Supplementary Materials: Generative Active Learning for Image Synthesis Personalization

Anonymous Author(s)*

To provide a more comprehensive understanding of the method, we have included additional details in the following sections. The source code can be accessed at https://github.com/(open_upon_acceptance)/.

1 ADDITIONAL IMPLEMENTATION DETAILS

For a fair comparison, we utilize the third-party implementation of HuggingFace for all SOTA methods in our experiments. We use Stable Diffusion 1.5 version as the pre-trained model. Besides, we apply 50 steps of DDPM (Denoising Diffusion Probabilistic Model) sampling with a guidance scale of 7.5 in all experiments. All experiments are conducted using A-100 GPUs.

In each round, we adopt DreamBooth to retrain the diffusion model. Specifically, the parameters of the text Transformer are frozen and the U-Net diffusion model is fine-tuned with a learning rate 5×10^{-6} and the batch size is 1. The training step for content learning is 800 steps (wooden pot for 600 steps to avoid overfitting) and style is 500 steps.

2 TEXT PROMPTS FOR ADDITIONAL TRAINING SAMPLES GENERATION

Here we provide the prompts used to synthesize the images for iterative training in Table 1 and Table 2. In each round, 3 prompts paired images will be selected as additional training samples.

3 STOPPING CRITERION

Figure 1 illustrates a case study to show the effectiveness of the early stopping setting. We can find that the personalized model performs well toward text fidelity and image fidelity in the termination round. However, beyond this point, introducing non-informative samples results in a decline in performance. We can find the prominent edge features and unnatural brushstrokes in the generated outputs. These deteriorated results suggest that the model struggles to maintain its initial level of quality when confronted with less informative prompts. The stopping criterion plays an important role in preventing the inclusion of samples that could negatively impact its output quality when there are few anchor directions to explore.

4 ABLATIONS ON OBJECT-DRIVEN ISP

We provide ablations on object-driven ISP. As shown in Figure 2, we can draw similar conclusions on these hyperparameter selections.

5 DETAILS ON USER STUDY

In the conducted user study, we perform 2 pair comparison tasks between Final Round (Ours) and Round 1, Final Round (Ours) and Oracle (Human + Balance). For each of 10 Objects and 5 Style, 20 queries are presented, which results in a total of $2\times(10\times20+5\times20)$ = 600 queries. In each query, 2 questions are included for assessing the object/style and text fidelity, respectively. Totally, 4800 responses are collected from 8 participants.

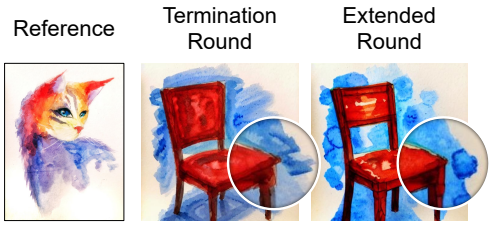


Figure 1: Comparison between the termination round and extended round in the iterative training process. The details are better preserved in the termination round.

The user study interface is shown in Figure 3. It includes the reference object/style image, prompt text, 2 image candidates, 2 questions with 3 options for each, instruction on tasks, and other functional components. The order of the queries and image candidates is randomly set for unbiased comparisons.

Questions:

- Task1. Choose one image that most aligns with the style of / object in the reference image.
  - Options: `Image A`, `Equal`, `Image B`
- Task2. Choose one image that most describes the Prompt text.
  - Options: `Image A`, `Equal`, `Image B`

Instructions:

- Task 1: First view the reference object/style image, then choose one image (`Image A`, or `Image B`) that most aligns with the style of / object in the reference image. If you cannot determine, choose `Equal`.
- Task 2: Then view the Prompt text. Choose one image (`Image A`, or `Image B`) that most describes the Prompt text. If you cannot determine, choose `Equal`.

6 ADDITIONAL RESULTS

This section provides samples of generated images. Figure 4 and Figure 5 show the generated images in each round based on object and style references. Figure 6 gives a more visualized comparison with SOTA methods. Figure 7, Figure 8, and Figure 9 show the generalization of our method on a wide range of prompts.

Table 1: The anchor prompts of object personalization to generate synthetic samples.

"S^* in park"	"S^* near a pool"	"S^* on street"
"S^* in a forest"	"S^* in a kitchen"	"S^* on the beach"
"S^* in a restaurant"	"S^* on snowy ground"	"S^* on the moon"
"S^* in desert"	"S^* in library"	"S^* under the night sky"
"S^* in front of the castle"	"S^* on the bridge"	"S^* next to waterfall"
"S^* in the cave"	"S^* at a concert venue"	"S^* on a balcony"

Table 2: The anchor prompts of style personalization to generate synthetic samples.

"A tie in style S^*"	"A table in style S^*"	"A broccoli in style S^*"
"A box in style S^*"	"Candy in style S^*"	"A billboard in style S^*"
"A drum in style S^*"	"A cell phone in style S^*"	"An empty bottle in style S^*"
"A glass in style S^*"	"A golden key in style S^*"	"A koala in style S^*"
"A toy in style S^*"	"A skateboard in style S^*"	"An apple on the table in style S^*"
"A bridge in style S^*"	"Beach scene in style S^*"	"A banana on the table in style S^*"

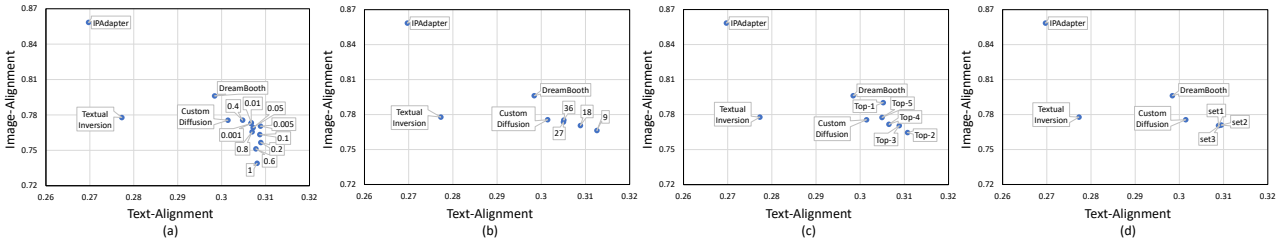
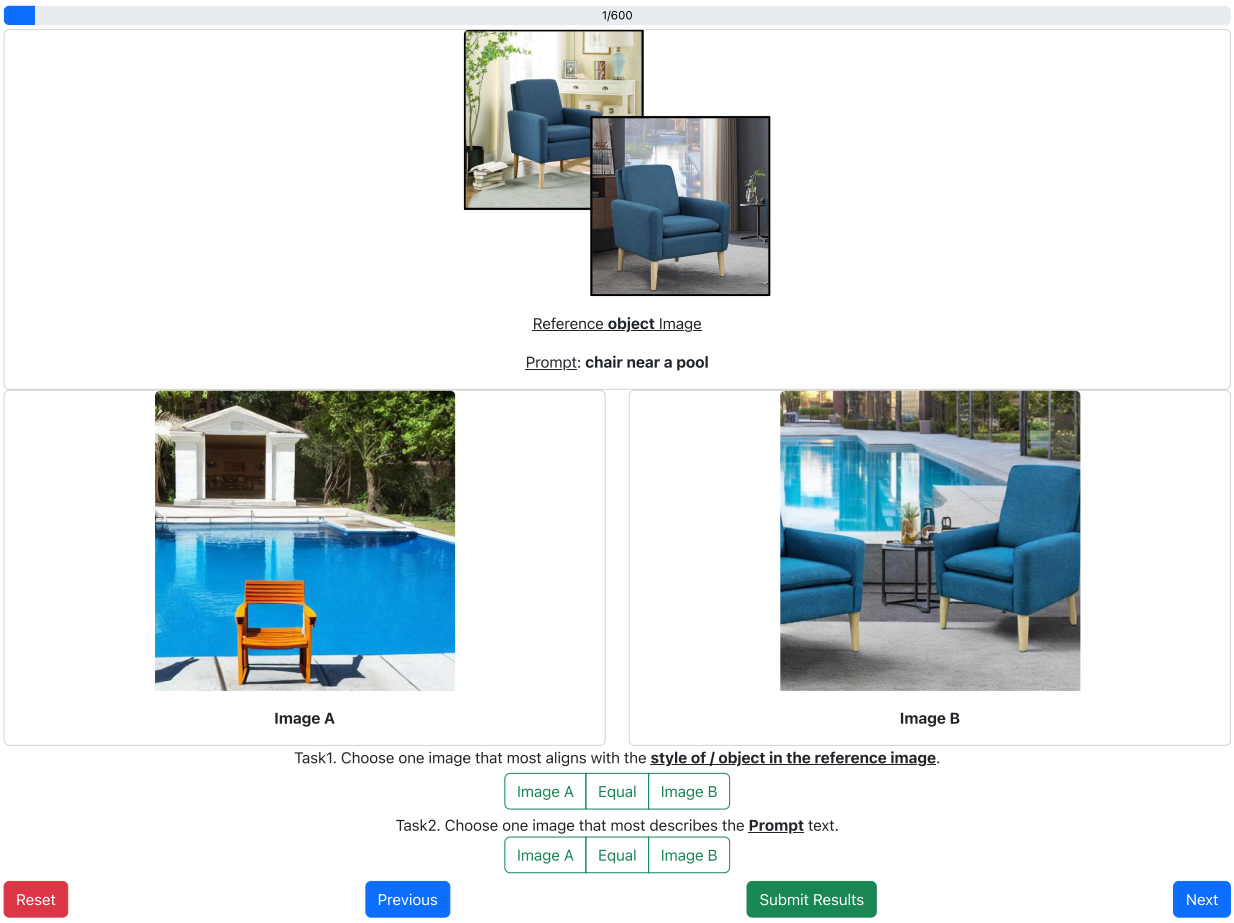


Figure 2: Illustration of ablation experiments on object-driven ISP. (a) Variation in performance with the parameter λ. (b) Effects of different anchor set sizes. (c) Impact of selecting the top-k prompts per iteration. (d) Results from varying the prompt composition within the anchor set.



Instructions

Task 1: First view the reference object/style image, then choose one image (i.e. **Image A**, **Image B**) that most aligns with the **style of / object** in the reference image. If you cannot determine, choose **Equal**.

Task 2: Then view the Prompt text. Choose one image (i.e. **Image A**, **Image B**) that most describes the **Prompt** text. If you cannot determine, choose **Equal**.

Notes:

- After selecting answers in two tasks, click the **Next** button to move to the next question.
- You can click the **Previous** button to go back to the previous question.
- Click the **Reset** button to clear all your selections and restart.
- After finishing all questions, please click the **Submit Results** button to submit your results.

Figure 3: The interface for user preference study. The references and a generation prompt are shown at the top. Participants are required to select their preferred output by following the given instructions.

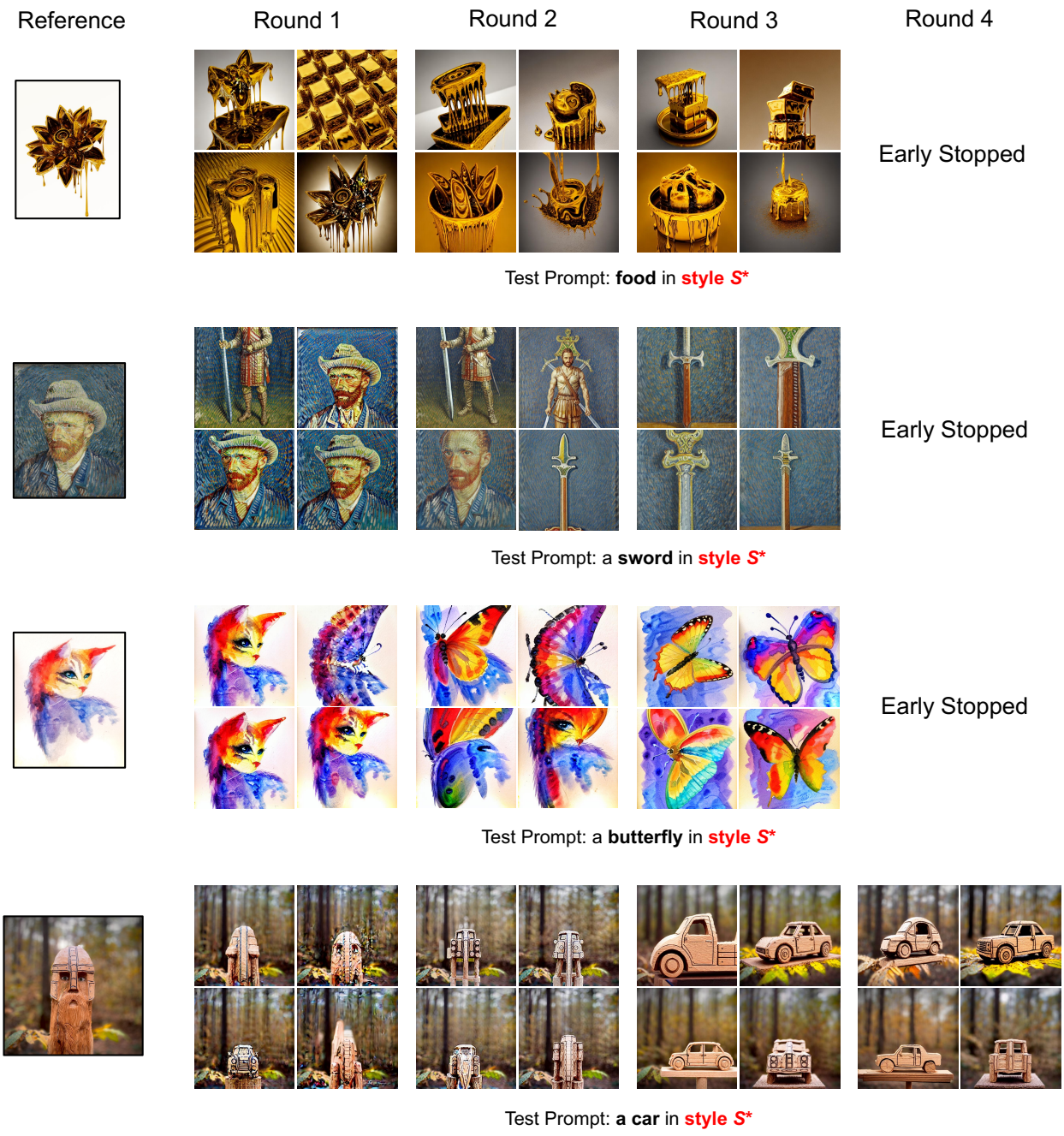


Figure 4: Examples of images generated in each round. The model in round 1 fails to produce the desired subjects, including food, sword, butterfly, and car. These subjects are successfully generated after using generative active learning.

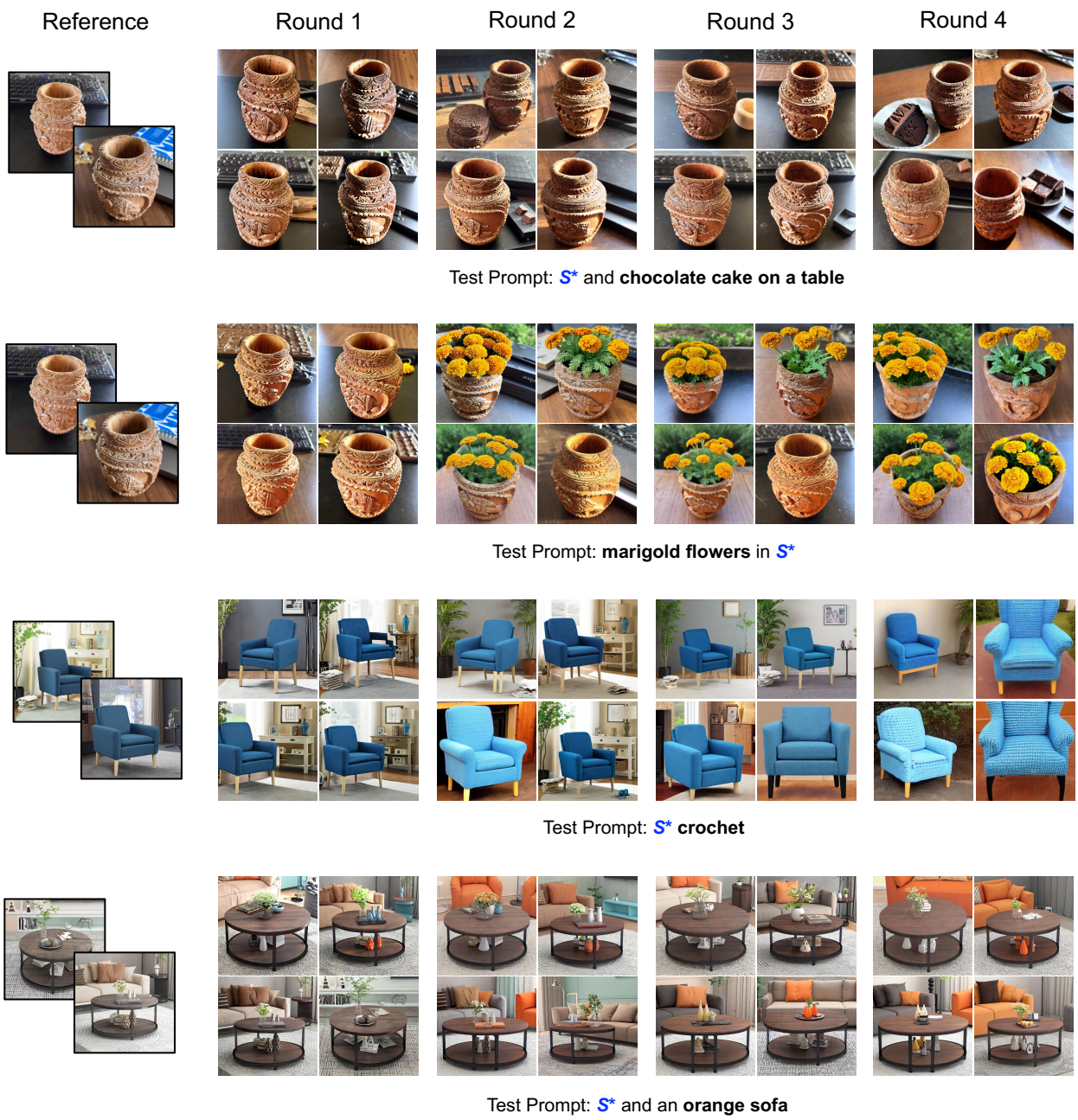


Figure 5: Examples of images generated in each round. The model in round 1 fails to produce the desired subjects, including chocolate cake, marigold flowers, crochet, and orange sofa. These subjects are successfully generated after using generative active learning.

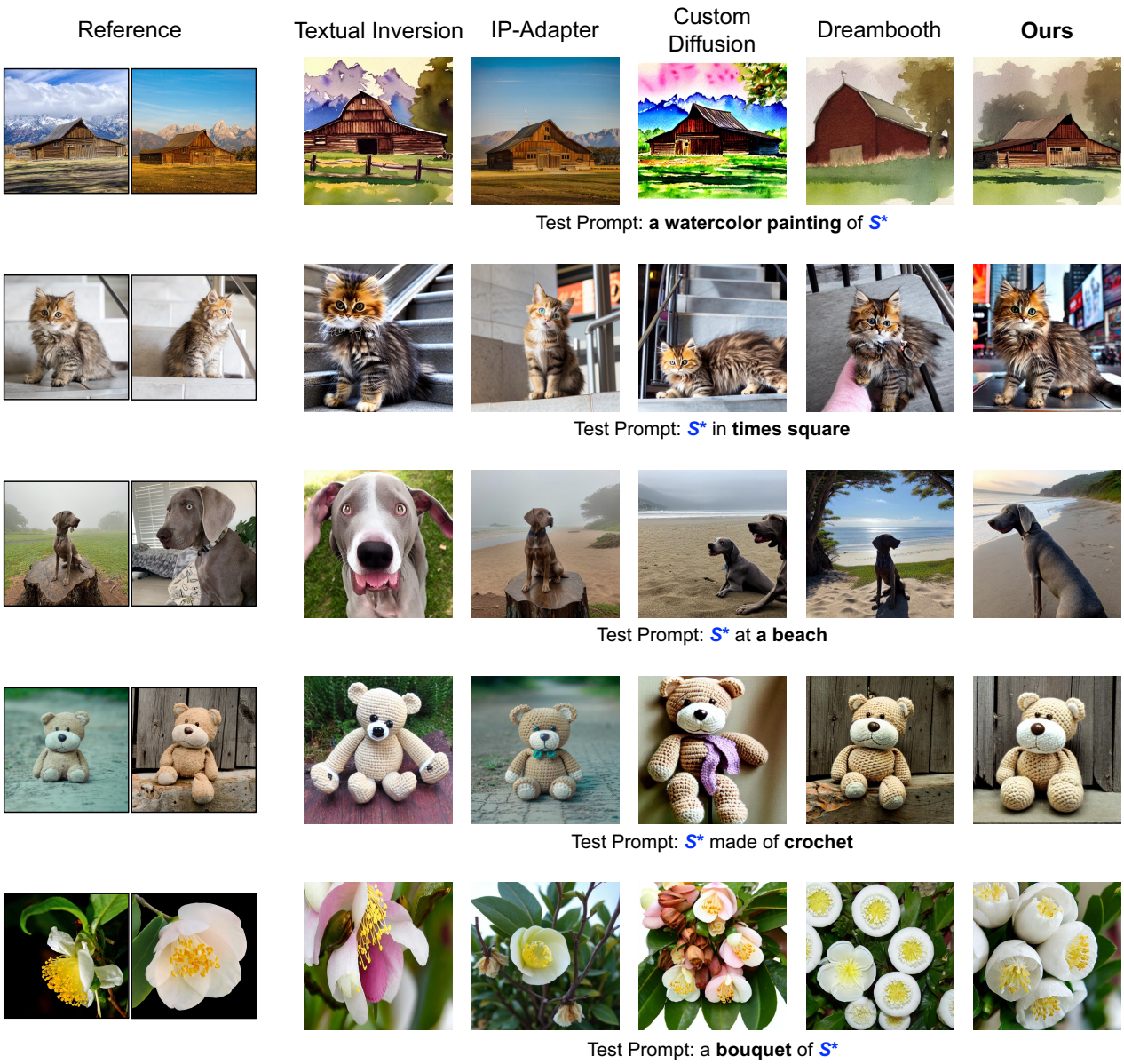


Figure 6: Comparison with other SOTA methods. Our method achieves better text and object fidelity.

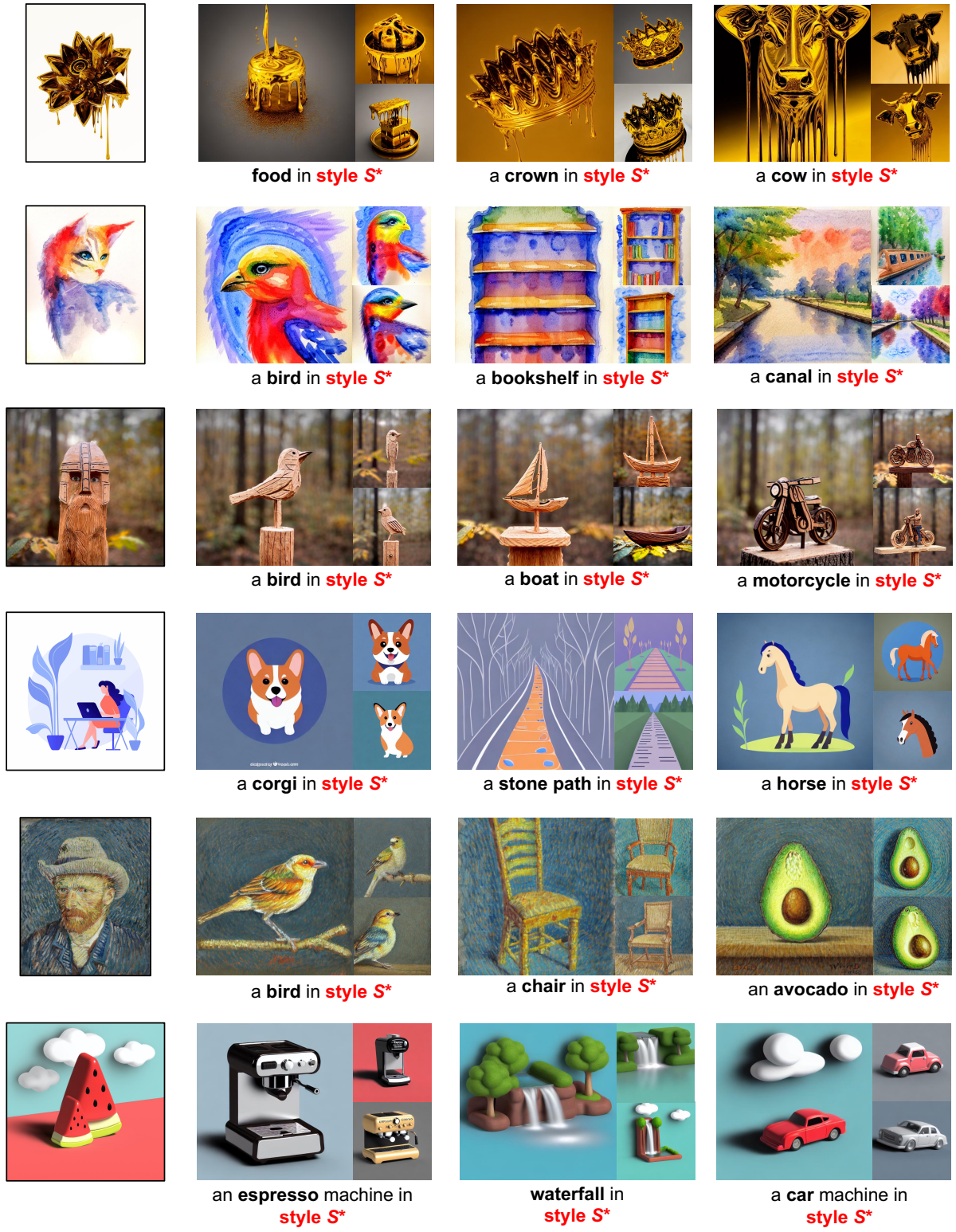


Figure 7: Generated images using style references. Our method generalizes well on a diverse range of prompts.

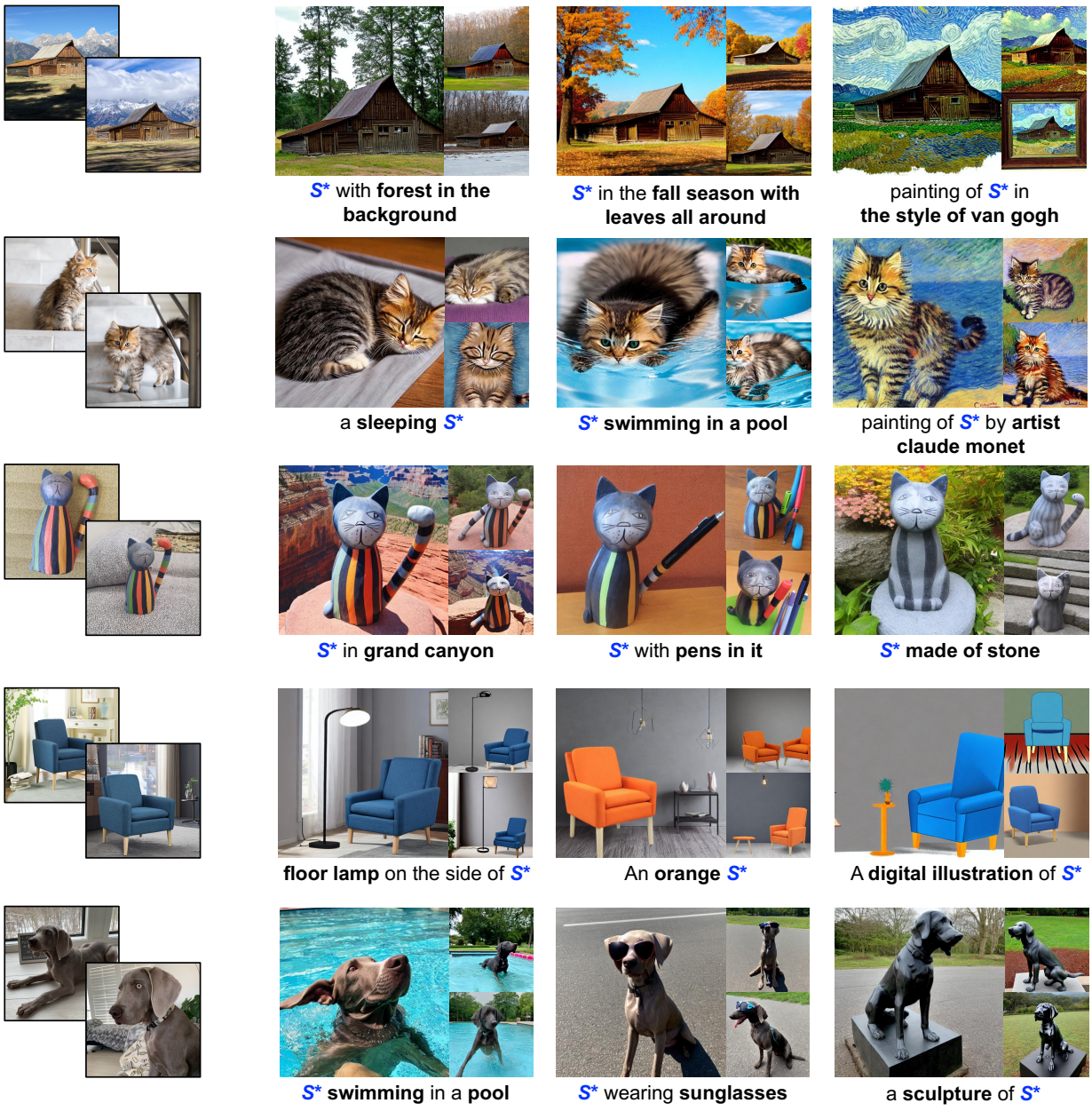


Figure 8: Generated images using object references. Our method can be used for a variety of personalization purposes, such as background replacement, artistic rendering, and attribute editing.

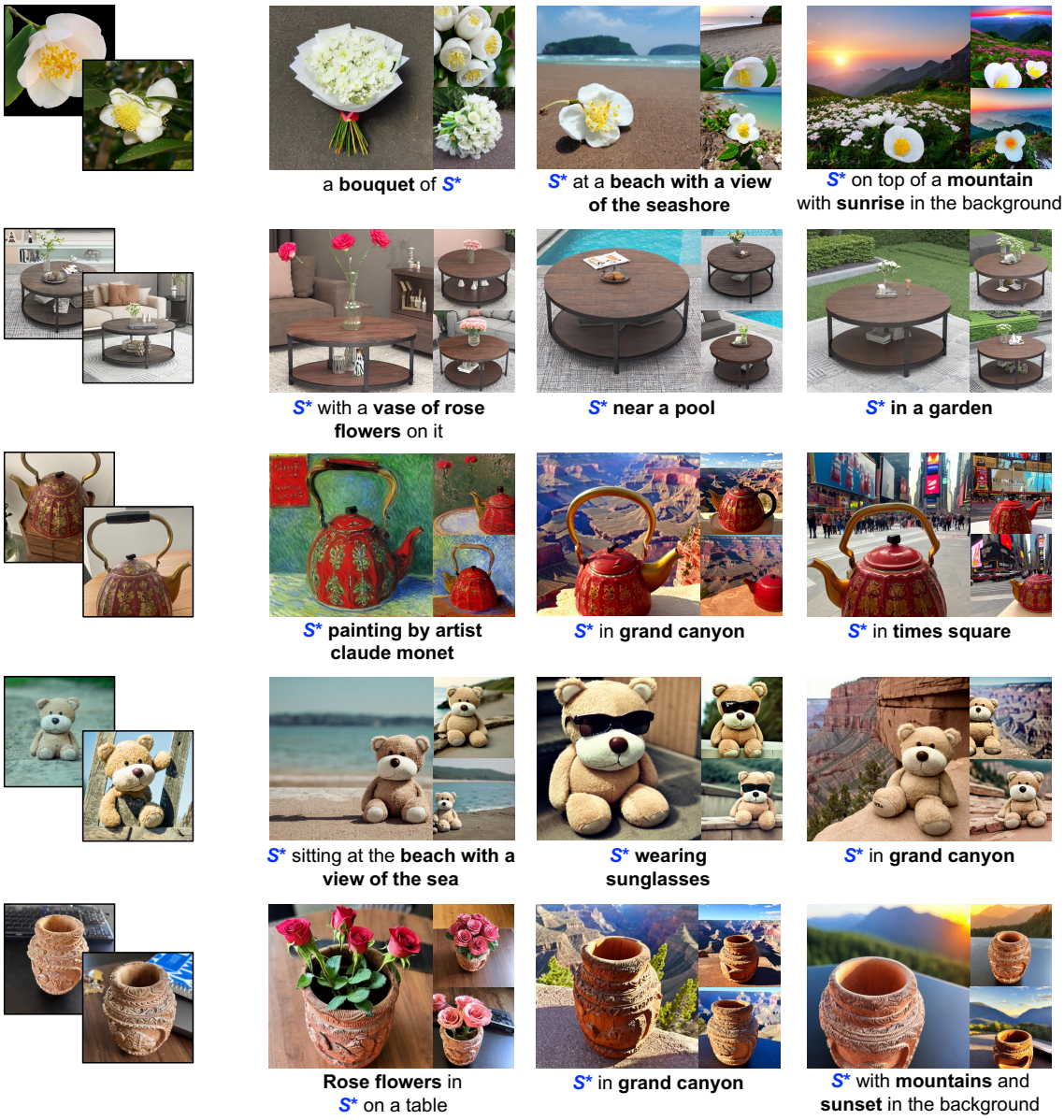


Figure 9: Generated images using object references. Our method can be used for a variety of personalization purposes, such as background replacement, artistic rendering, and attribute editing.