

# The Supplementary of PrimeComposer: Faster Progressively Combined Diffusion for Image Composition with Attention Steering

Yibin Wang\*

yibinwang1121@163.com

School of Computer Science, Fudan University  
Shanghai, China

Jianwei Zheng

zjw@zjut.edu.cn

College of Computer Science and Technology, Zhejiang  
University of Technology  
Hangzhou, Zhejiang, China

Weizhong Zhang\*

weizhongzhang@fudan.edu.cn

School of Data Science, Fudan University  
Shanghai, China

Cheng Jin†

jc@fudan.edu.cn

School of Computer Science, Fudan University  
Shanghai, China

## 1 ALGORITHMS

The computation pipeline of our PrimeComposer and RCA are illustrated in Algorithm 1 and Algorithm 2, respectively.

## 2 PREPROCESSING THE TEST BENCHMARK

To effectively alleviate the unwanted artifacts appearing around the synthesized objects, we propose RCA to restrict the impact of object-specific tokens. To identify these tokens, we adjust the prompts by incorporating special tags, denoted as  $\langle ref \rangle$ , placed before and after target words through manual annotation. This adjustment facilitates the precise marking of object-specific tokens. For instance, the original caption 'a cartoon animation of a white fox in the forest' is adjusted to 'a cartoon animation of a  $\langle ref \rangle$  white fox  $\langle ref \rangle$  in the forest'. Before the composition process, we identify and record the indices of these specially tagged tokens for each input sample, ensuring targeted and effective region-constrained attention during synthesis. We will release the preprocessed benchmark to the public.

## 3 ADDITIONAL INFERENCE TIME COMPARISON

Given that most training baselines are primarily trained in the photorealism domain, we exclusively compare the inference speed with them within photorealism domains using an NVIDIA A100 40GB PCIe. As depicted in Table 1, our PrimeComposer demonstrates faster inference speed than all the considered baselines, underscoring our superior efficiency in this task.

\*Equal contribution.

†Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3680848>

## Algorithm 1 PrimeComposer

**Input:** The background image  $I^{bg}$ , the object image  $I^{obj}$ , the foreground mask  $M^{fg}$ , the object mask  $M^{obj}$ , the caption embedding  $\epsilon$ , the thresholds  $\alpha$ , the Correlation Diffuser  $\theta_{CD}$ , the LDM  $\theta_{LDM}$

**Output:** The composite image  $I^*$

```

1:  $z_0^{bg*} = \text{VAE-Encoder}(I^{bg}); z_0^{fg*} = \text{VAE-Encoder}(I^{fg*})$ 
2: for  $t = 1, \dots, T$  do
3:    $z_t^{bg*} \leftarrow \text{Inverse}(z_{t-1}^{bg*}, t - 1)$ 
4:    $z_t^{fg*} \leftarrow \text{Inverse}(z_{t-1}^{fg*}, t - 1)$ 
5: end for
6:  $\text{noise} \sim \mathcal{N}(0, 1)$ 
7:  $z_t^{init} \leftarrow z_t^{fg*} \odot M^{fg} + z_t^{bg*} \odot (1 - M^{fg}) + \text{noise} \odot (M^{obj} \text{ xor } M^{fg})$ 
8: for  $t = T, \dots, 1$  do
9:   if  $t \leq \alpha T$  then
10:     $z_t^{pc*} \leftarrow z_t^{fg*} \odot M^{obj} + z_t^{bg*} \odot (1 - M^{obj})$ 
11:     $z_{t,obj} \leftarrow \text{Segement}(z_t, M^{obj})$ 
12:     $\{A_t^{cross}, A_t^{obj}\} \leftarrow \theta_{CD}(z_{t,obj}, z_t^{pc*}, \epsilon)$ 
13:     $z_{t-1} \leftarrow \theta_{LDM}(z_t, \{A_t^{cross}, A_t^{obj}\}, \epsilon) \odot M^{fg} + z_{t-1}^{bg*} \odot (1 - M^{fg})$ 
14:   else
15:     $z_{t-1} \leftarrow \theta_{LDM}(z_t) \odot M^{fg} + z_{t-1}^{bg*} \odot (1 - M^{fg})$ 
16:   end if
17: end for
18:  $I^* = \text{VAE-Decoder}(z_0)$ 
19: return  $I^*$ 

```

Methods	Blended [1]	SDEdit [2]	Paint [3]	DIB [5]	Ours
PI in <i>Photorealism</i>	22.41 sec	20.68 sec	19.74 sec	18.57 sec	<b>16.23 sec</b>

**Table 1: Inference time comparison with the SOTA training baselines on photorealism domains. PI and sec mean Per Image and seconds, respectively.**



**Figure 1: Additional cases of challenges in preserving the objects' appearance (left) and synthesizing natural coherence (right). The problematic areas of coherence are indicated by red dotted lines**

#### Algorithm 2 Region-constrained Cross-Attention

**Input:** The input image feature  $F$ , the caption embeddings  $\epsilon$ , and the object mask  $M^{obj}$

**Output:** The output image feature  $F_{out}$

- 1: Get image queries  $Q$ , text keys  $K$ , text values  $V$  via linear projections of  $F$  and  $\epsilon$
- 2: Resize  $M^{obj}$  to match the spatial size of  $F$
- 3:  $A \in \mathbb{R}^{h \times w \times p} \leftarrow Q \cdot K^T / \sqrt{d}$
- 4: Initialize a mask  $\hat{A} \in \mathbb{R}^{h \times w \times p}$
- 5: **for**  $k = 1, \dots, p$  **do**
- 6:   **if** the  $k$ -th text embedding corresponds to the object-specific word **then**
- 7:      $\hat{A}^k \leftarrow A^k \odot M^{obj} + (-\infty) \odot (1 - M^{obj})$
- 8:   **else**
- 9:      $\hat{A}^k \leftarrow A^k$
- 10:   **end if**
- 11: **end for**
- 12:  $F_{out} \leftarrow \text{Softmax}(\hat{A}) \cdot V$
- 13: **return**  $F_{out}$

## 4 SOCIETAL IMPACTS

The widespread use of PrimeComposer in image composition has some interesting effects on how we create and see pictures. One

potential impact is that it might lead to misunderstandings or misrepresentations of different cultures. People could unintentionally use this tool to mix and match cultural symbols, possibly spreading stereotypes or watering down the true meaning behind these symbols. To avoid this, it's important to educate users on cultural sensitivity. Another thing to think about is how easy it becomes to create fake pictures that look real. As more and more people use tools like PrimeComposer, it might get harder to tell if a picture is genuine or if it has been altered. This could make it challenging for people to trust what they see online and might require us to be more careful and critical when looking at pictures. The way we think about art and creativity might also change. With tools like PrimeComposer, anyone can create unique and diverse images easily. While this is exciting, it might challenge traditional ideas about who gets credit for creating something. It raises questions about who owns the rights to these images and what it means to be a creator in a world where machines assist in the creative process.

In essence, PrimeComposer opens up new possibilities for creativity, but it also brings up important issues around cultural understanding, trust in images, and the evolving nature of art and creativity in a tech-driven world. Addressing these concerns ensures that technology contributes positively to how we express ourselves and understand the world around us.

## 5 ADDITIONAL CASES OF CHALLENGES

Additional cases of challenges the current SOTA method encountered are exhibited in Fig. 1.

## 6 ADDITIONAL QUALITATIVE RESULTS

Further qualitative comparisons of image composition across various domains are exhibited in Fig. 2 and 3.

## 7 ADDITIONAL ABLATION

Additional ablation studies are exhibited in Fig. 4.

## 8 VISUALIZATIN OF OUR EXTENDED CFG

Classifier-free guidance is extended in our work to reinforce the steering effect of the infused prior weights in foreground generation. The extended CFG is defined as

$$\hat{\epsilon} = \epsilon_{\theta}(\mathbf{z}_t|\emptyset) + s[\epsilon_{\theta}(\mathbf{z}_t|c, f) - \epsilon_{\theta}(\mathbf{z}_t|f) + \underbrace{\epsilon_{\theta}(\mathbf{z}_t|c, f) - \epsilon_{\theta}(\mathbf{z}_t|c)}_{SM}].$$

To qualitatively assess its effectiveness, we present the averaged saliency map (SM) in Fig. 5. These visualizations showcase the extended CFG facilitates establishing coherent relations and preserving the object’s appearance. This phenomenon aligns with our design philosophy.

## 9 LIMITATION

Firstly, our approach faces a common challenge in the field, which is the limited ability to freely control the object’s viewpoint. While efforts have been made to address this concern, as demonstrated in [4], it typically involves resource-intensive training processes. Secondly, our current methodology cannot seamlessly integrate multiple objects into the background simultaneously. This aspect poses a significant challenge in application scenarios where compositions involve more complex scenes or diverse elements. Thirdly, although our method demonstrates accelerated inference times compared to the previous approaches, we acknowledge that the current speed of reasoning may not fully meet the demands of practical applications. We recognize the need for further optimizations to enhance the efficiency of our method and make it more suitable for real-time or near-real-time applications.

## REFERENCES

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*. 18208–18218.
- [2] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*.
- [3] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*. 18381–18391.
- [4] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. 2023. ControlCom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040* (2023).
- [5] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. 2020. Deep image blending. In *WACV*. 231–240.



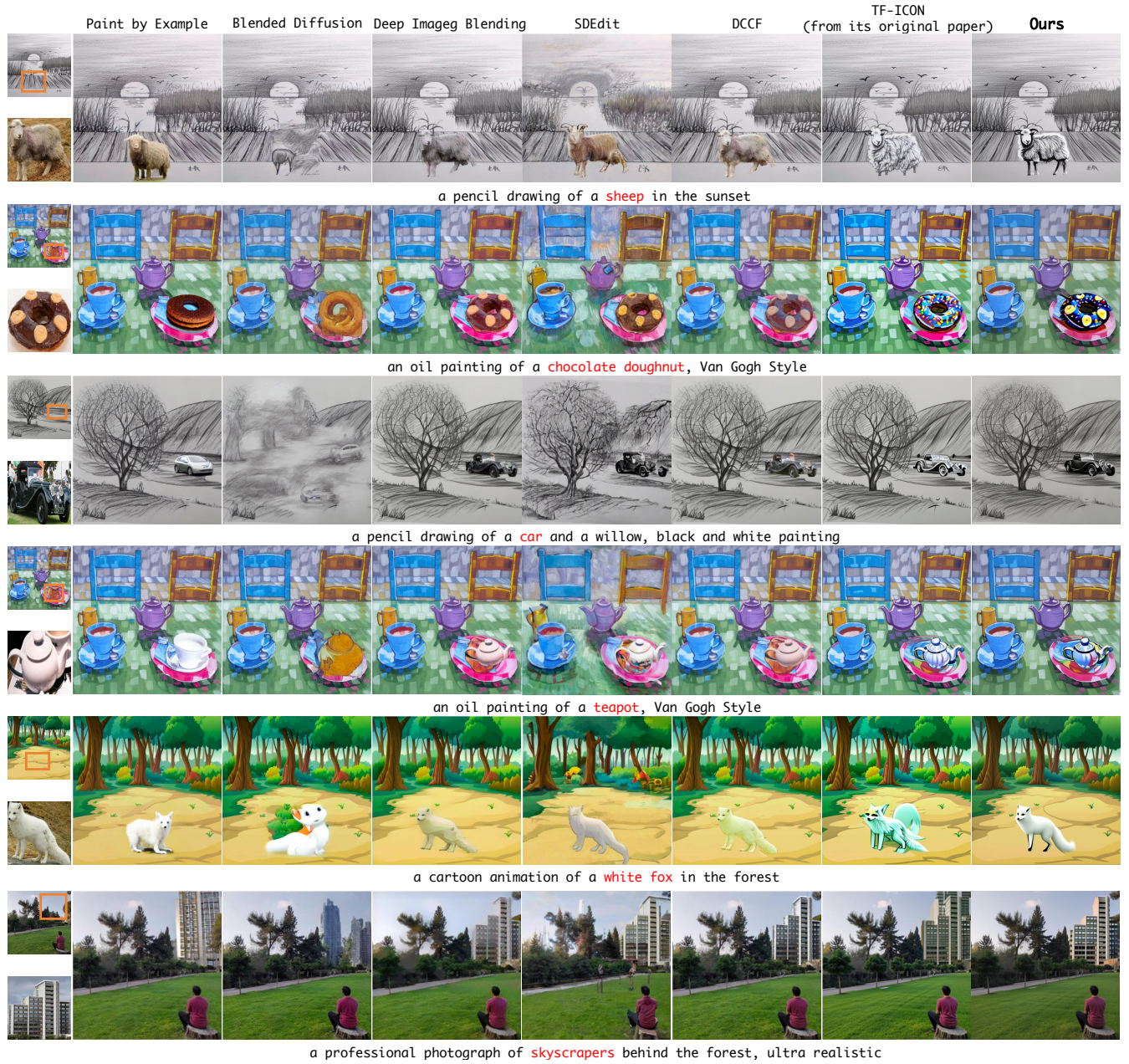


Figure 2: Additional qualitative comparison with SOTA baselines in cross-domain image composition. All the results of TF-ICON come from its originary paper.



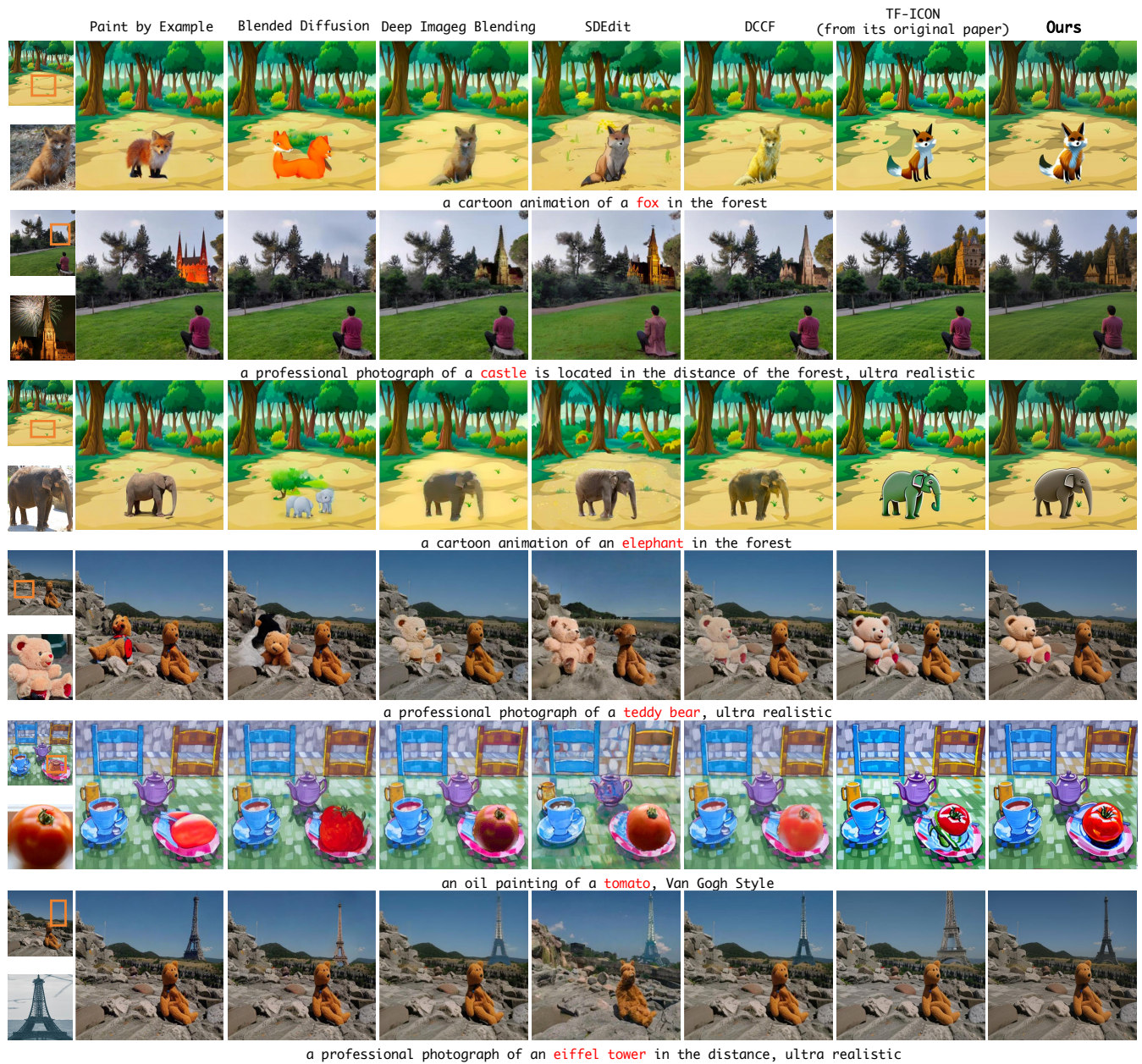


Figure 3: Additional qualitative comparison with SOTA baselines in cross-domain image composition. All the results of TF-ICON come from its originary paper.



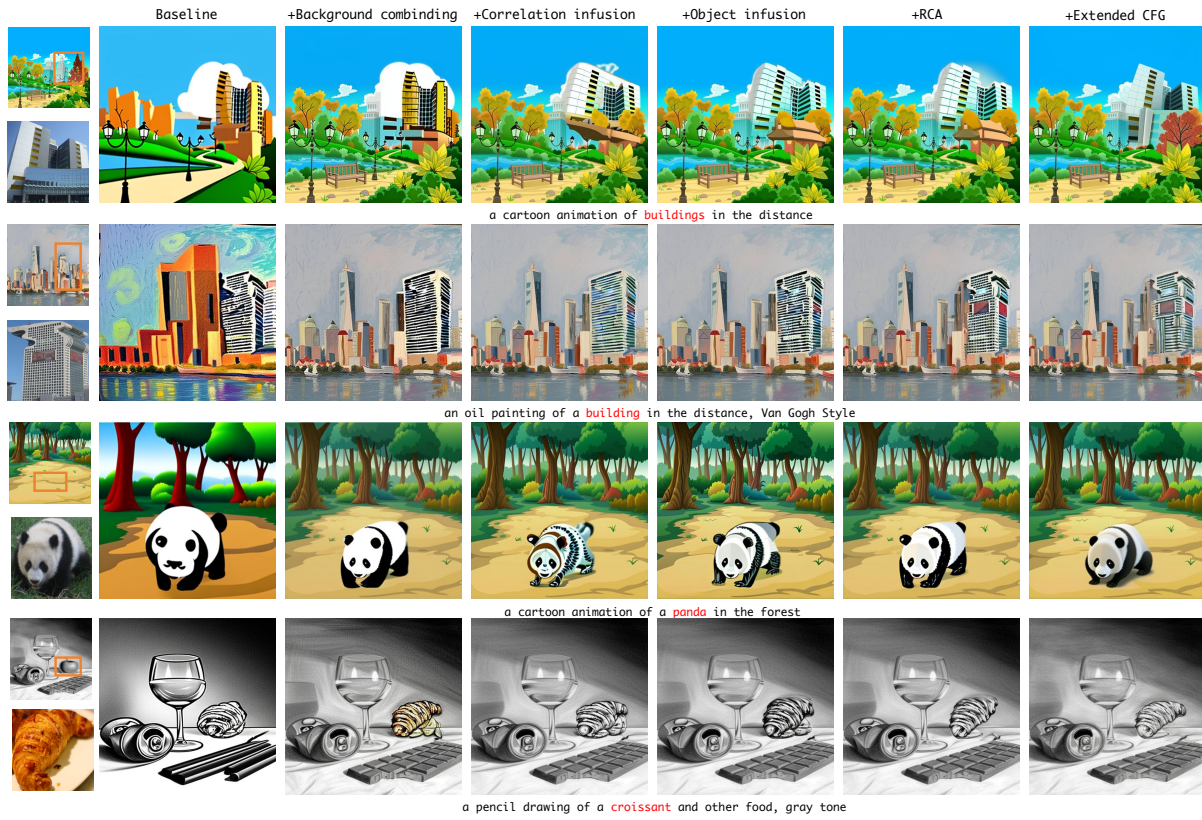


Figure 4: Additional ablation study of different variants of our framework. RCA: Region-constrained Cross-Attention. CFG: Classifier-free Guidance.

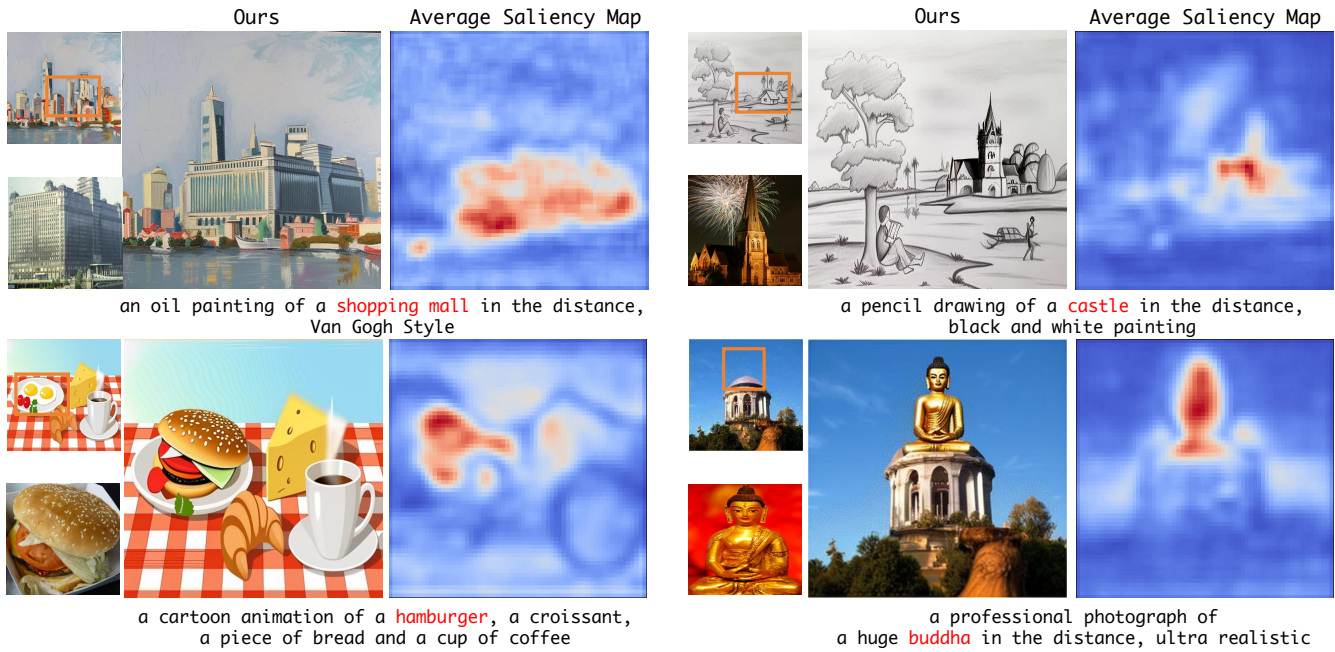


Figure 5: The visualization of average saliency maps derived from our extended classifier-free guidance.