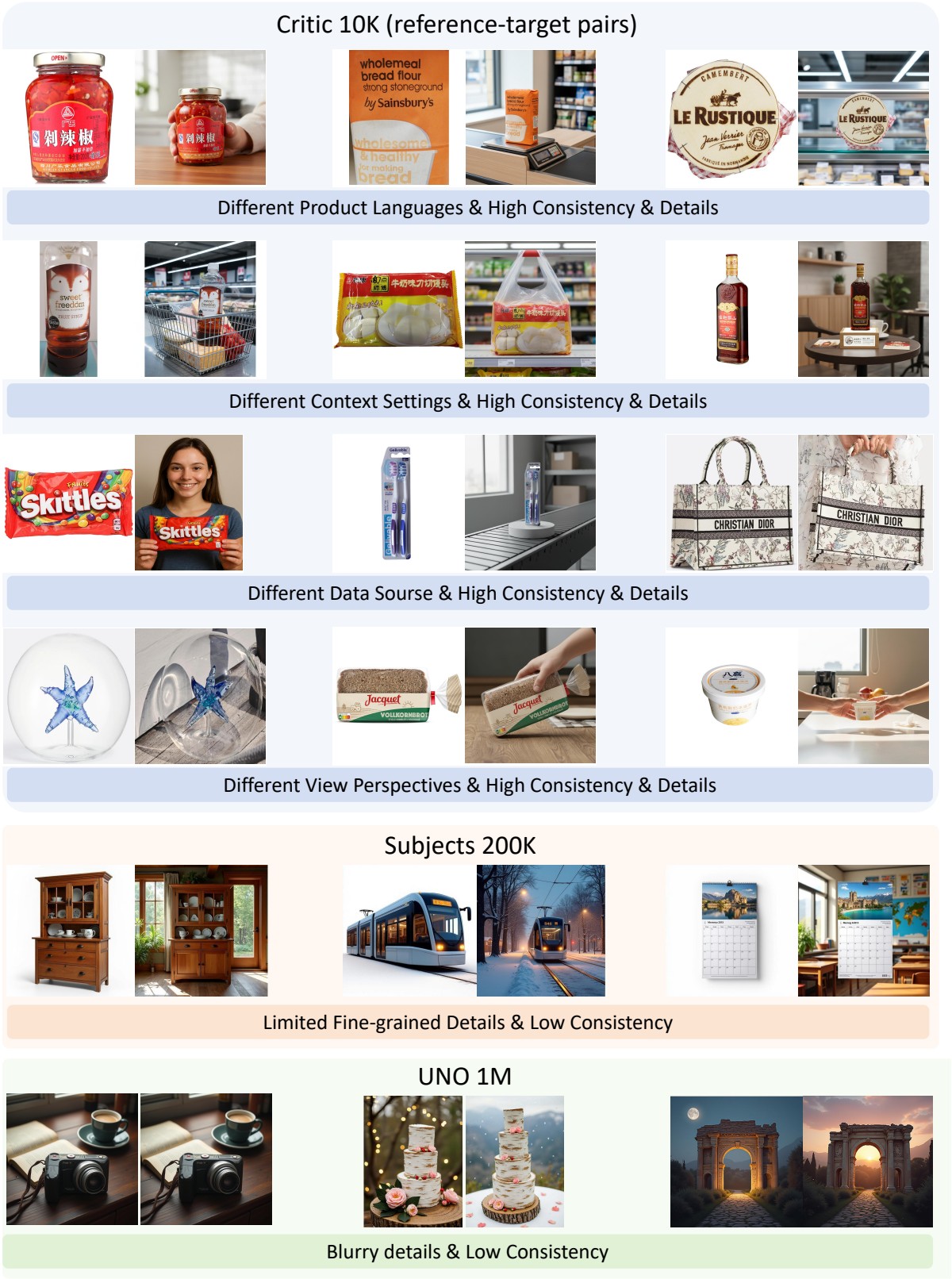

Figure 3. **Comparion with dataset.** Comparison of data samples from our Critic-10k dataset, the Subjects-200K [10], and the UNO-1M [13] dataset.

Figure 4. **Additional visual result.** We present the visual results of our proposed ImageCritic under multilingual, multi-view, and multi-scene settings. By applying an OCR [5] model to the generated images, we observe that after correction with our method, all recognized text perfectly matches the reference images. This demonstrates that our approach achieves precise and comprehensive detail correction without disrupting the overall structural or contextual integrity of the images.

Figure 5. **Additional visual result.** We present the visual results of our proposed ImageCritic under multilingual, multi-view, and multi-scene settings. By applying an OCR [5] model to the generated images, we observe that after correction with our method, all recognized text perfectly matches the reference images. This demonstrates that our approach achieves precise and comprehensive detail correction without disrupting the overall structural or contextual integrity of the images.

Figure 6. **Additional visual result.** We present the visual results of our proposed ImageCritic under multilingual, multi-view, and multi-scene settings. By applying an OCR [5] model to the generated images, we observe that after correction with our method, all recognized text perfectly matches the reference images. This demonstrates that our approach achieves precise and comprehensive detail correction without disrupting the overall structural or contextual integrity of the images.

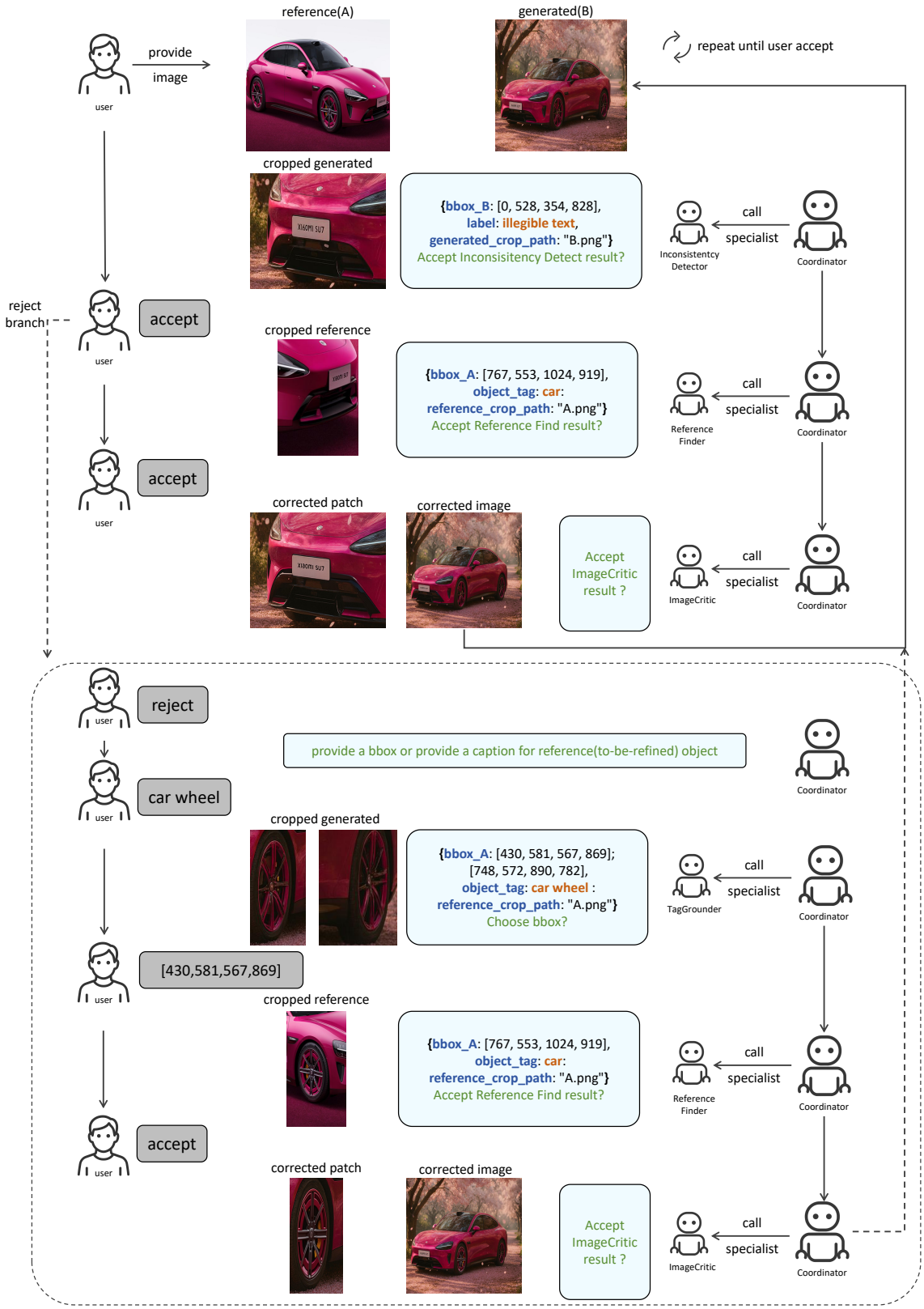

Figure 7. **Illustration of a multi-agent image correcting workflow.** The system performs localized detection, reference matching, and iterative region-level corrections driven by user feedback, progressively correcting the generated image until it is accepted.

Listing 1. The prompt for Inconsistency Detector.

```
prompt = (
   "Carefully compare the two images. Image 1 is the reference image (correct version),
       and Image 2 is the target image that may contain defects. Focus only on the main
       subject of the image, ignoring any differences in the background. Identify the
       region in Image 2 that differs from the corresponding area in Image 1.
       Differences may include blur, illegible text, texture inconsistency, artifacts,
       missing parts, or any other visual discrepancies.Return ONLY the bounding box of
       the different region in Image 2 in the strict format:[xmin, ymin, xmax, ymax]"
)

messages = [
   {
      "role": "user",
      "content": [
         {"type": "text", "text": "Reference Image1 (correct version)"},
         {"type": "image", "image 1": image_A_path},
         {"type": "text", "text": "Target Image2 (may have defects)"},
         {"type": "image", "image 2": image_B_path},
         {"type": "text", "text": prompt},
      ],
   }
]
```

Listing 2. The prompt for Reference Finder.

```
   prompt = (
      "I will show you a problematic region from image1 and a reference image2. "
      "Please find the corresponding region in image2 that matches the same area as the "
      "problematic region from image1. Return only the bounding box coordinates in the "
      "format [xmin, xmax, ymin, ymax], No additional text, just the coordinates inside
          []"
   )

   messages = [
      {
         "role": "user",
         "content": [
            {"type": "text", "text": "image1 (problem region)"},
            {"type": "image", "image": problem_crop_path},
            {"type": "text", "text": "Reference image2"},
            {"type": "image", "image": image_A_path},
            {"type": "text", "text": prompt},
         ],
      }
   ]
```

Listing 3. The prompt for TagGrounder.

```
prompt = (
   f"Find the region in this image that best matches the product tag: \"{tag}\". "
   "Return ONLY the bounding box in Image in the strict format: "
   "[xmin, ymin, xmax, ymax]. No extra text."
)

messages = [
   {
      "role": "user",
      "content": [
         {"type": "text", "text": f"Image to search product tag: {tag}"},
         {"type": "image", "image": image_path},
         {"type": "text", "text": prompt},
      ],
   }
]
```

Listing 4. The prompt for Coordinator.

```
"""
You are the Coordinator Agent for an image restoration workflow.

The workflow has three sequential steps:

1. Inconsistency Detector: compare images and detect the difference region
  - Input: image_A (reference image path), image_B (target image path)
  - Output: bbox_B (difference region), prompt (problem description)

2. Reference Finder: find a clean reference region
  - Input: image_A (reference), image_B (target), bbox_B (problem region)
  - Output: bbox_A (reference region), cropped reference region, object_tag

3. ImageCritic: perform correction
  - Input: prompt, image_A, image_B, bbox_A, bbox_B, object_tag
  - Output: image_path (final corrected image), patch info

Additionally, there is an auxiliary specialist:
- TagGrounder: given an image and a user-provided product tag, locate a bbox in that
    image.

Workflow rules:
- Execute strictly in the order 1 - 2 - 3.
- After each step, you MUST ask the user:
  "Accept [STEP_NAME] result? (yes/no):"
- If the user answers "yes" or "y", continue to the next step.
- If the user answers "no":
  - You must wait for the user to provide either:
    (a) a new bbox in the format [xmin, ymin, xmax, ymax], or
    (b) a product tag (e.g. 'shoe', 'bag', 'logo') to re-locate the region.
  - If the user gives a bbox, this bbox has the highest priority and must override
      previous bbox_B or bbox_A.
  - If the user gives a product tag, you MUST call the TagGrounder specialist via the
      `delegate_to_specialist` tool to convert this tag into a bbox, then use this
      bbox in the current or next step.
- Always delegate concrete image-processing work to the appropriate specialist via
    the `delegate_to_specialist` tool.
- Maintain data flow between steps (outputs from a previous step feed into the next).
- When all three steps are done and the final restored image is produced, output:
  "Image restoration workflow completed!"
  and then stop.

Current task:
The user provides image_A and image_B paths. Please run the three-step correction
    workflow.
"""
```

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 2, 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025. 1

[3] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. arXiv preprint arXiv:2506.21416, 2025. 2

[4] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025. 2, 3

[5] Mi Jian, Yumeng Li, Bowen Wang, Xiaomin He, Zheyuan Gu, Qing Yan, Colin Zhang, and Lei Zhang. dots.ocr: Multilingual document layout parsing in a single vision-language model. https://github.com/rednote-hilab/dots.ocr/blob/master/assets/blog.md, 2025. 3, 6, 7, 8

[6] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[7] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. arXiv preprint arXiv:2504.16915, 2025. 2

[8] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rdle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollr, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 1, 2

[9] Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. arXiv preprint arXiv:2509.01977, 2025. 2

[10] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098, 2024. 2, 5

[11] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 2

[12] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871, 2025. 2

[13] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160, 2025. 2, 5