RegionRoute: Regional Style Transfer with Diffusion Model

Supplementary Material

1. Pseudo Ground-Truth Generation

We construct our training subset from the Grounded COCO dataset introduced in TokenCompose [4], which augments MS-COCO [3] image–caption pairs with object-level grounding. Concretely, we randomly sample 150 image–caption pairs from the training split to serve as our base data for LoRA-MoE fine-tuning [5] and analysis.

For each selected image, we first choose a single target object to undergo localized style transfer. The target object is sampled uniformly at random from the annotated instances in that image. We then extract the corresponding binary segmentation mask at the original image resolution. Given the target image and its mask, we synthesize a pseudo ground-truth (pseudo-GT) image in a desired artistic style using a diffusion-based image-to-image style transfer model. Specifically, we employ Flux.1-Kontext [2]: the original COCO image is used as the visual input, and the model is driven by a style instruction in the prompt, such as *"make the image into pixel-art style"*, *"make the image into cyberpunk style."*. The diffusion model processes the entire image, producing a fully stylized version that preserves the global scene layout and object semantics while changing the overall appearance according to the requested style. After obtaining this stylized image, we use the target object mask to extract the corresponding stylized region and composite it back onto the original image. Concretely, we crop the stylized image using the binary mask, and replace the masked area in the original image with the stylized content.

Because no existing dataset provides direct supervision for localized and mask-conditioned style transfer, our pseudo-GT construction offers a practical way to obtain spatially grounded training pairs. Although the composited pseudo-GT images may contain imperfect boundaries or slight inconsistencies due to mask inaccuracies, this does not hinder the learning process. The Flux.1-Kontext [2] model used in our fine-tuning exhibits strong semantic and object-level understanding, enabling it to correctly recognize and localize the target object even when the pseudo-GT supervision is not perfectly aligned. As a result, the model learns to generate smooth and coherent object boundaries during training and maintains robust object awareness at inference time, despite the approximate nature of the pseudo-GT masks.

To encourage stylistic diversity and avoid overfitting to a single visual domain, we generate pseudo-GT images in four representative styles: pixel art, cyberpunk, expressionism, and line art. For each of the 150 base images, we repeat the above procedure once for each style, using a style-specific instruction prompt (e.g., *"make the image into line-art style"*) while keeping the underlying image and target object fixed. This yields four distinct pseudo-GT variants per image, resulting in a total of 600 stylized training samples (150 images × 4 styles). Each training sample thus consists of: (i) the original image, (ii) a binary mask for the target object, (iii) the corresponding pseudo-GT image where only the target region has been stylized, and (iv) a regional style editing instruction. Some training samples are shown in Figure 2. This construction provides the spatially localized supervision necessary for learning style transfer that is both content-aware and region-specific.

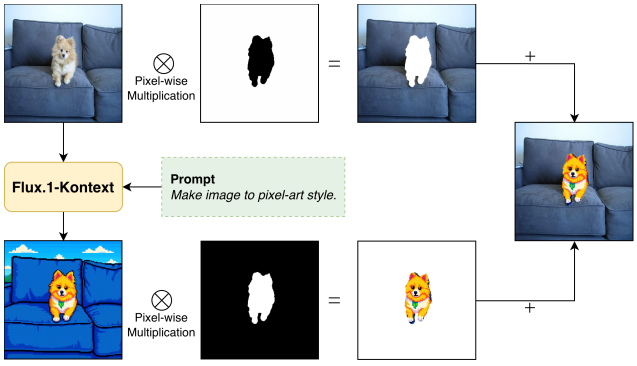

Figure 1. Illustration of the pseudo ground-truth (pseudo-GT) generation process. A diffusion-based style transfer model generates a fully stylized version of the image according to a given style prompt. The stylized region corresponding to the mask is then blended back into the original image using seamless cloning, producing an aligned input–target pair for localized style learning.

2. More Examples and Analysis

2.1. Qualitative Analysis of Localized Style Transfer

Figure 3 shows additional qualitative results on object categories that appear in our training set. These examples correspond to the standard in-distribution case. As can be observed, the model consistently (i) identifies the target object region, (ii) applies the desired style within the masked area, and (iii) preserves the structural and geometric attributes of the object. The stylized region aligns well with the original content, and transitions at the mask boundaries remain visually smooth. These results further confirm that the model behaves reliably on object categories it has been exposed to during pseudo-GT fine-tuning.

Figure 4 presents results on object categories that do not appear in the training set, providing additional evidence on the model's generalization behavior. The first four examples

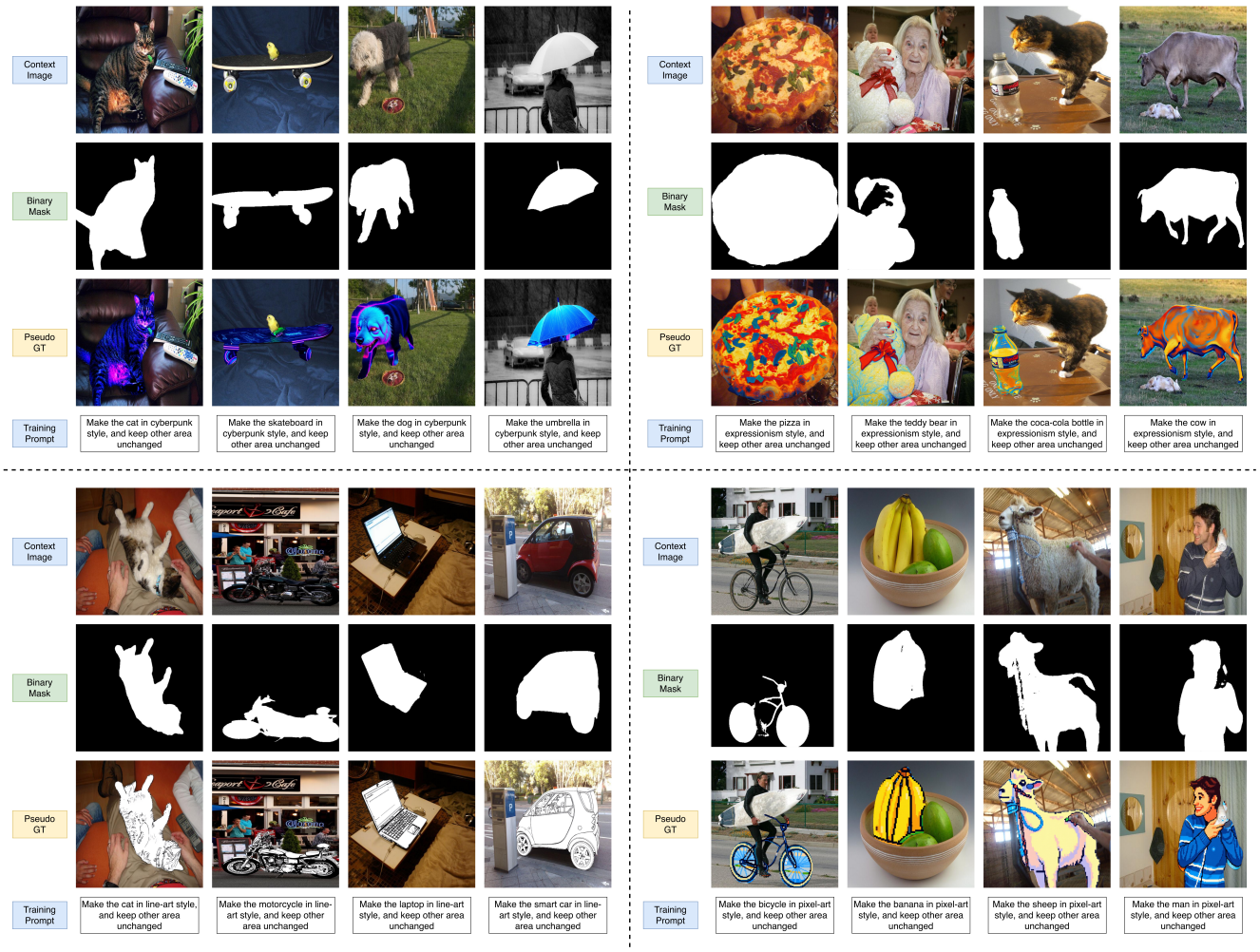

Figure 2. Examples from our pseudo ground truth dataset for localized style transfer. Each example consists of four elements: the original context image, the corresponding binary mask that specifies the target object, the pseudo ground truth image generated by applying a global style transformation to the entire image and compositing the stylized region back onto the original image, and the training prompt used for fine-tuning.

involve categories entirely absent from training. The last two examples involve categories where only a synonym-level variant appears in training (e.g., motorbike appears only as motorcycle, and aeroplane appears only as airplane).

Despite the lack of category-level supervision during training, the model is able to accurately locate the target object and apply the style transformation in a structurally coherent manner. The transferred style remains consistent with the request while maintaining correct object semantics and boundaries. These results suggest that the model is not tied to specific object classes but instead has learned a more general mechanism for localized style transfer. The strong semantic understanding of the underlying Flux.1-Kontext [2] model helps ensure that even imperfect pseudo-GT supervision does not significantly impair the model's ability to generalize to new object categories.

2.2. Qualitative Analysis for Ablation Study

Figure 5 provides additional qualitative results comparing several ablated variants of RegionRoute. These examples help illustrate how different components contribute to regional style editing. As a reference, the original Flux.1-Kontext [2] model is included. Without our designed modules, the base model tends to apply the requested style globally. While the stylization remains visually consistent, the absence of spatial selectivity highlights the need for additional mechanisms to enable region-level control.

When LoRA weights are removed from either the double stream blocks or the single stream blocks, the model is still able to stylize the target object in a number of cases, indicating that some degree of regional control is retained even when one LoRA branch is removed. However, these vari-

Figure 3. Additional examples of localized style transfer on objects that also appear in the training data. The model produces clean, structurally consistent stylization within the target regions, showing stable behavior on in-distribution categories.

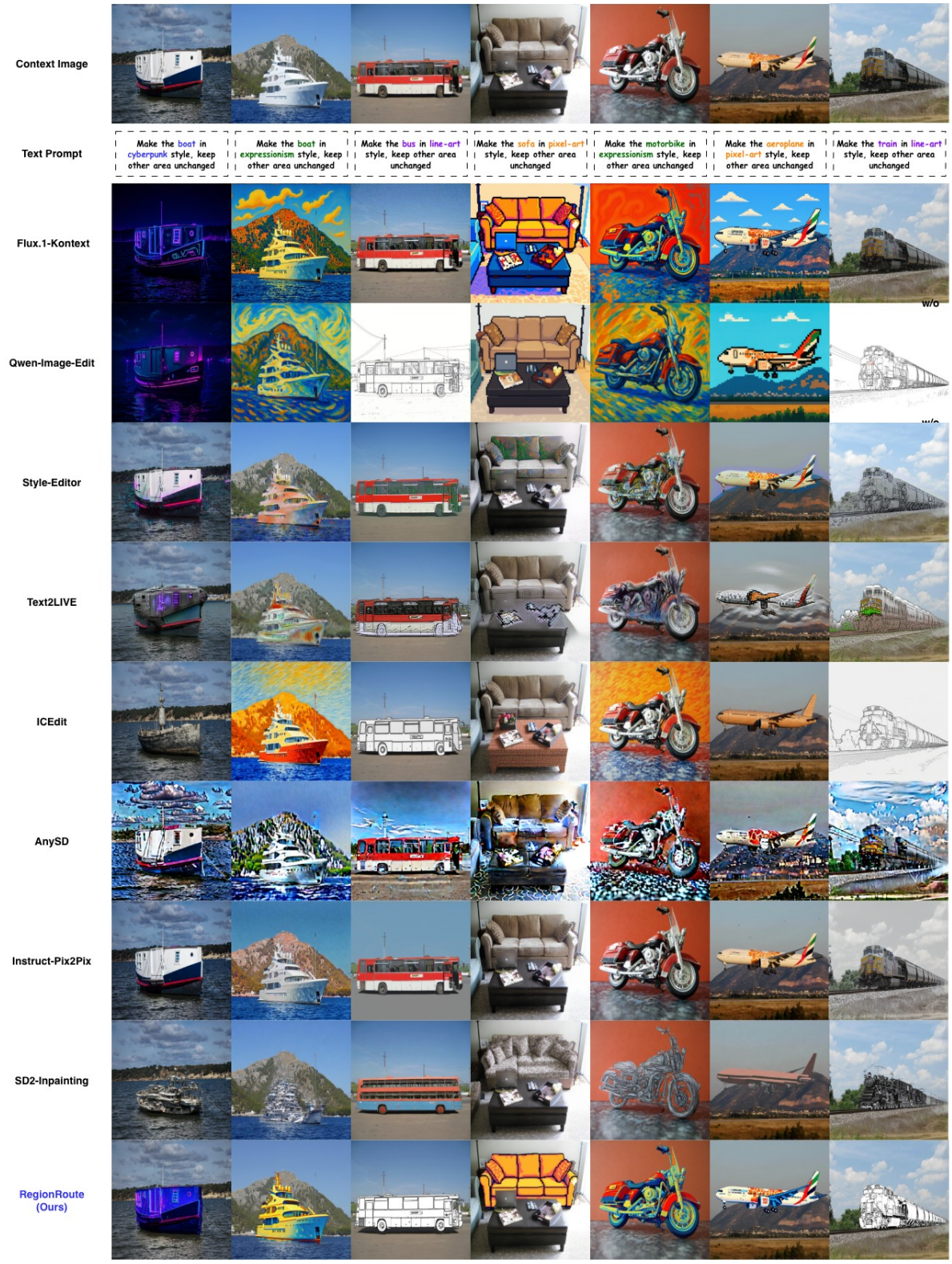

Figure 4. The first four columns show results on object categories that never appear in the training set. The last two show categories with only synonym-level presence (motorbike → motorcycle, aeroplane → airplane). The model successfully performs localized style transfer even without direct supervision for these object types, indicating strong semantic generalization.

ants more frequently exhibit style leakage into non-target areas and less stable boundary behavior. The fact that both successful and unsuccessful examples appear suggests that the single stream and double stream LoRA fine-tuning each make a meaningful contribution to suppressing global style propagation during diffusion, with both branches providing complementary support for achieving cleaner regional editing.

We study the effect of removing the two training objectives used in our fine-tuning process. When the cover loss $\mathcal{L}_{\text{cover}}$ is removed, the model can still perform regional editing in a number of cases, but incomplete stylization becomes more common and parts of the target object may retain their original appearance. This indicates that $\mathcal{L}_{\text{cover}}$ encourages more uniform coverage within the masked region, although certain examples remain successful even without this objective. When the focus loss $\mathcal{L}_{\text{focus}}$ is removed, the model also retains the ability to produce correct localized stylization in several cases, but it shows a higher tendency to introduce spillover into the background. These behaviors suggest that both losses contribute to improving the consistency and spatial precision of regional editing, with each offering complementary benefits while not being strictly required for the model to succeed in some cases.

Overall, the full RegionRoute model, which combines both routing streams and both training objectives, produces the most consistent results across all examples. However, the ablations collectively show that each module contributes to regional style editing to some extent, and each variant retains certain successful cases. The observations suggest that style localization emerges from the combined effects of all components, with each module providing complementary improvements rather than acting as an isolated requirement.

2.3. Qualitative Analysis for Attention Control

Figure 6 analyzes how RegionRoute and the Flux.1-Kontext [2] baseline allocate style-token attention during localized style-editing tasks. All visualizations correspond specifically to the attention from the style token toward spatial features at various diffusion steps, thereby revealing how each model propagates style information through the image.

Across all examples, RegionRoute demonstrates precise and stable routing of the style token to the intended object, producing well-defined attention patterns that match the object boundaries. In contrast, the baseline model consistently fails to form such a connection. Its style-token attention is diffuse, unstable, and often allocated to background regions or unrelated objects. This indicates that the baseline model lacks the capability to effectively anchor or attach the style token to the designated object, causing the style influence to spread globally rather than locally. As a result, the baseline's final outputs exhibit undesirable global stylization and style leakage into regions that should remain unchanged.

The temporal attention evolution further reinforces this observation: RegionRoute maintains object-aligned style-token attention from early to late diffusion steps, whereas the baseline's attention drifts or expands over time. These findings collectively demonstrate that RegionRoute provides a significant advantage in controlling the spatial propagation of style information by ensuring a persistent and accurate linkage between the style token and the target object.

2.4. Qualitative Analysis for Failure Cases

We summarize the major failure modes observed in our regional style transfer results, organized according to the underlying challenges. Figure 7 presents representative failure cases observed in our regional style transfer results. The first category involves small or hard-to-recognize objects. When the target object occupies very few pixels or is heavily occluded, such as a tiny potted plant, a small bottle, a distant person, or a table that is largely hidden behind other items, the model sometimes fails to identify the object, and the stylization does not occur. A related challenge arises when multiple instances of the target category appear in the same context image. In such cases, the model typically stylizes only the largest or most visually salient instances while neglecting smaller ones, suggesting that object-level recognition becomes unreliable for low-scale or low-visibility instances.

A second category of failures concerns unintended stylization of regions near the target object. When the target object is physically connected to or closely surrounded by other structures, the model may extend the stylization to these adjacent areas. Examples include stylizing portions of a bird perch along with the bird, a section of railway track along with a train, or the carpet underneath a chair. These behaviors indicate that spatial proximity and shared edges can sometimes be interpreted as part of a single coherent object.

A third type of failure involves objects that contain or enclose other objects, leading to ambiguous semantic grouping. When stylizing the car in an image where a person is sitting inside it, the person may also be stylized. Similarly, stylizing a sofa can unintentionally alter the appearance of pillows placed on it. These cases reflect ambiguity in the semantic interpretation of "the object" as described by the prompt.

Finally, we observe cases where only part of the target object is stylized. This typically occurs when certain regions of the object are visually ambiguous or difficult to separate from the background, resulting in incomplete localization.

Together, these failure modes illustrate the limitations that arise from object detection difficulty, small object scale,

Figure 5. We evaluate the impact of removing single-stream LoRA blocks, double-stream LoRA blocks, the cover loss $\mathcal{L}_{\text{cover}}$, and the focus loss $\mathcal{L}_{\text{focus}}$. The full RegionRoute model achieves the clearest and most accurate localized style transfer, while ablated variants exhibit style leakage, incomplete stylization, or degraded unedited object consistency.

occlusion, close physical coupling of objects, and ambiguous semantic boundaries implied by user prompts.

3. Details for VLM Evaluation

We evaluate controllability and semantic reliability using the Qwen2.5-VL-7B-Instruct [1] vision-language model

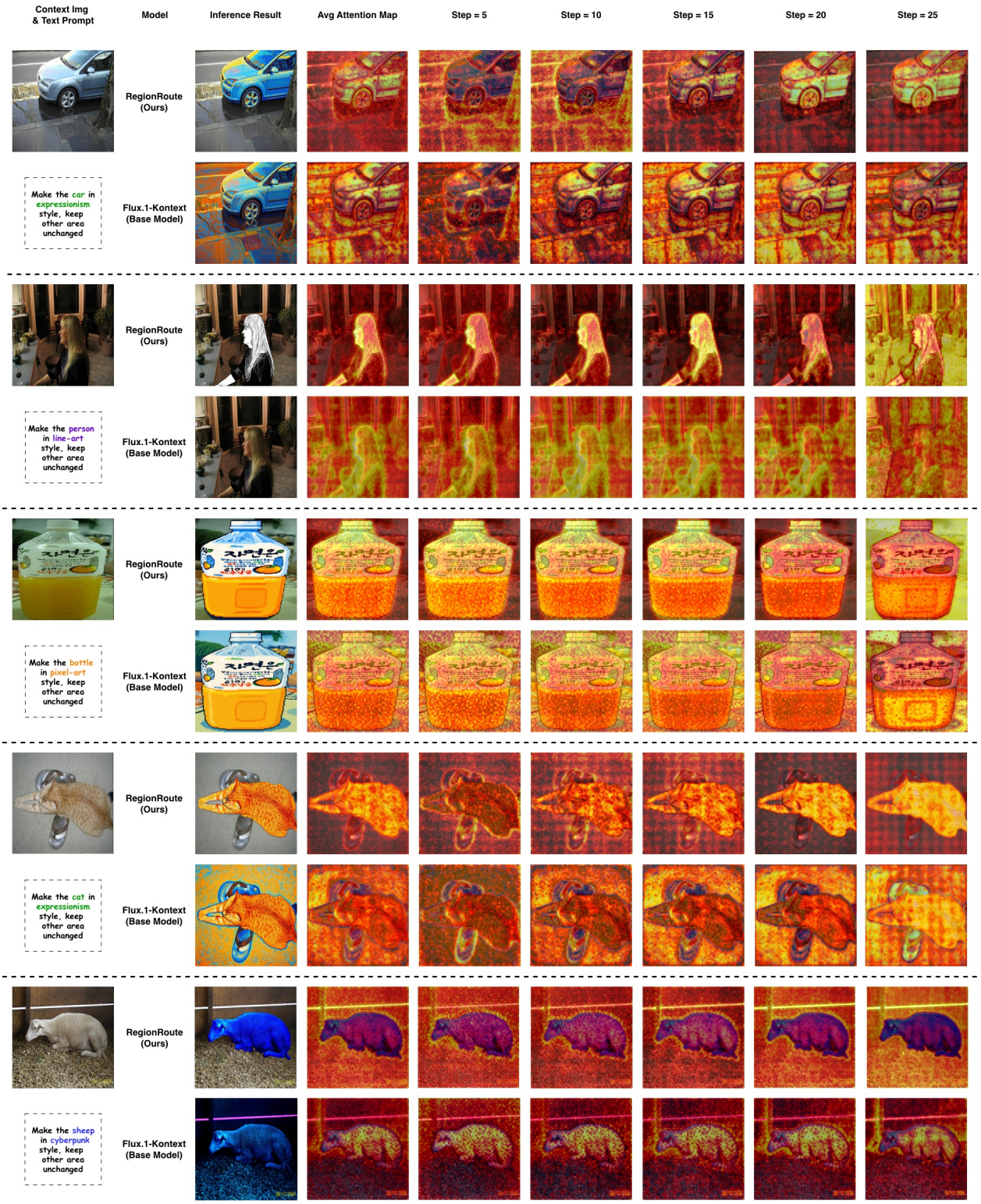

Figure 6. Comparison of the style-token attention maps for RegionRoute (ours) and the Flux.1-Kontext baseline. For each task, we show the edited output, the averaged style-token attention map, and the step-wise evolution of style-token attention at diffusion steps 5, 10, 15, 20, and 25.

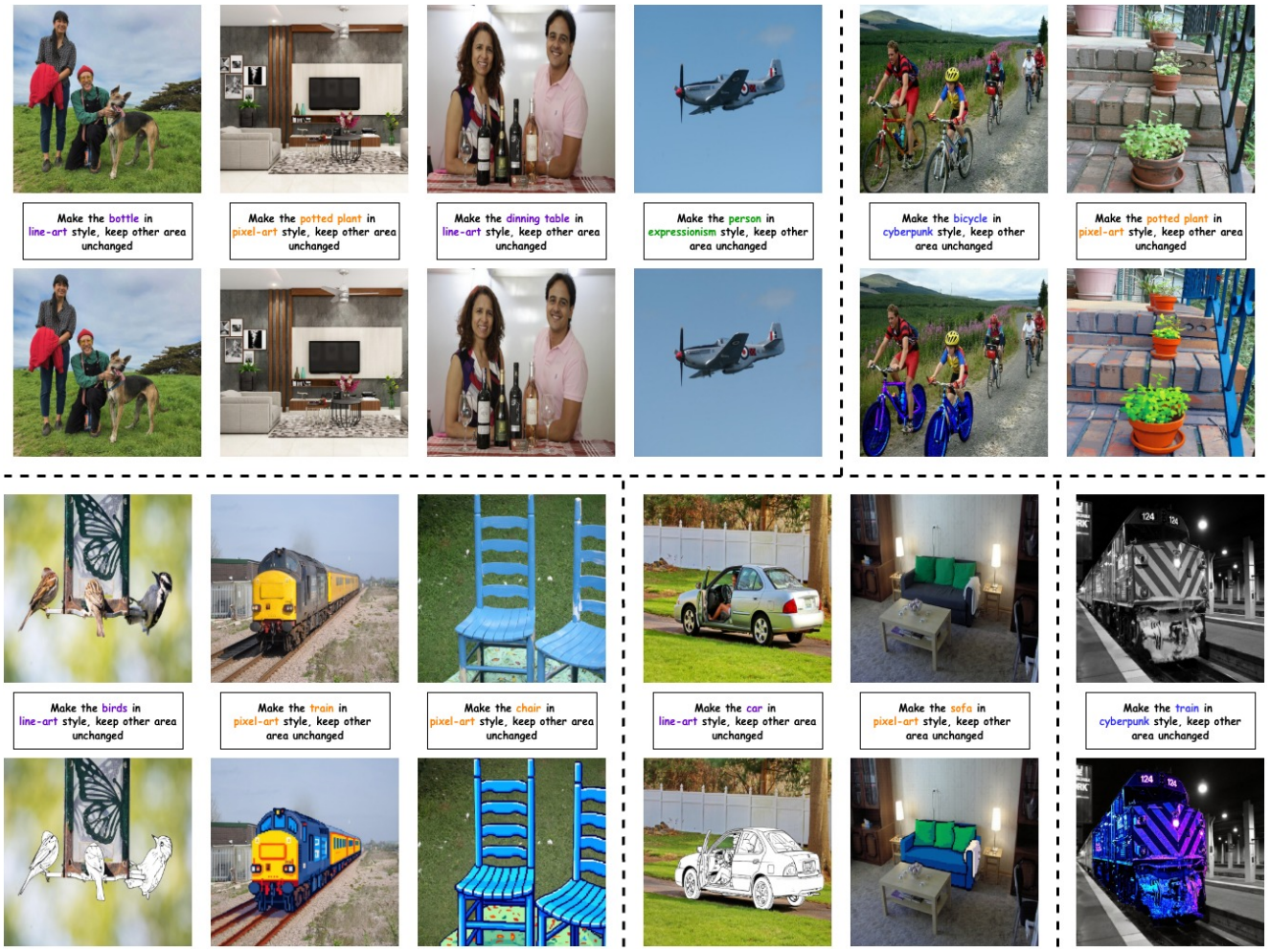

Figure 7. Examples illustrating several types of failure: (1) incomplete stylization when multiple target objects are present, where only the most salient instances are stylized; (2) failure to stylize very small or heavily occluded objects; (3) unintended stylization of regions adjacent to the target object, especially when objects are physically connected; (4) stylization leakage onto objects that are contained within the target object, such as a person inside a car or pillows on a sofa; and (5) partial stylization of a single object in cases where some object parts are difficult to recognize.

(VLM). For each edited image, the VLM answers four binary questions. These questions assess whether the target style is correctly applied to the intended object, whether the background remains unchanged, and whether the model inadvertently introduces stylistic attributes unrelated to the user instruction.

For each sample, we load the target style and additionally sample a *negative style*, which is randomly chosen from all styles except the target. The VLM then receives the edited image and answers:

- **Q1:** "Is the object in the *target* style?"
- **Q2:** "Is the background in the *target* style?"
- **Q3:** "Is the object in the *negative* style?"
- **Q4:** "Is the background in the *negative* style?"

The negative style used in Q3–Q4 is deliberately unrelated to the target style. These two questions operate as sanity checks for both the editing model and the VLM. If the VLM tended to respond "yes" for arbitrary style queries, Q3–Q4 would expose such failure by producing uniformly high scores. In practice, Q3 and Q4 yield consistently low probabilities across almost all methods, demonstrating that the VLM is not biased toward affirmative answers and is capable of distinguishing stylistic attributes reliably.

For each image, the script constructs the four questions, packages them with the image into the VLM input, and enforces strict binary outputs by instructing the model to answer only "yes" or "no." Responses are parsed and saved for aggregation, forming the probabilities reported in Table 2. This evaluation provides a scalable measure of object-level stylization accuracy, and background preservation.

References

[1] 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, 2025. 6

[2] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 2, 5

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[4] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8553–8564, 2024. 1

[5] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024. 1