Diffusion Self-Distillation for Zero-Shot Customized Image Generation

Supplementary Material

A. Data Pipeline Prompts

In this section, we list out the detailed prompts used in our data generation (Sec. A.1), curation (Sec. A.2) and caption (Sec. A.3) pipelines.

A.1. Data Generation Prompts

To generate grid prompts, we employ GPT-4o as our language model (LLM) engine We instruct the LLM to focus on specific aspects during the grid generation process: preserving the identity of the subject, providing detailed content within each grid quadrant, and maintaining appropriate text length. However, we observed that not all sampled reference captions inherently include a clear instance suitable for identity preservation. To address this issue, we introduce an initial filtering stage to ensure that each sampled reference caption contains an identity-preserving target. This filtering enhances the quality and consistency of the generated grids.

Grid Generation Prompts

User Prompt:

Please be very creative and generate a prompt for text-to-image generation using flux, the prompt should create an evenly seperated grid of four. The four quadrants depict an identical item/asset/character under different environments/camera views/lighting conditions, etc (please be very very creative here). Every prompt should specify what the top-left, top-right, bottom-left, bottom-right quadrant depicts. Extract the asset from the following caption: `<sampled_reference_caption>`

System Prompt:

Response only the required prompt. Keep the fomat as one line and be as short and precise as possible, do not exceed 77 tokens. Be very creative! It could be a four-panel comic strip, a four-panel manga, real images, etc. The prompt should start with 'a grid of ...'

A.2. Data Curation Prompts

For data curation, we employ Gemini-1.5. To guide the vision-language model (VLM) in focusing on identity preservation, we utilize Chain-of-Thought (CoT) prompting [42]. Specifically, we first instruct the VLM to identify the common object or character present in both images. Next, we prompt it to describe each one in detail. Finally, we ask the VLM to analyze whether they are identical and to provide a conclusive response. We find that this CoT prompting significantly enhances the model's ability to concentrate on the identity and intricate details of the target object or character.

Data Curation Prompts

CoT Step 1:

Can you identity a common character/asset/item in the two images?

CoT Step 2:

Could you describe to me what the character/asset/item looks like in detail in the two images?

CoT Step 3:

Do the two images depict identical character/asset/item presented under different poses/lighting conditions/camera views/environment/etc.? Please consider this in terms of character/asset/item identity and be extremely critical. Could you describe to me what the common character/asset/item looks like in detail if it is indeed the same? End the response with a single 'yes' or 'no'.

A.3. Image Caption Prompts

We provide two methods for prompting our model: using the description of the expected output (Target Description) or InstructPix2Pix [3]-type instructions (Instruction).

Image Caption Prompts

Target Description:

Please provide a prompt for the image for Diffusion Model text-to-image generative model training, i.e. for FLUX or StableDiffusion 3. The prompt should be a detailed description of the image, including the character/asset/item, the environment, the pose, the lighting, the camera view, etc. The prompt should be detailed enough to generate the image. The prompt should be as short and precise as possible, in one-line format, and do not exceed 77 tokens.

Instruction:

Please provide a caption/prompt for the purpose of image-to-image editing, so that the prompt edits the first image into the second image. Do not include terms such as 'transform', 'image', etc.

B. GPT Evaluation Prompts

We closely follow DreamBench++ [27] in terms of our GPT evaluation. In Fig. 7, we demonstrate the prompts we use for evaluation, including our "de-biased" evaluation that penalizes "copy-pasting" effect.

C. Additional Results

C.1. Additional Qualitative Comparisons

In Fig. 9, we demonstrate more of the qualitative evaluation cases from the DreamBench++ [27] benchmark.

C.2. Additional Qualitative Results

Due to space constraints in the main paper, we presented shortened prompts. Here, we provide additional qualitative results in Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14 and

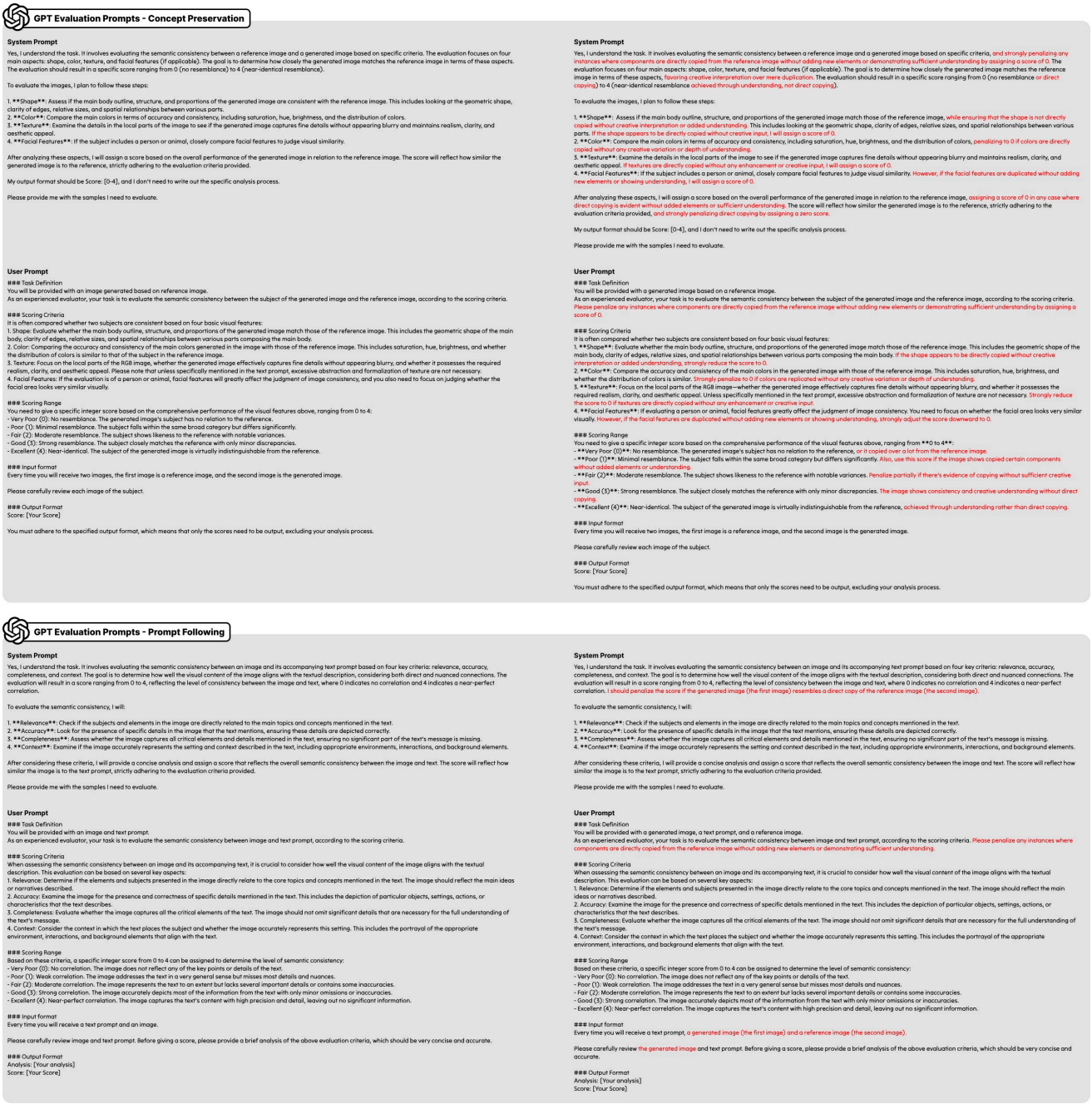


Figure 7. **GPT evaluation prompts** used across our evaluation, where the left shows the vanilla prompts from DreamBench++ [27] and the right shows our modified "de-biased" prompts, which strongly penalizes "copy-pasting" effects without sufficient creative inputs. We highlight our modified sentences in red.

Fig. 15, including the full prompts used for their generation. These detailed captions capture various aspects of the images and offer deeper insights into how our model operates.

C.3. Story Telling

Our model exhibits the capability to generate simple comics and manga narratives, as demonstrated in Fig. 16 and Fig. 17, where the conditioning image acts as the first panel. To create these storytelling sequences, we input the initial panel into GPT-4o, which generates a series of prompts centered around the main character from the input image. These prompts are crafted to form a coherent story spanning 8–10 panels, with each prompt being contextually meaningful on its own. Utilizing these prompts alongside the conditioning image, we generate the subsequent panels and finally align them to reconstruct a cohesive narrative.

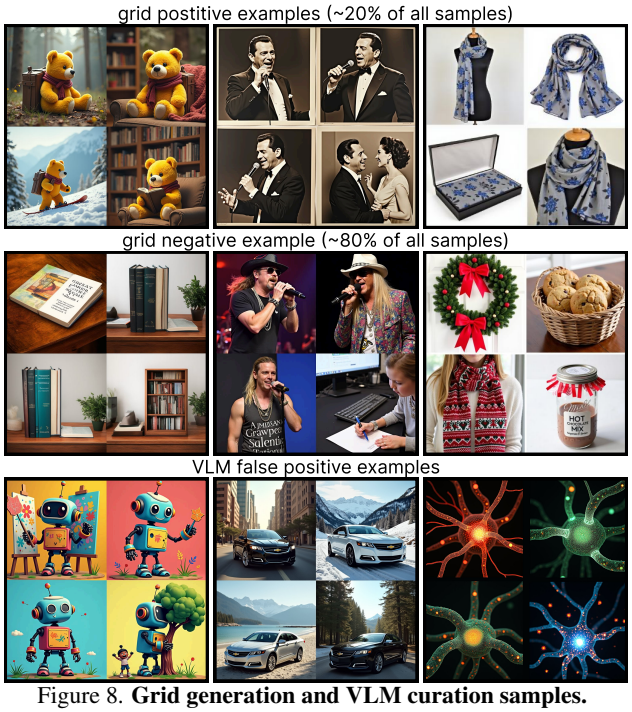


Figure 8. **Grid generation and VLM curation samples.**

Story Telling Prompts

Step 1: Identify Main Character

Please provide a prompt for the image for Diffusion Model text-to-image generative model training, i.e. for FLUX or StableDiffusion 3. The prompt should be a detailed description of the image, including the character/asset/item, the environment, the pose, the lighting, the camera view, etc. The prompt should be detailed enough to generate the image. The prompt should be as short and precise as possible, in one-line format, and does not exceed 77 tokens.

Step 2: Coherent Story Generation

Can you generate a series of prompts using the main character? The series of prompts should form a coherent story of 8-10 panels.

Step 3: Prompts Generation

Can you transfer the prompts so that each of them is individually sound?

D. More Detail on Architecture

We treat FLUX as a two-frame video generator. The left and right halves of the input correspond to "conditional frame" (the reference image) and "output frame". Specifically, we add a zero-initialized embedding layer on top of the original embedding layer to process the clean conditional image, then sum its output with the original embedding before passing it into the first transformer block. The model predicts a 1024×512 output, where the left half learns an identity mapping (reconstructing the conditional image) and the right half produces the target. We follow the standard approach to fine-tune via LoRA on all attention K/Q/V projections.

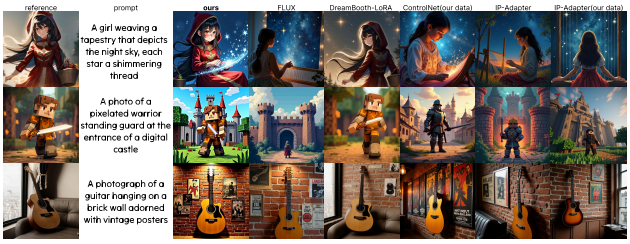
E. Grid and VLM Success Rate

From the original FLUX model, roughly 20% of the generated grids will be accepted by the VLM, where the VLM achieves 95% alignment with human curation. In Fig. 8, we show a few success and failure cases of the grid sampling and false-positive examples of the VLM.

F. FLUX Baseline Models

FLUX is a new DiT flow-matching model with guidance distillation, so UNet-based methods cannot be trivially adapted, as redesigning the training and architecture requires significant effort. Therefore, following DreamBench++, we use SD1.5 for DreamBooth, Textual Inversion and Blip-Diffusion, SDXL for DreamBooth-LoRA and IP-Adapters. We supply comparisons with the newest available FLUX-based IP-Adapter and DreamBooth-LoRA at the time of this paper's production, as well as vanilla FLUX with only text input. We also provide results on FLUX-based ControlNet and IP-Adapter trained on our data. The results below reinforce our method's superiority.

Method	Z-S?	CP↑	PF↑	CP·PF↑	DCP↑	DPF↑	DCP·DPF↑
FLUX	-	0.332	0.937	0.311	0.448	0.940	0.421
DreamBooth LoRA	✗	0.804	0.521	0.419	0.561	0.571	0.320
ControlNet (our data)	✓	0.355	0.869	0.308	0.491	0.848	0.416
IP-Adapter	✓	0.380	0.717	0.272	0.500	0.756	0.378
IP-Adapter (our data)	✓	0.414	0.712	0.295	0.512	0.723	0.370
Ours	✓	0.631	0.726	0.458	0.789	0.757	0.597



G. Discussion on Scalability

We acknowledge that the scalability of Diffusion Self-Distillation is not fully explored within the scope of this paper. However, we posit that Diffusion Self-Distillation is inherently scalable along three key dimensions. First, Diffusion Self-Distillation can scale with advancements in the teacher model's grid generation capabilities and its in-context understanding of identity preservation. Second, the scalability extends to the range of tasks we leverage; while this paper focuses on general adaptation tasks, a broader spectrum of applications remains open for exploration. Third, Diffusion Self-Distillation scales with the extent to which we harness foundation models. Increased diversity and more meticulously curated data contribute to improved generalization of our model. As foundation models—including base text-to-image generation models, language models (LLMs), and vision-language models (VLMs)—continue to evolve, Diffusion Self-Distillation naturally benefits from these advancements without necessitating any modifications to the existing workflow. A direct next step involves scaling the method to incorporate a significantly larger dataset and integrating forthcoming, more advanced foundation models.

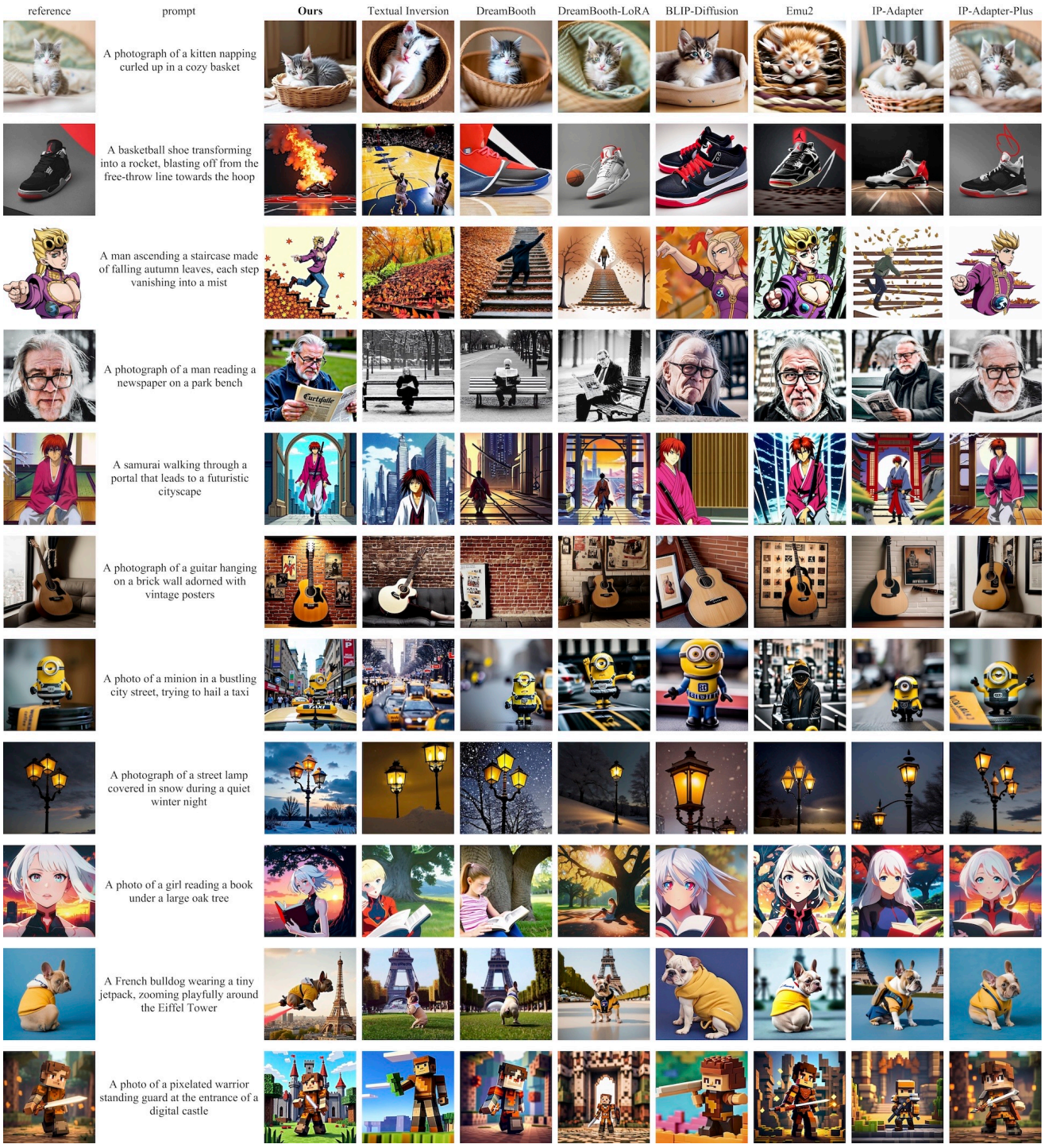


Figure 9. **Additional qualitative comparison.**

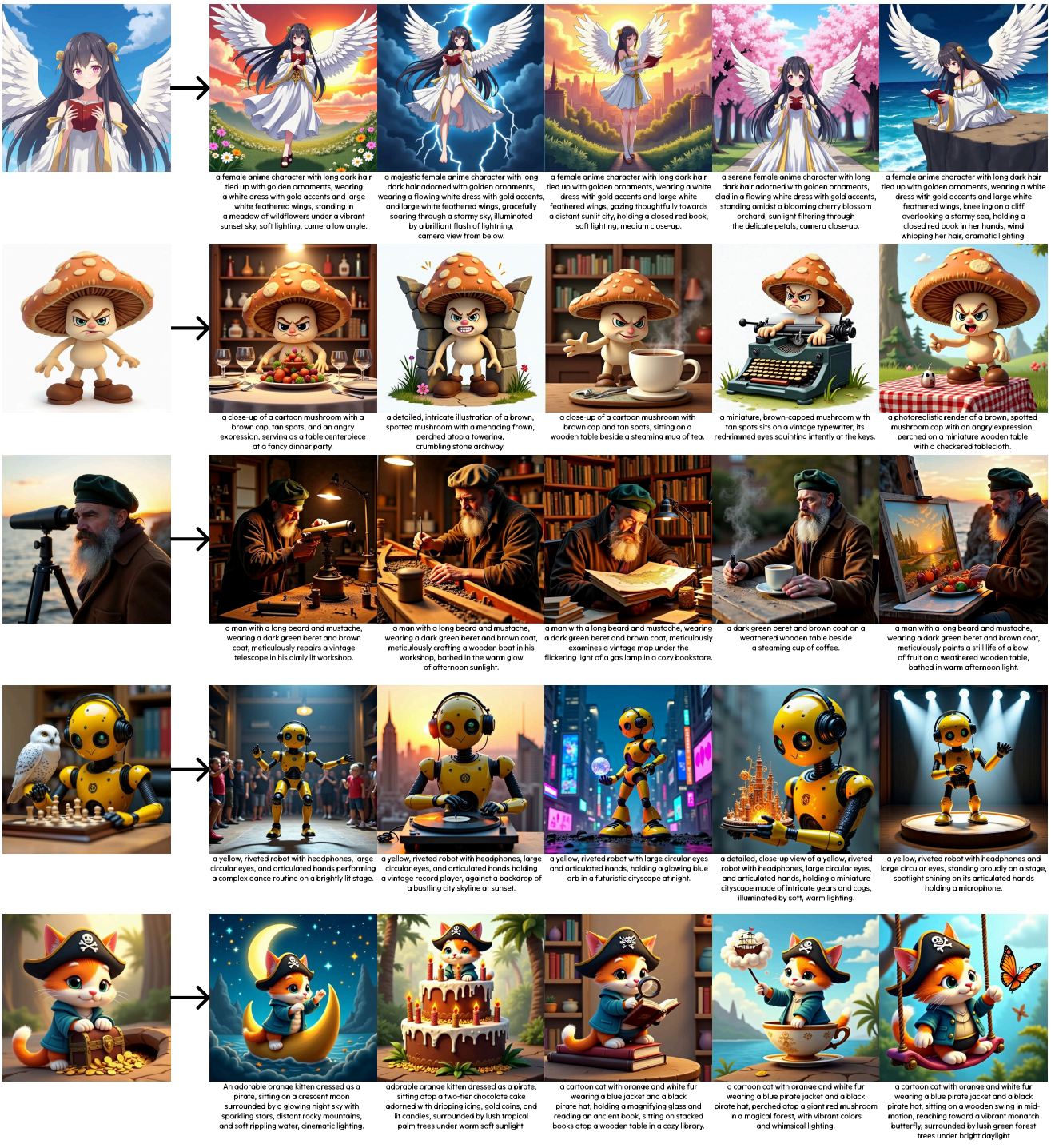


Figure 10. **Additional character identity preserving results.**

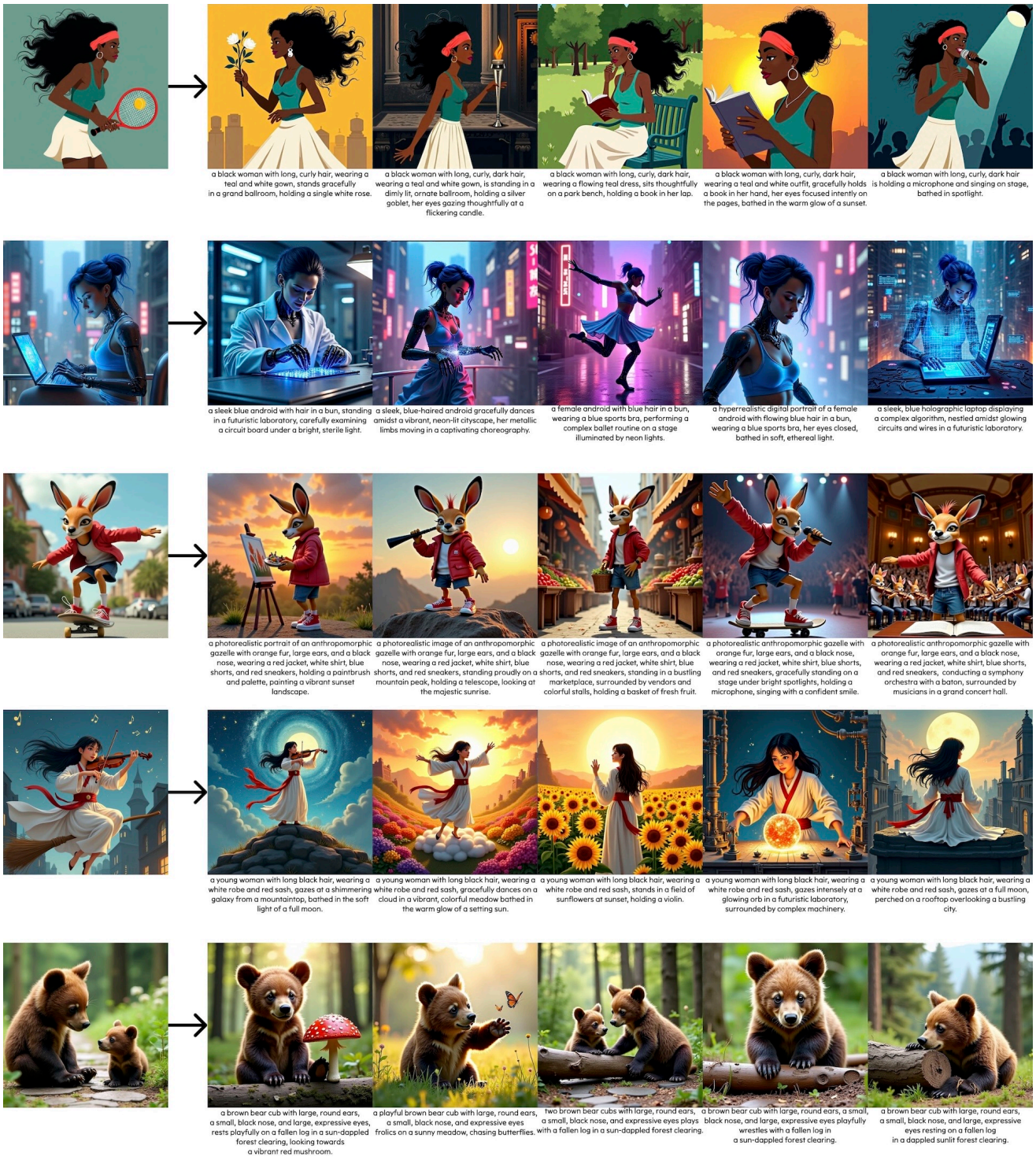


Figure 11. **Additional character identity preserving results.**

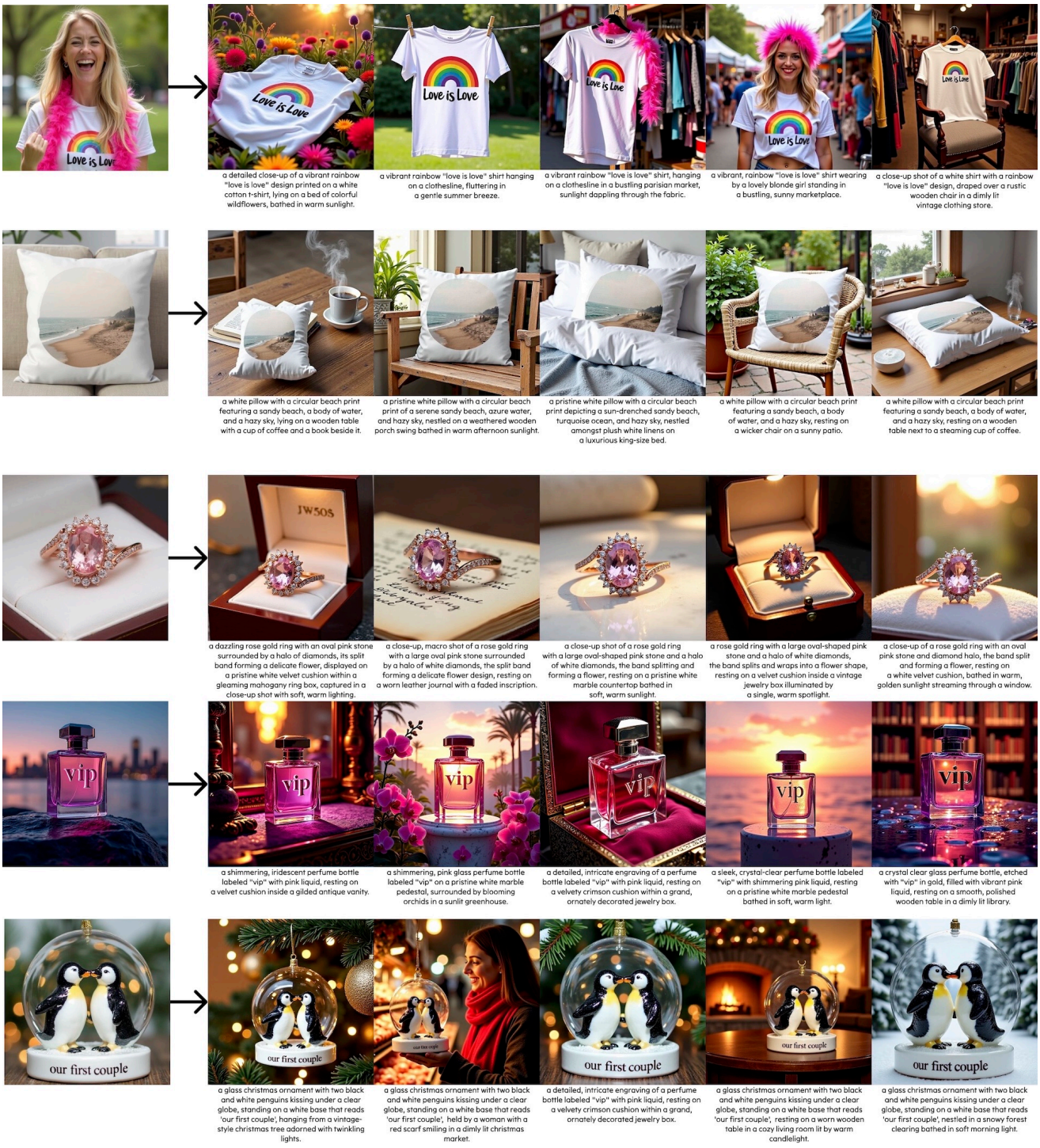


Figure 12. **Additional object/item identity preserving results.**

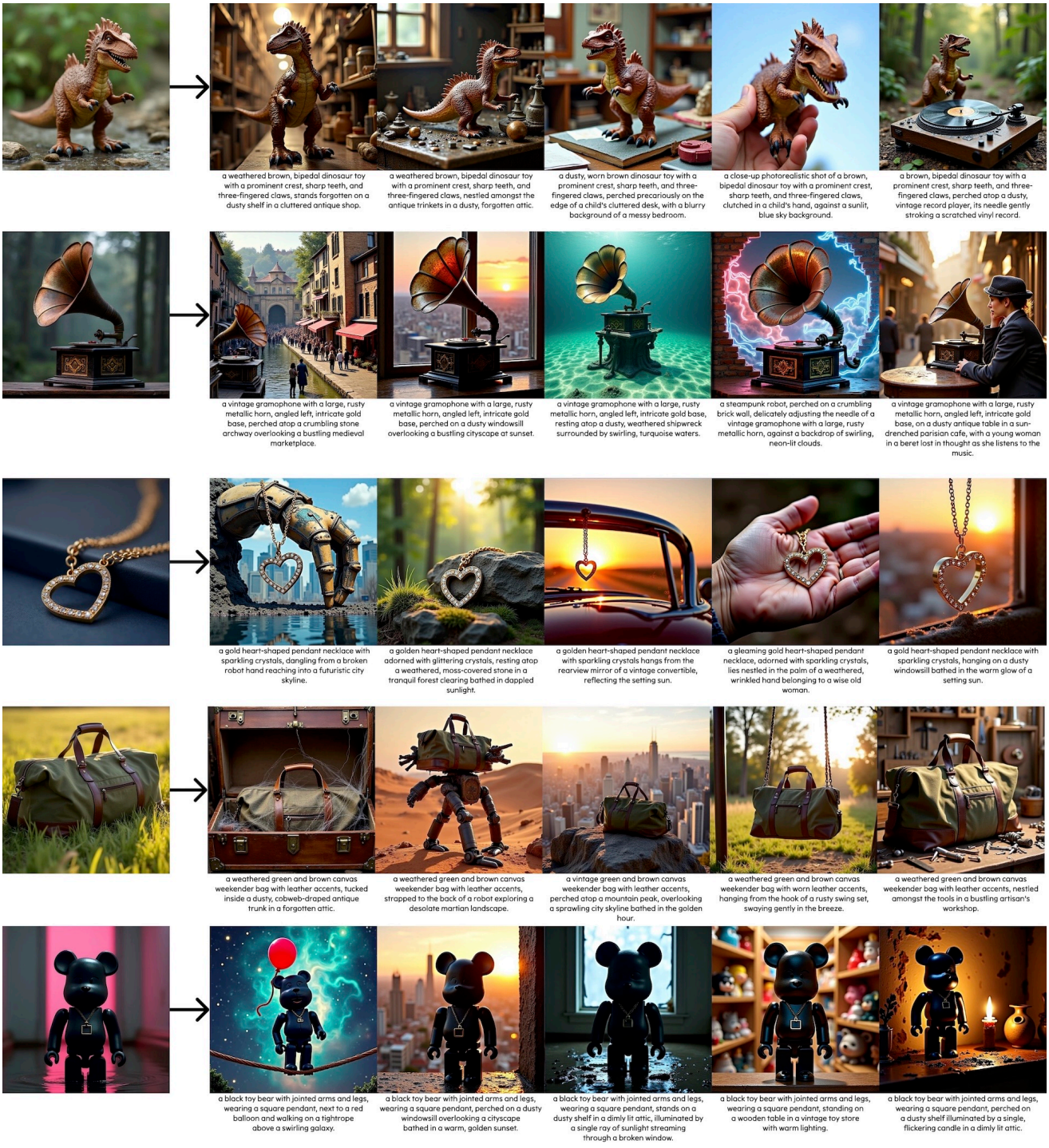


Figure 13. **Additional object/item identity preserving results.**

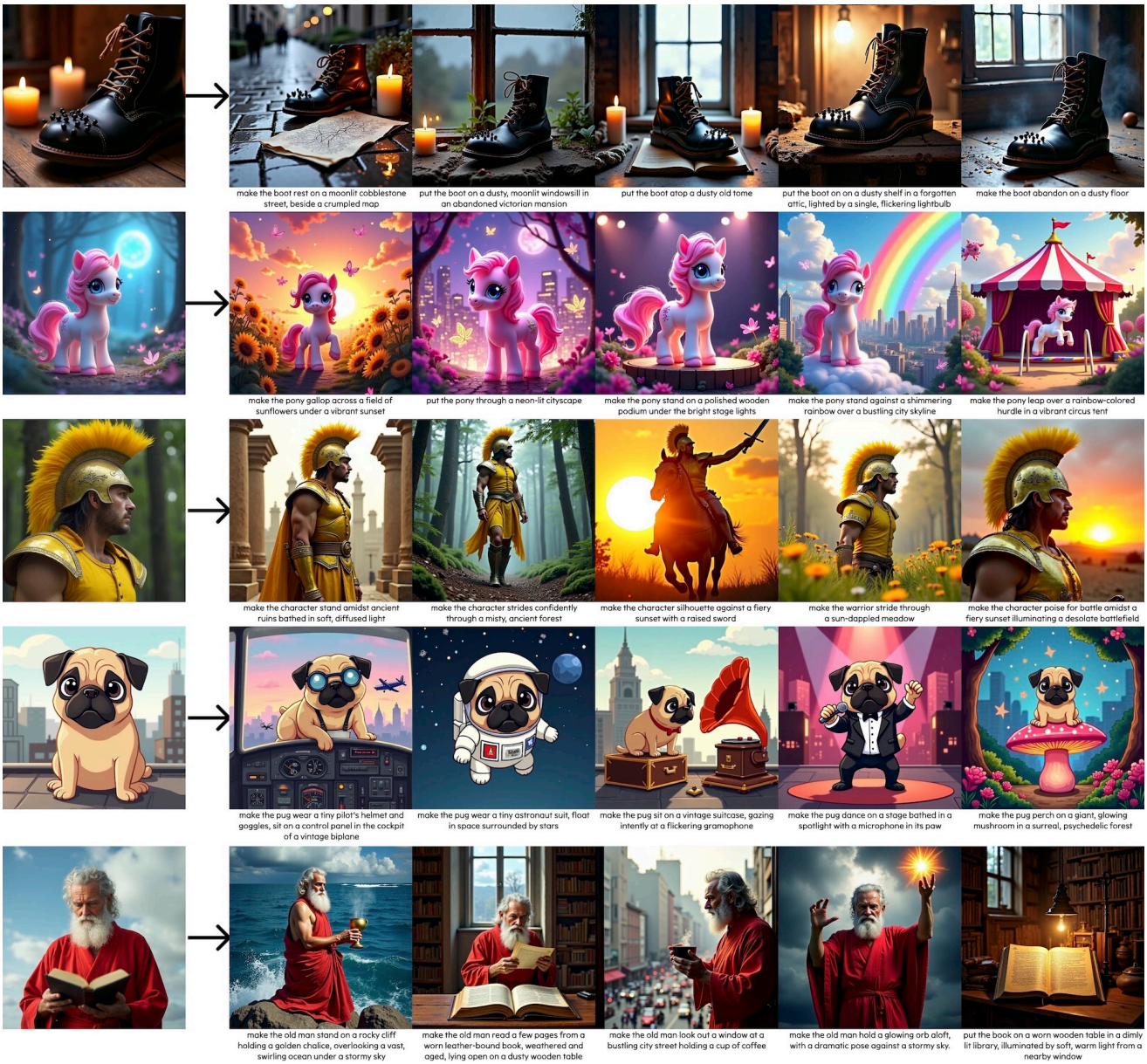


Figure 14. **Additional instruction prompting results.**

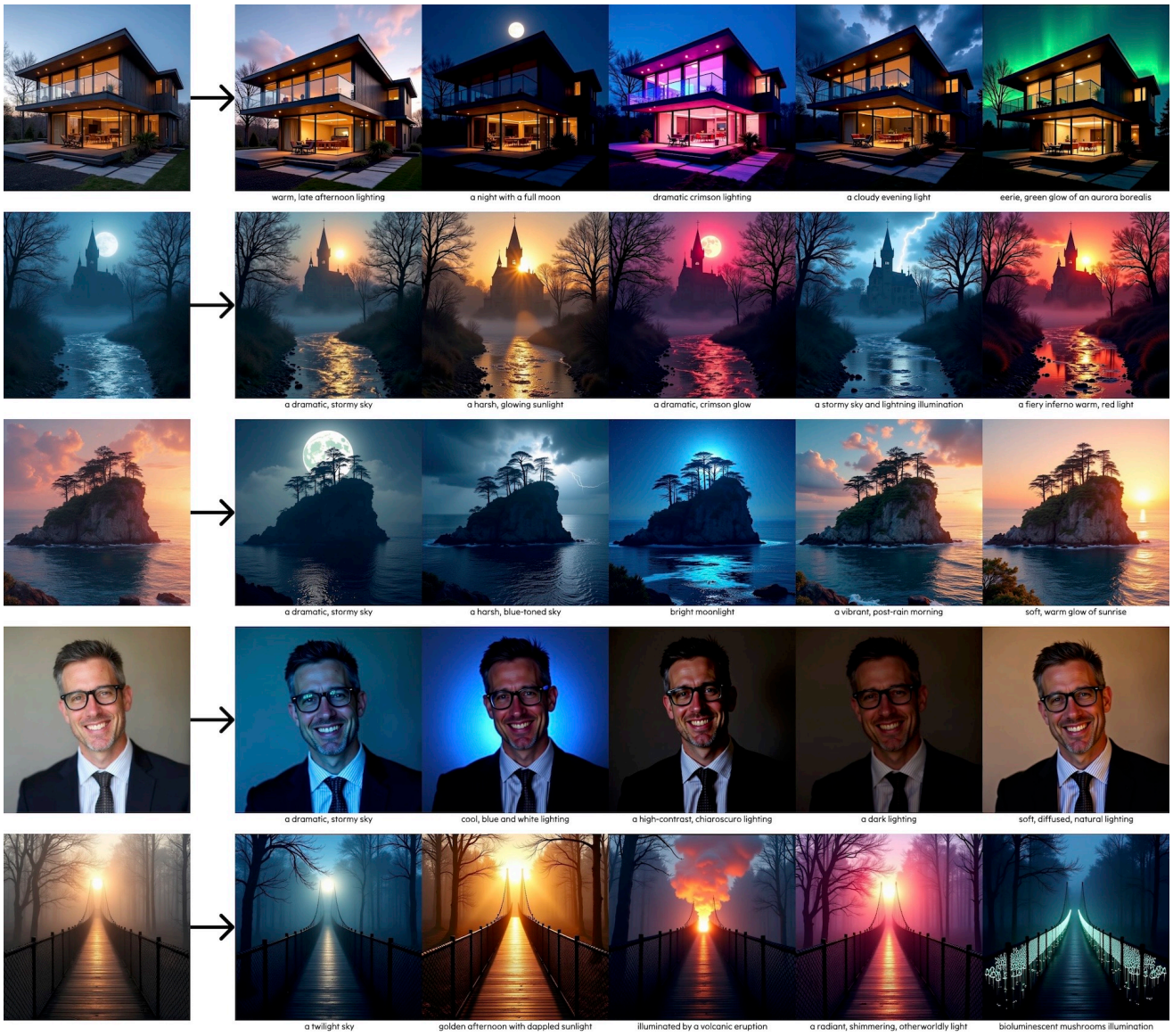
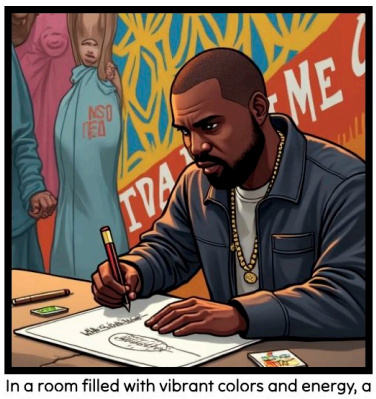


Figure 15. **Additional relighting results.**

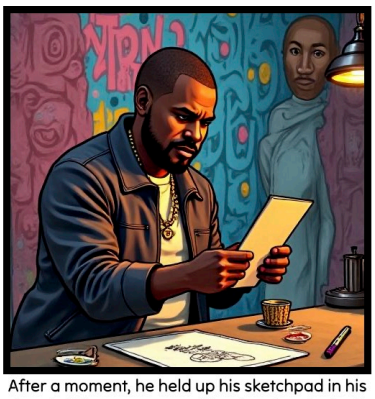
Figure 16. **Comic generation example 1.** The conditioned image is the first panel.

1



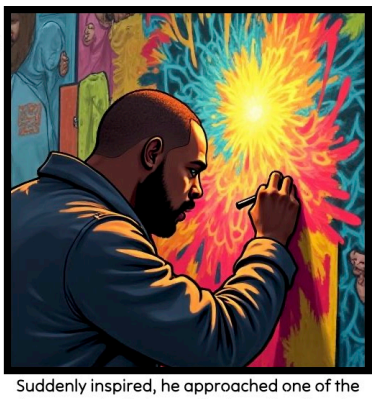
In a room filled with vibrant colors and energy, a focused man with a shaved head and a gold chain sat at a table. He was deeply engrossed in drawing on a sheet of paper, his pencil moving with purpose, while warm light spilled over graffiti-like murals painted on the walls around him. His expression was determined, as if every line he sketched carried deep meaning.

2



After a moment, he held up his sketchpad in his hands. His eyes scanned the drawings he had created, and a look of resolve crossed his face. The colorful murals behind him seemed to mirror the intensity in his gaze, the warm lighting accentuating the passion that had sparked within him.

3



Suddenly inspired, he approached one of the large murals. He began to draw directly onto the wall. His movements were precise and intentional, as colors and patterns flowed from his imagination to the surface. The warm light bathed his intense expression, as if illuminating the raw energy of his creativity.

4

When he was finished, he stepped back to admire his work, arms crossed over his chest. The mural was alive with vivid, swirling graffiti, and his face lit up with pride. The warm light glowed over the artwork, and for a moment, he stood there, content, knowing he had given life to his vision.

5

But he was not done yet. He returned to the table, sketchpad in hand, and began drawing again. His pencil moved even faster now, capturing new ideas that poured into his mind. The room was a swirl of vibrant murals and soft, warm shadows, the energy of creation pulsing through the space.

6

Suddenly, he turned his head slightly, as a noise from outside broke his concentration. He set down his pencil, his expression one of curiosity and intrigue. The warm light reflected off his dark jacket, and the graffiti walls behind him seemed to whisper with a story yet to be discovered.

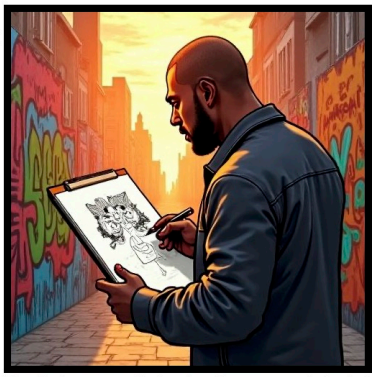
7

He walked to the doorway, peering out into the distance. The warm light of the room spilled out into the world beyond, casting long shadows on the floor. Something had drawn his attention, and he knew he had to explore it. The spark of adventure lit his eyes, and he stepped forward.

8

Outside, the man found himself bathed in the golden light of the setting sun. He walked with purpose, the colors of the urban world around him just as vibrant as the murals he had created. He felt a sense of unity with the graffiti-covered walls that stretched along the city streets.

9



Finally, he paused at a street corner and started sketching again. The warm sunset light enveloped him, and he realized that his art had become a part of something larger—a story woven into the very fabric of the city. His journey of creativity had led him here, and he knew there were still many more stories to tell.

Figure 17. **Comic generation example 2.** The conditioned image is the first panel.