Diffusion Self-Distillation for Zero-Shot Customized Image Generation

Shengqu Cai Eric Ryan Chan Yunzhi Zhang
Leonidas Guibas Jiajun Wu Gordon Wetzstein
Stanford University

Figure 1. Given an input image, Diffusion Self-Distillation is a novel diffusion-based approach that generates diverse images that maintain the input's identity across various contexts. Unlike prior approaches that require fine-tuning or are limited to specific domains, Diffusion Self-Distillation offers instant customization without any additional inference-stage training, enabling precise control and editability in text-to-image diffusion models. This ability makes Diffusion Self-Distillation a valuable tool for general AI content creation.

Abstract

Text-to-image diffusion models produce impressive results but are frustrating tools for artists who desire fine-grained control. For example, a common use case is to create images of a specific concept in novel contexts, i.e., "identity-preserving generation". This setting, along with many other tasks (e.g., relighting), is a natural fit for image+text-conditional generative models. However, there is insufficient high-quality paired data to train such a model directly. We propose Diffusion Self-Distillation, a method for using a pre-trained text-to-image model to generate its own dataset for text-conditioned image-to-image tasks. We first leverage a text-to-image diffusion model's in-context generation ability to create grids of images and curate a large paired dataset with the help of a vision-language model. We then fine-tune the text-to-image model into a text+image-to-image model using the curated paired dataset. We demonstrate that Diffusion Self-Distillation outperforms existing zero-shot methods and is competitive with per-instance tuning techniques on a wide range of identity-preserving generation tasks, without requiring test-time optimization. Project page: primecai.github.io/dsd.

1. Introduction

In recent years, text-to-image diffusion models [24, 28, 29, 32] have set new standards in image synthesis, generating high-quality and diverse images from textual prompts. However, while their ability to generate images from text is impressive, these models often fall short in offering precise control, editability, and consistency—key features that are crucial for real-world applications. Text input alone can be insufficient to convey specific details, leading to variations that may not fully align with the user's intent, especially in scenarios that require faithful adaptation of a character or asset's identity across different contexts.

Maintaining the instance's identity is challenging, however. We distinguish *structure-preserving* edits, in which the target and source image share the general layout, but may differ in style, texture, or other local features, and *identity-preserving* edits, where assets are recognizably the same across target and source images despite potentially large-scale changes in image structure (Fig. 3). The latter task is a superset of the former and requires the model to have a significantly more profound understanding of the input image and concepts to extract and customize the desired

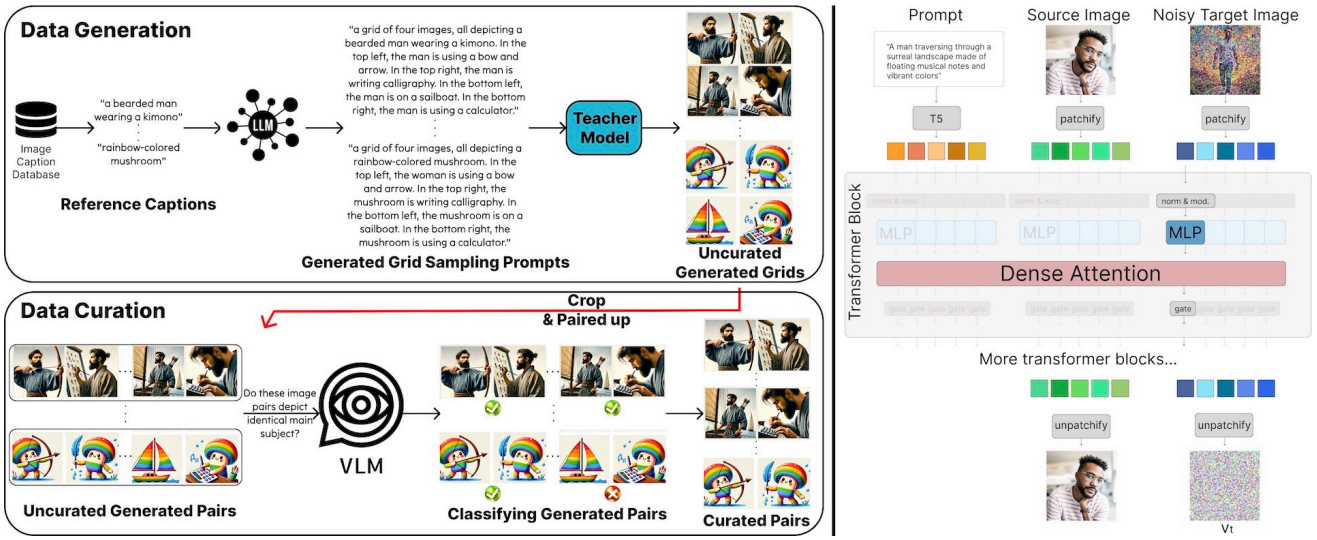


Figure 2. **Overview of our pipeline.** *Left:* the top shows our vanilla paired data generation wheel (Sec. 3.1). We first sample reference image captions from the LAION [33] dataset. These reference captions are parsed through an LLM to be translated into identity-preserved grid generation prompts (Sec. 3.1.2). We feed these enhanced prompts to a pretrained text-to-image diffusion model to sample potentially identity-preserved grids of images, which are then cropped and composed into vanilla image pairs (Sec. 3.1.1). On the bottom, we show our data curation pipeline (Sec. 3.1.3), where the vanilla image paired are fed into a VLM to classify whether they depict identical main subjects. This process mimics a human annotation/curation process while being fully automatic; we use the curated data as our final training data. *Right:* we extend the diffusion transformer model into an image-conditioned framework by treating the input image as the first frame of a two-frame sequence. The model generates both frames simultaneously—the first reconstructs the input, while the second is the edited output—allowing effective information exchange between the conditioning image and the desired output.

identity. For example, image editing [2, 22, 43], such as local content editing, re-lighting, and semantic image synthesis, etc. are all *structure-preserving* and *identity-preserving* edits, but novel-view synthesis and character-consistent generation under pose variations, are *identity-preserving* but not *structure-preserving*. We aim to address the general case, maintaining identity without constraining structure.

For *structure-preserving* edits, adding layers, as in ControlNet [43], introduces spatial conditioning controls but is limited to structure guidance and does not address consistent identity adaptation across diverse contexts. For *identity-preserving* edits, fine-tuning methods such as DreamBooth [31] and LoRA [13] can improve consistency using a few reference samples but are time consuming and computationally intensive, requiring training for each reference. Zero-shot alternatives like IP-Adapter [42] and InstantID [37] offer faster solutions without the need for retraining but fall short in providing the desired level of consistency and customization; IP-Adapter [42] lacks full customization capabilities, and InstantID [37] is restricted to facial identity.

In this paper, we propose a novel approach called Diffusion Self-Distillation, designed to address the core challenge of zero-shot instant customization and adaptation of any character or asset in text-to-image diffusion models. We identify the primary obstacle that hinders prior methods, such as IP-Adapter [42] and InstantID [37], from achieving better identity preservation or generalizing beyond facial contexts: the absence of large-scale paired datasets and corresponding supervised identity-preserving training pipelines. With recent advancements in foundational model capabilities, we are now positioned to exploit these strengths further. Specifically, we can generate consistent grids of identical characters or assets, opening a new pathway for customization that eliminates the need for pre-existing, handcrafted paired datasets—which are expensive and time consuming to collect. The ability to generate these consistent grids likely emerged from foundational model training on diverse datasets, including photo albums, mangas, and comics. Our approach harnesses Vision-Language Models (VLMs) to automatically curate many generated grids, producing a diverse set of grid images with consistent identity features across various contexts. This curated synthetic dataset then serves as the foundation for fine-tuning and adapting any identity, transforming the task of zero-shot customized image generation from unsupervised to supervised. Diffusion Self-Distillation offers transformative potential for applications like consistent character generation, camera control, relighting, and asset customization in fields such as comics and digital art. This flexibility allows artists to rapidly iterate and adapt their work, reducing effort and enhancing creative freedom, making Diffusion Self-Distillation a valuable tool for AI-generated content.

We summarize our contributions as follows:

- We propose Diffusion Self-Distillation, a *zero-shot*

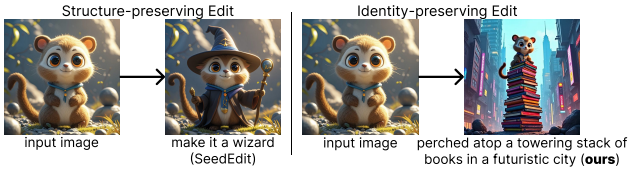


Figure 3. **Difference between *structure-preserving* and *identity-preserving* edits.** In *structure-preserving* editing, the main structures of the image are preserved, and only local edits or stylizations are performed. In *identity-preserving* editing, the global structure of the image may change radically.

identity-preserving customized image generation model that scales to any instance under any context, with performances on par with inference-stage tuning methods;

- We provide a self-distillation pipeline to obtain identity-preserving data pairs purely from pretrained text-to-image diffusion models, LLMs, and VLMs, without any human effort involved in the entire data creation wheel;
- We correspondingly design a unified architecture for image-to-image translation tasks involving *both* identity- and structure-preserving edits, including personalization, relighting, depth controls, and instruction following.

2. Related work

Recent advancements in diffusion models have underscored the need for enhanced control and customization in image-generation tasks. Various methods have been proposed to address these challenges through additional conditioning mechanisms, personalization, and rapid adaptation [26].

Control Mechanisms in Diffusion Models. To move beyond purely text-based controls, approaches like ControlNet [43] introduce spatial conditioning via inputs such as sketches, depth maps, and segmentation masks, enabling fine-grained structure control. ControlNet++ [19] refines this by enhancing the integration of spatial inputs for more nuanced control. Uni-ControlNet [44] unifies various control types within a single framework, standardizing the handling of diverse signals. T2I-Adapter [23] employs lightweight adapters to align pretrained models with external control signals without altering the core architecture. While these methods offer increased flexibility, they often focus on structural conditioning types such as depths and lack capabilities for concept extraction or identity preservation.

Personalization and Fine-Tuning. Techniques like DreamBooth [31] and LoRA [13] enhance the consistency and relevance of generated images by fine-tuning models with small sets of reference images. DreamBooth [31] personalizes models to maintain a subject's identity across different contexts, while LoRA [13] provides an efficient approach to fine-tuning large models without extensive retraining. However, these methods require multiple images and test-time optimization for each reference, which can be computationally expensive—especially with the exponential growth in model sizes (12 billion parameters for FLUX).

Zero-Shot and Fast Adaptation. IP-Adapter [42] incorporates image prompts into diffusion models using image embeddings, allowing for generations that align closely with reference visuals. InstantID [37] ensures zero-shot face preservation, maintaining a subject's key features across various contexts. While effective as zero-shot methods without user training, IP-Adapter [42] struggles to adapt specific targets like unique characters or assets, and InstantID [37] is limited to facial identity preservation. IPAdapter-Instruct [30] enhances image-based conditioning with instruct prompts but relies heavily on specific instructions and task-specific pretrained models. Other methods, such as SuTI [7] and GDT [15], handcrafted corresponding datasets, which are expensive and challenging to collect and scale. Another work along this line is Subject-Diffusion [21], which uses segmentation masks to create synthetic data for training but is bounded by achieving only simple attributes and accessories copying and editing within input images.

Existing methods contribute valuable advancements but often target specific domains or require user-stage tuning. Diffusion Self-Distillation bridges these gaps by offering a unified, zero-shot approach for consistent customization of characters and assets using minimal input. By leveraging self-distillation assisted by vision-language models, Diffusion Self-Distillation provides a comprehensive and adaptable solution for a wide range of creative applications.

3. Diffusion Self-Distillation

We discover that recent text-to-image generation models offer the surprising ability to generate in-context, consistent image grids (see Fig. 2, left). Motivated by this insight, we develop a zero-shot adaptation network that offers fast, diverse, high-quality, and identity-preserving, i.e., consistent image generation conditioned on a reference image. For this purpose, we first generate and curate sets of images that exhibit the desired consistency using pretrained text-to-image diffusion models, large language models (LLMs), and vision-language models (VLMs) (Sec. 3.1). Then, we finetune the same pretrained diffusion model with these consistent image sets, employing our newly proposed parallel processing architecture (Sec. 3.2) to create a conditional model. By this end, Diffusion Self-Distillation finetunes a pretrained text-to-image diffusion model into a zero-shot customized image generator in a supervised manner.

3.1. Generating a Pairwise Dataset

To create a pairwise dataset for supervised Diffusion Self-Distillation training, we leverage the emerging multi-image generation capabilities of pretrained text-to-image diffusion models to produce potentially consistent

vanilla images (Sec. 3.1.1) created by LLM-generated prompts (Sec. 3.1.2). We then use VLMs to curate these vanilla samples, obtaining clean sets of images that share the desired identity consistency (Sec. 3.1.3). The data generation and curation pipeline is shown in Fig. 2, left.

3.1.1. Vanilla Data Generation via Teacher Model

To generate sets of images that fulfill the desired identity preservation, we prompt the teacher pretrained text-to-image diffusion model to create images containing multiple panels featuring the same subject with variations in expression, pose, lighting conditions, and more, for training purposes. Such prompting can be as simple as specifying the desired identity preservation in the output, such as "*a grid of 4 images representing the same* $< object/character/scene/etc. >$", "*an evenly separated 4 panels, depicting identical* $< object/character/scene/etc. >$", etc. We additionally specify the expected content in each sub-image/panel. The full set of prompts is provided in our supplemental material Sec. A. Our analysis shows that current state-of-the-art text-to-image diffusion models (e.g., SD3 [8], DALL·E 3, FLUX) demonstrate this identity-preserving capability, likely emerging from their training data, which includes comics, mangas, photo albums, and video frames. Such in-context generation ability is crucial to our data generation wheel.

3.1.2. Prompt Generation via LLMs

We rely on an LLM to "brainstorm" a large dataset of diverse prompts, from which we derive our image grid dataset. By defining a prompt structure, we prompt the LLM to produce text prompts that describe image grids. A challenge we encountered is that when prompted to create large sets of prompts, LLMs tend to produce prompts of low diversity. For example, we noticed that without additional guidance, GPT-4o has a strong preference for prompts with cars and robots, resulting in highly repetitive outputs. To address this issue, we utilize the available image captions in the LAION [33] dataset, feeding them into the LLM as content references. These references from real image captions dramatically improve the diversity of generated prompts. Optionally, we also use the LLM to filter these reference captions, ensuring they contain a clear target for identity preservation. We find that this significantly improves the hit rate of generating consistent multi-image outputs.

3.1.3. Dataset Curation and Caption with VLMs

While the aforementioned data generation scheme provides identity-preserving multi-image samples of decent quality and quantity, these initial "uncurated" images tend to be noisy and unsuitable for direct use. Therefore, we leverage the strong capabilities of VLMs to curate a clean dataset. We extract pairs of images from the generated samples intended to preserve the identity and ask the VLM whether the two images depict the same object, character, scene, etc. We find that employing Chain-of-Thought prompting [38] is particularly helpful in this context. Specifically, we first prompt the VLM to identify the common object, character, or scene present in both images, then have it describe each one in detail, and finally analyze whether they are identical, providing a conclusive response. This process yields pairs of images that share the same identity.

3.2. Parallel Processing Architecture

We desire a conditional architecture suitable for general image-to-image tasks, including transformations in which structure is preserved, and transformations in which concepts/identities are preserved but image structure is not. This is a challenging problem because it may necessitate the transfer of fine details without guaranteeing spatial correspondences. While the ControlNet [43] architecture is excellent at structure-preserving edits, such as depth-to-image or segmentation-map-to-image, it struggles to preserve details under more complex identity-preserving edits, where the source and target images are not pixel-aligned. On the other hand, IP-Adapter [42] can extract certain concepts, such as styles, from the input image. Still, it strongly relies on a task-specific image encoder and often fails to preserve more complex concepts and identities. Drawing inspiration from the success of multi-view and video diffusion models [1, 3–5, 10–12, 14, 16, 18, 34, 35, 39, 41], we propose a simple yet effective method to extend the vanilla diffusion transformer model into an image-conditioned diffusion model. Specifically, we treat the input image as the first frame of a video and produce a two-frame video as output. The final loss is computed over the two-frame video, establishing an identity mapping for the first frame and a conditionally editing target for the second frame. Our architecture design allows generality for generic image-to-image translation tasks, since it enables effective information exchange between the two frames, allowing the model to capture complex semantics and perform sophisticated edits, as shown in Fig. 2, right.

4. Experiments

Implementation details. We use FLUX1.0 DEV as both our teacher and student models, achieving self-distillation. For prompt generation, we use GPT-4o; for dataset curation and captioning, we use Gemini-1.5. We train all models on 8 NVIDIA H100 80GB GPUs with an effective batch size of 160 for 100k iterations, using AdamW optimizer [20] with a learning rate 10^{-4}. Our parallel processing architecture uses LoRAs with rank 512 on the base model.

Datasets. Our final training dataset contains $\sim$ 400k subject-consistent image pairs generated from our teacher model, FLUX1.0 DEV. Generating and curating the dataset is

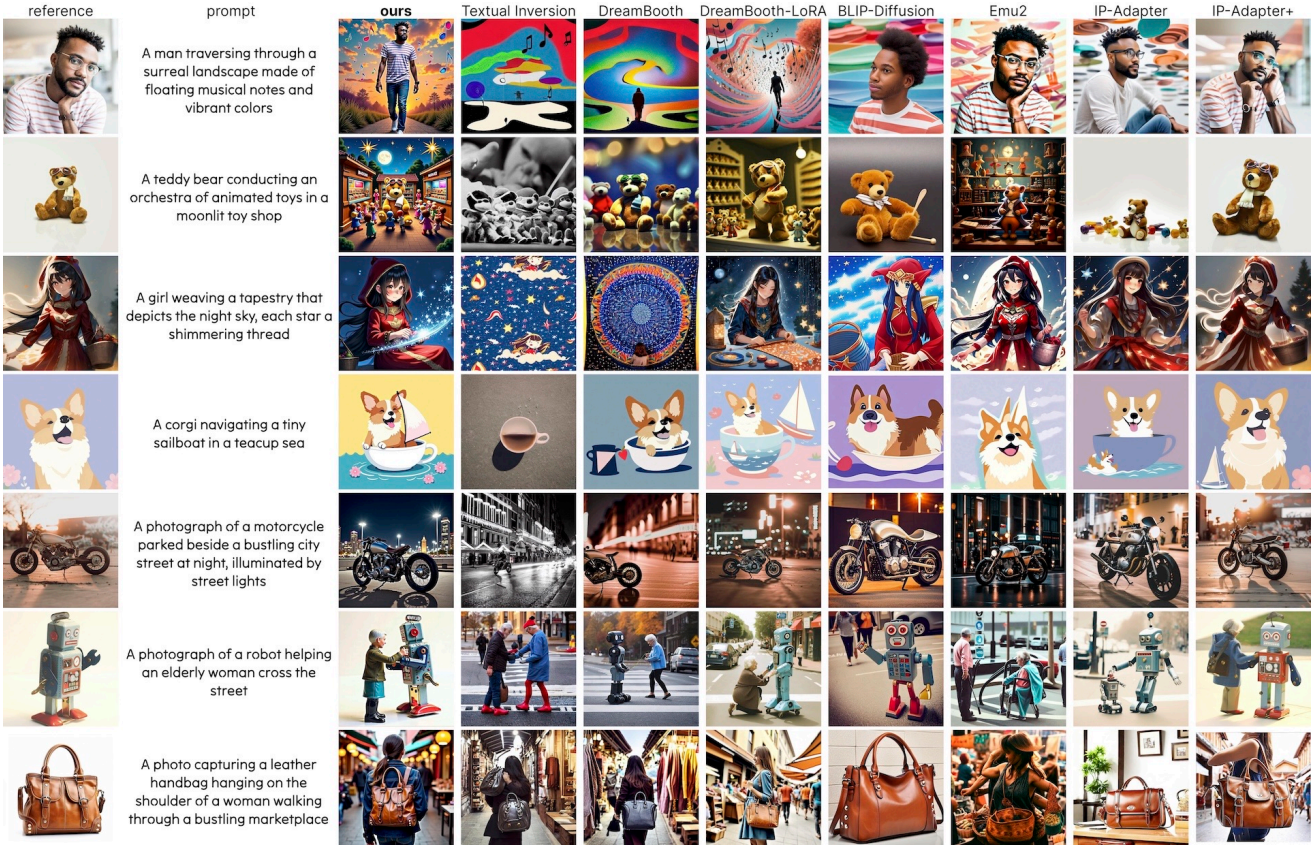


Figure 4. **Qualitative comparison.** Overall, our method achieves high subject identity preservation and prompt-aligned diversity while not suffering from a "copy-paste" effect, such as the results of IP-Adapter+ [42]. This is largely thanks to our supervised training pipeline, which alleviates the base model's in-context generation ability.

fully automated and requires no human effort, so its size could be further scaled. We use the publicly available DreamBench++ [25] dataset and follow their protocols for evaluation. DreamBench++ [25] is a comprehensive and diverse dataset for evaluating personalized image generation, consisting of 150 high-quality images and 1,350 prompts—significantly more than previous benchmarks like DreamBench [31]. The dataset covers various categories such as animals, humans, objects, etc., including photorealistic and non-photorealistic images, with prompts designed to span different difficulty levels (simple/imaginative). In contrast, prompts are generated using GPT-4o and refined by human annotators to ensure diversity and ethical compliance.

Baselines. We follow the setups in DreamBench++ [25] and compare our model with two classes of baselines: inference-stage tuning models and zero-shot models. For inference-stage models, we compare against Textual Inversion [9], DreamBooth [31] and its LoRA [13] version. For zero-shot models, we compare with BLIP-Diffusion [17], Emu2 [36], IP-Adapter [42], IP-Adapter+ [42].

Evaluation metrics. The evaluation protocol of prior works [7, 30, 31, 42] typically involves comparing the CLIP [27] and DINO [6] feature similarities. However, we note that the metrics mentioned above capture only global semantic similarity, are extremely noisy, and are biased towards "copy-pasting" the input image. This is especially troublesome when the input image or the prompt is complex. We refer to DreamBench++ [25] for a detailed analysis of their limitations. Therefore, we follow the metrics designed in DreamBench++ [25] and report GPT-4o scores on the more diverse DreamBench++ [25] benchmark for both concept preservation (CP) with different categories of subjects and prompt following (PF) with photorealistic (Real.) and Imaginative (Imag.) prompts, then use their product as a final evaluation score. This evaluation protocol emulates a human user study using VLMs. We additionally slightly modify the GPT evaluation prompts so that penalization can be applied if the generated contents show no internal understanding and creative output but instead naively copy over components from the reference image. The modified metrics are named "de-biased concept preservation (Debiased CP)" and "de-biased prompt following (Debiased PF)". The full set of GPT evaluation prompts will be provided in our supplementary Sec. B.

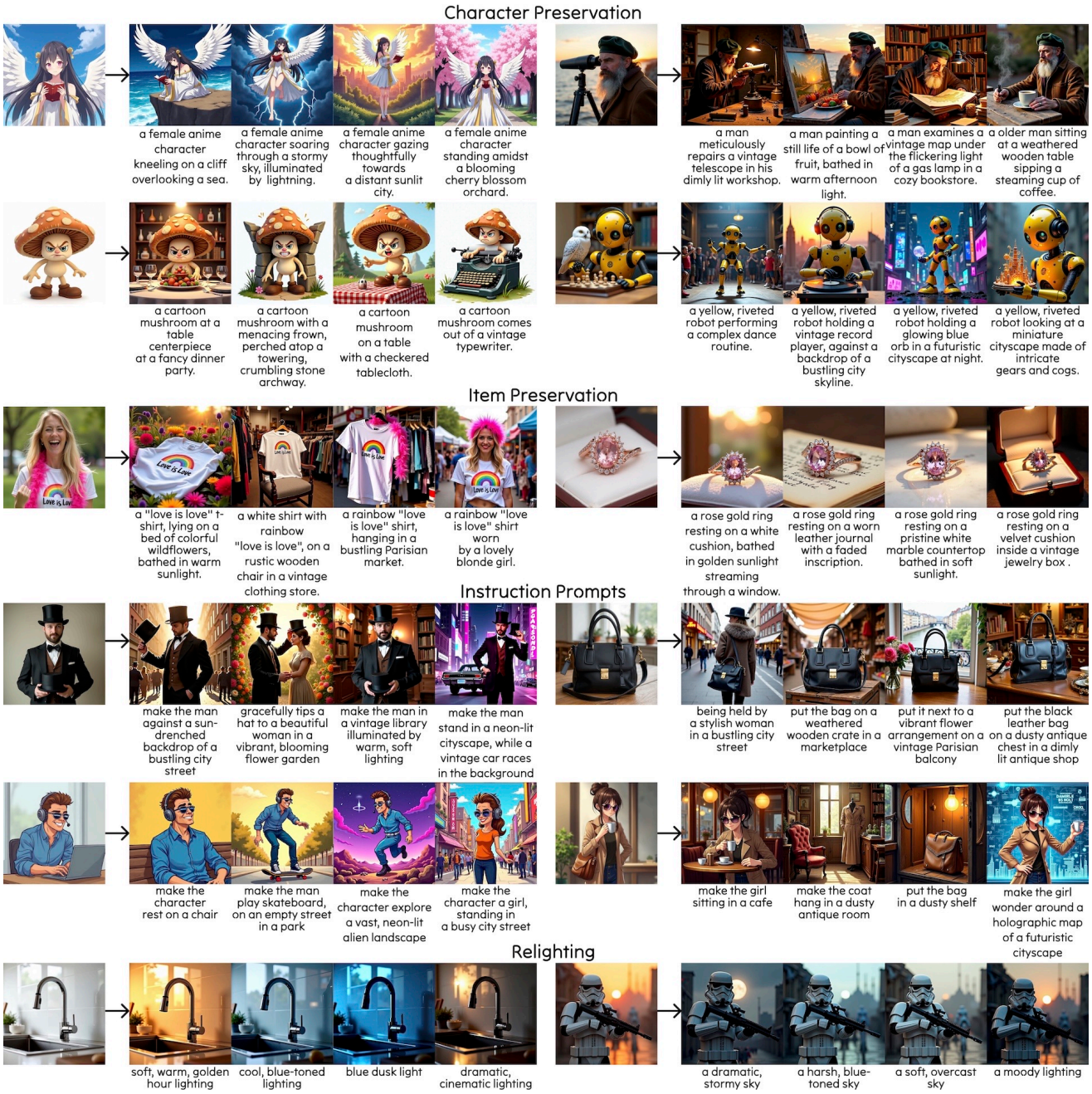


Figure 5. **Qualitative result.** Our Diffusion Self-Distillation is capable of various customization targets across different tasks and styles, for instance, characters or objects, photorealistic or animated. Diffusion Self-Distillation can also take instruction types of prompts as input, similar to InstructPix2Pix [2]. Further, our model exhibits relighting capabilities without significantly altering the scene's content.

Qualitative results. Fig. 4 presents our qualitative comparison results, demonstrating that our model significantly outperforms all baselines in subject adaptation and concept consistency while exhibiting excellent prompt alignment and diversity in the outputs. Textual Inversion [9], as an early concept extraction method, captures only vague semantics from the input image, making it unsuitable for zero-shot customization tasks that require precise subject adaptation. DreamBooth [31] and DreamBooth-LoRA [13, 31] face challenges in maintaining consistency, primarily because they perform better with multiple input images. This dependency limits their effectiveness when only a single reference image is available. In contrast, our method achieves robust results even with just one input image, highlighting its efficiency and practicality. BLIP-Diffusion [17], operating as a self-supervised representation learning framework, can extract

Method	Z-S?	Concept Preservation				Prompt Following			CP·PF↑	Debiased Concept Preservation				Debiased Prompt Following			Debiased CP·PF↑
		Animal↑	Human↑	Object↑	Overall↑	Real.↑	Imag.↑	Overall↑		Animal↑	Human↑	Object↑	Overall↑	Real.↑	Imag.↑	Overall↑	
Textual Inversion	✗	0.502	0.358	0.305	0.388	0.671	0.437	0.598	0.232	0.741	0.694	0.717	0.722	0.619	0.385	0.541	0.391
DreamBooth	✗	0.640	0.199	0.488	0.442	0.798	0.504	0.692	0.306	0.670	0.362	0.676	0.626	0.750	0.467	0.656	0.411
DreamBooth LoRA	✗	0.751	0.311	0.543	0.535	0.898	0.754	0.849	0.450	0.681	0.675	0.761	0.720	0.865	0.718	0.816	0.588
BLIP-Diffusion	✓	0.637	0.557	0.469	0.554	0.581	0.303	0.464	0.257	0.771	0.733	0.745	0.750	0.529	0.266	0.442	0.332
Emu2	✓	0.670	0.546	0.447	0.554	0.732	0.560	0.670	0.371	0.652	0.683	0.701	0.681	0.686	0.494	0.622	0.424
IP-Adapter	✓	0.667	0.558	0.504	0.576	0.743	0.446	0.607	0.350	0.790	0.764	0.743	0.766	0.695	0.377	0.589	0.451
IP-Adapter+	✓	0.900	0.845	0.759	0.834	0.502	0.279	0.388	0.324	0.481	0.473	0.530	0.504	0.442	0.229	0.371	0.187
Ours	✓	0.647	0.567	0.640	0.631	0.777	0.625	0.726	0.458	0.852	0.774	0.750	0.789	0.808	0.681	0.757	0.597

Table 1. **Quantitative result**. On the human-aligned GPT score metrics, our method is only inferior to IP-Adapter+ [42] for concept preservation (largely because of IP-Adapter families' "copy-pasting" effect) and the tuning-base DreamBooth-LoRA [13, 31] for prompt following, but outperforms every other baseline, achieving the best overall performance considering both concept preservation and prompt following. We also note that on the de-biased GPT evaluation, which penalizes "copy-pasting" the reference image without significant creative interpretation or transformation, the advantages of IP-Adaper+ [42] no longer hold. This can also be partly observed by their bad prompt following scores, meaning they are biased towards the reference input and are not accommodating the input prompt. The first, second, and third values are highlighted, where Diffusion Self-Distillation is the best overall performing model.

concepts from the input in a zero-shot manner but is confined to capturing overall semantic concepts without the ability to customize specific subjects. Similarly, Emu2 [36], a multimodal foundation model, excels at extracting semantic concepts but lacks mechanisms for specific subject customization, limiting its utility in personalized image generation. IP-Adapter [42] and IP-Adapter+ [42] employ self-supervised learning schemes aimed at reconstructing the input from encoded signals. While effective in extracting global concepts, they suffer from a pronounced "copy-paste" effect, where the generated images closely resemble the input without meaningful transformation. Notably, IP-Adapter+ [42], which utilizes a stronger input image encoder, exacerbates this issue, leading to less diversity and adaptability in the outputs. In contrast, our approach effectively preserves the subject's core identity while enabling diverse and contextually appropriate transformations. As illustrated in Fig. 5, our Diffusion Self-Distillation demonstrates remarkable versatility, adeptly handling various customization targets across different targets (characters, objects, etc.) and styles (photorealistic, animated, etc.). Moreover, Diffusion Self-Distillation generalizes well to a wide range of prompts, including instructions similar to InstructPix2Pix [2], underscoring its robustness and adaptability in diverse customization tasks.

Quantitative results. Quantitative comparison with the baselines are shown in Tab. 1, where we report GPT evaluation following DreamBench++ [25]. Such an evaluation protocol is similar to human score but uses automatic multimodal LLMs. Our method achieves the best overall performances accommodating both concept preservation and prompt following, while only being inferior to IP-Adapter+ [42] for the former (mainly because of the "copy-paste" effect again), and the per-instance tuning DreamBooth-LoRA [13, 31] for the latter. We note that the concept preservation evaluation of DreamBench++ [25] is still biased towards favoring a "copy-paste" effect, especially on more challenging and diverse prompts. For instance, the outstanding concept preservation performances of the IP-Adapter family [42] are primarily because of their strong "copy-paste" effect, which copies over the input image without considering relevant essential changes in the prompts. This can also be partly observed by their underperforming prompt following scores, which means they are biased towards the reference input and do not accommodate the input prompt. Therefore, we also present our "de-biased" version of GPT scores, which are as simple as telling GPT to penalize if the generated image resembles a direct copy of the reference image. We observe that the advantages of IP-Adaper+ [42] no longer hold. Overall, Diffusion Self-Distillation is the best-performing model.

Ablation studies. **(1)** *Data curation:* During dataset generation, we first synthesize grids using a frozen pre-trained FLUX model and then filter the images via VLM curation. Why not fine-tune the FLUX model on image grids to improve the hit rate? To study this, we fit a LoRA [13] using $>$ 7000 consistent grids (Fig. 6, left). Though a more significant proportion of samples are consistent grids, we find that the teacher model loses diversity in its output. Therefore, we choose to rely entirely on VLMs to help us curate from large numbers of diverse but potentially noisy grids. **(2)** *Parallel processing architecture:* We compare the parallel processing architecture to three alternative image-to-image architectures: 1) concatenating the source image to the noise image ("concatenation"); 2) a ControlNet [43]-based design, and 3) an IP-Adapter [42]-based design. We train each architecture using the same data as our parallel processing model (Fig. 6, middle). For ControlNet [43], we draw the same conclusion as prior work [14], in that it works best for structure-aligned edits, but generally struggles to preserve details when the source image and target image differ in camera pose. IP-Adapter [42] struggles to effectively transfer details and styles from the source image due to the limited capacity of its image encoder. **(3)** *Other image-to-image tasks:* Although not "self-distillation", since

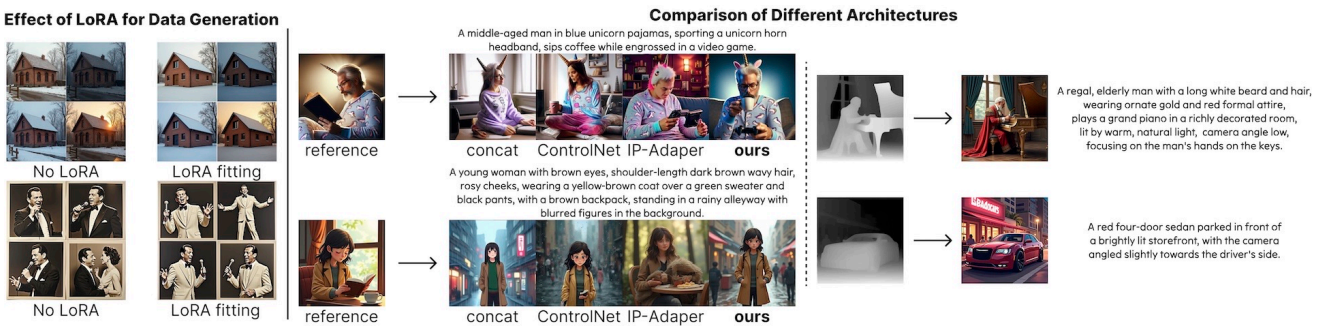


Figure 6. **Ablation study.** Left: We compare the base model's in-context sampling ability with a consistent grid LoRA-overfitted model. We observe that although applying LoRA to the base model can increase the likelihood of outputs being consistent grids, it may adversely affect output diversity. Therefore, we rely on vision-language models (VLMs) to curate from a large number of diverse but potentially noisy grids. Right: We compare our architectural design with a vanilla conditional model (by adding a few input channels), ControlNet [43], and IP-Adapter [42]. Our architecture learns the input concepts and identities significantly better. We also demonstrate that our architecture can effectively scale to depth-conditioned image generation similar to ControlNet [43].

Method	CP ↑	PF ↑	Creativity ↑
Textual Inversion [9]	1.693	1.924	2.850
DreamBooth [31]	2.329	2.883	3.597
DreamBooth LoRA [13, 31]	2.576	3.386	4.247
BLIP-Diffusion [17]	1.854	2.281	0.286
Emu2 [36]	1.843	2.096	2.965
IP-Adapter [42]	2.274	2.307	3.481
IP-Adapter+ [42]	3.733	1.959	2.428
Ours	3.661	3.328	4.453

Table 2. **User study**. "CP" refers to concept preservation scores and "PF" refers to prompt following scores. The first, second, and third values are highlighted. Our user study results mostly align with our GPT evaluation, where our Diffusion Self-Distillation is the best overall performing model.

it requires an externally-sourced paired dataset (generated with Depth Anything [40]), we additionally train our architecture on depth-to-image to demonstrate its utility for more general image-to-image tasks (Fig. 6, right).

User study. To evaluate the fidelity and prompt consistency of our generated images, we conducted a user study on a random subset of the DreamBench++ [25] test cases, selecting 20 samples. A total of 25 female and 29 male annotators, aged from 22 to 78 (average 34), independently scored each image from 1 to 5 based on three criteria: (1) concept preservation—the consistency with the reference image, (2) prompt alignment—the consistency with the given prompt, and (3) creativity—the level of internal understanding and transformation. The average scores are presented in Tab. 2. Our human annotations closely align with the GPT evaluation, demonstrating that our Diffusion Self-Distillation is slightly behind IP-Adapter+[42] in concept preservation and the inference-stage tuning method DreamBooth-LoRA [13, 31] in prompt alignment. Notably, our model achieved the highest creativity score, while IP-Adapter+ [42] scored lower in this metric due to its "copy-paste" effect. These results further confirm that our Diffusion Self-Distillation offers the most balanced and superior overall performance.

5. Discussion

We present Diffusion Self-Distillation, a zero-shot approach designed to achieve identity adaptation across a wide range of contexts using text-to-image diffusion models without any human effort. Our method effectively transforms zero-shot customized image generation into a supervised task, substantially reducing its difficulty. Empirical evaluations demonstrate that Diffusion Self-Distillation performs comparably to inference-stage tuning techniques while retaining the efficiency of zero-shot methods.

Limitations and future work. Our work focuses on identity-preserving edits of characters, objects, and scene relighting. Future directions could explore additional tasks and use cases. Integration with ControlNet [43], for example, could provide fine-grained and independent control of identity and structure. Additionally, extending our approach from image to video generation is a promising avenue of future work.

Ethics. We are mindful of the potential misuse, particularly in deepfakes. We oppose exploiting our work for purposes that infringe upon ethical standards or privacy.

Conclusion. Our Diffusion Self-Distillation democratizes content creation, enabling identity-preserving, high-quality, and fast customized image generation that adapts seamlessly to evolving foundational models, significantly expanding the creative boundaries of art, design, and digital storytelling.

Acknowledgements. This project was in part supported by Google, Kaiber AI, 3bodylabs, ONR N00014-23-1-2355, and NSF RI #2211258. ERC was supported by the Nvidia Graduate Fellowship and the Snap Research Fellowship. YZ was in part supported by the Stanford Interdisciplinary Graduate Fellowship.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. In *arXiv*, 2023. 4

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 6, 7, 1

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *https://openai.com/research/video-generation-models-as-world-simulators*, 2024. 4

[4] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *ICCV*, 2023.

[5] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Huang, Tuanfeng Wang, and Gordon. Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In *CVPR*, 2024. 4

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. 3, 5

[8] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 4

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 5, 6, 8

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *arXiv*, 2023. 4

[11] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. In *arXiv*, 2022.

[12] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *arXiv*, 2022. 4

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3, 5, 6, 7, 8

[14] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *arXiv*, 2023. 4, 7

[15] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multitask learners. In *arXiv*, 2024. 3

[16] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon. Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *NeurIPS*, 2024. 4

[17] Dongxu Li, Junnan Li, and Steven Hoi. BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 5, 6, 8

[18] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *arXiv*, 2023. 4

[19] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2024. 3

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[21] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *SIGGRAPH*, 2024. 3

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 3

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *arXiv*, 2021. 1

[25] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *arXiv*, 2023. 5, 7, 8, 1, 2

[26] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit H Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. In *arXiv*, 2023. 3

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *CoRR*, 2021. 5

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. In *arXiv*, 2022. 1

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[30] Ciara Rowles, Shimon Vainer, Dante De Nigris, Slava Elizarov, Konstantin Kutsy, and Simon Donné. Ipadapter-instruct: Resolving ambiguity in image-based conditioning using instruct prompts. In *arXiv*, 2024. 3, 5

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2, 3, 5, 6, 7, 8

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2, 4

[34] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. In *SIGGRAPH Asia*, 2024. 4

[35] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *arXiv*, 2023. 4

[36] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 5, 7, 8

[37] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. In *arXiv*, 2024. 2, 3

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 4, 1

[39] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. In *arXiv*, 2023. 4

[40] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 8

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *arXiv*, 2024. 4

[42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. In *arXiv*, 2023. 2, 3, 4, 5, 7, 8

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 7, 8

[44] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Unicontrolnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 3

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 [38]. 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 [2]-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++ [25] 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. 8, we demonstrate more of the qualitative evaluation cases from the DreamBench++ [25] 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. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13 and Fig. 14,

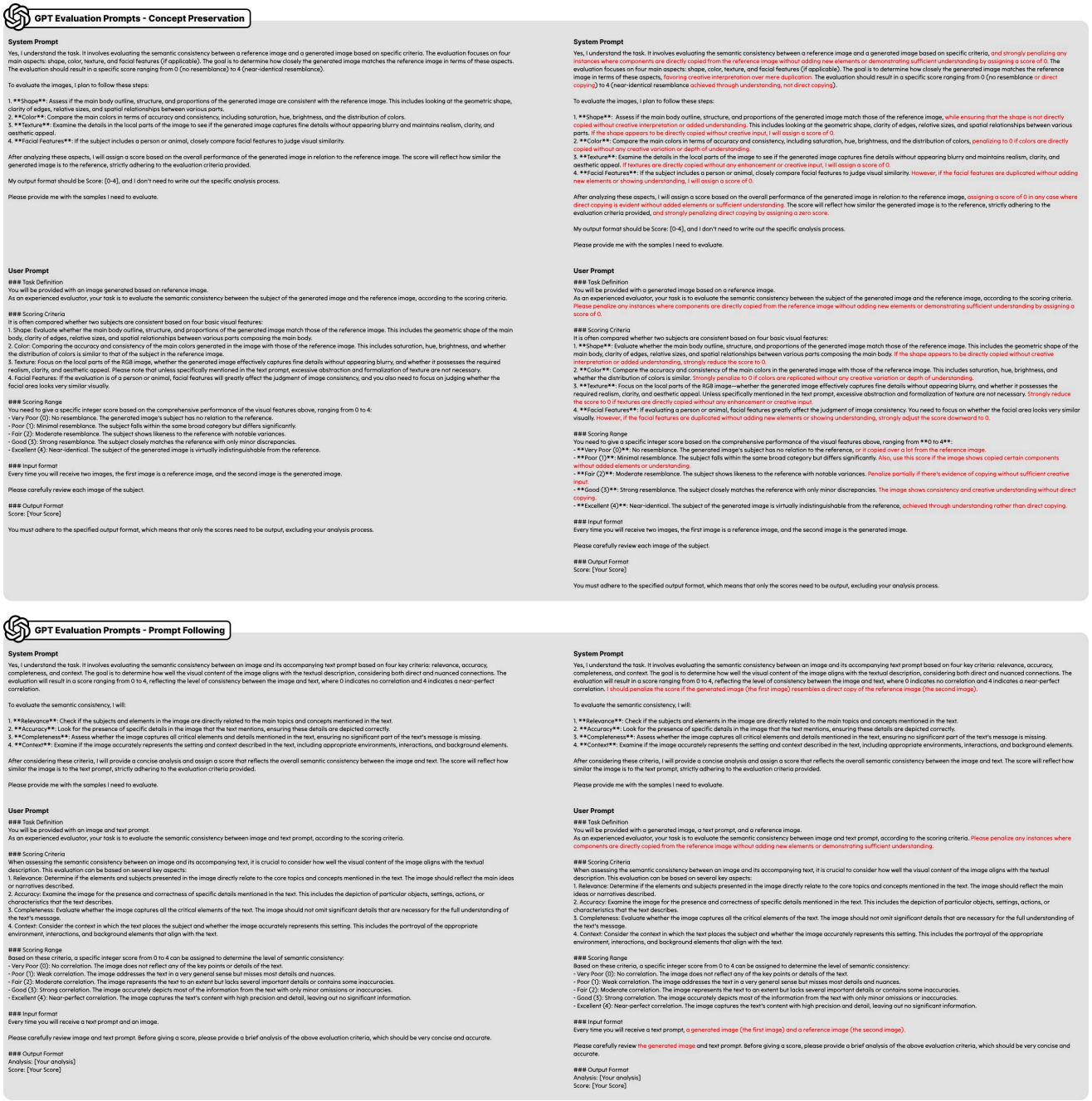


Figure 7. **GPT evaluation prompts** used across our evaluation, where the left shows the vanilla prompts from DreamBench++ [25] and the right shows our modified "de-biased" prompts, which strongly penalizes "copy-pasting" effects without sufficient creative inputs. We highlight our modified sentences in red.

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. 15 and Fig. 16, 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.

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. 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.

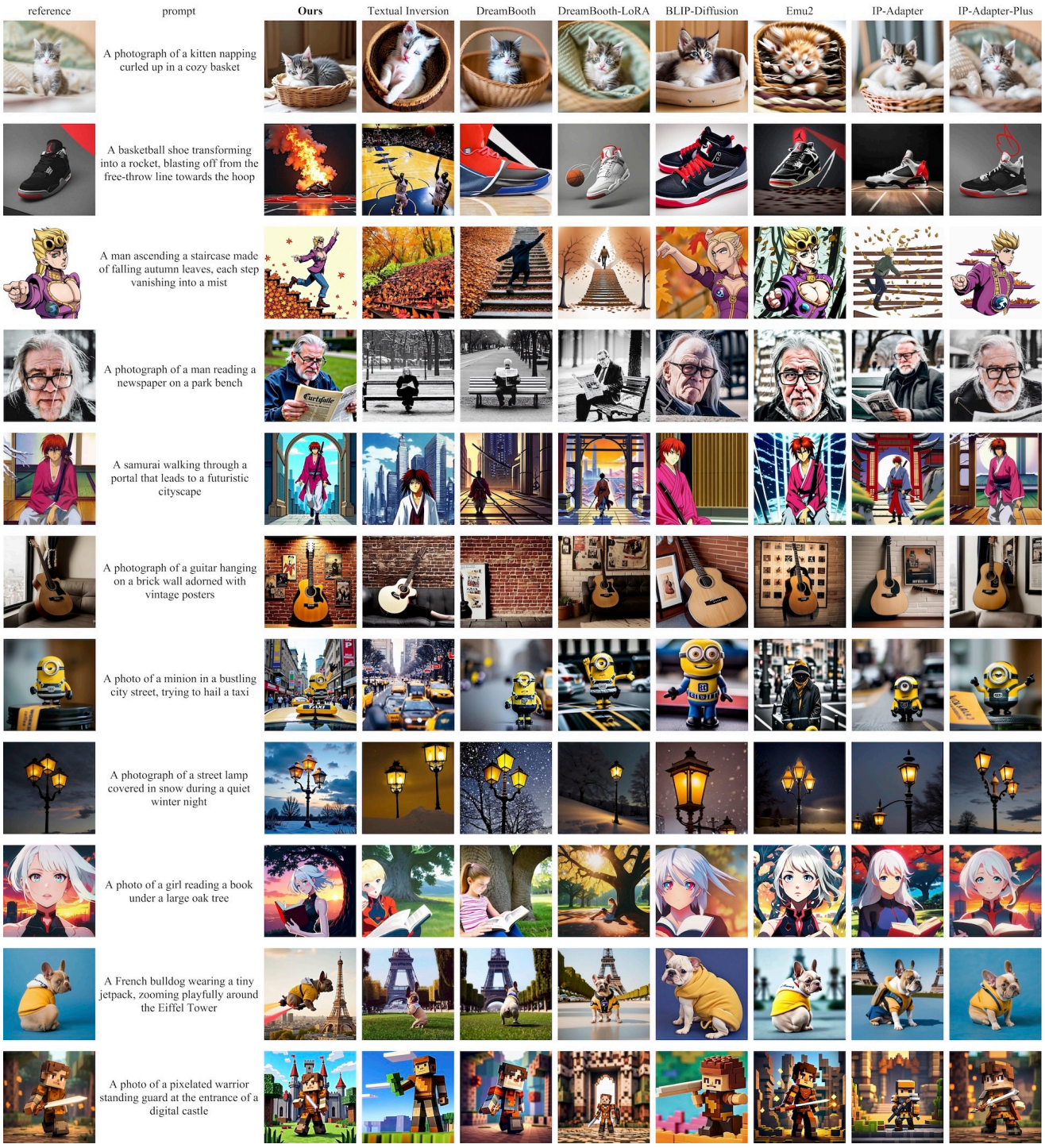


Figure 8. **Additional qualitative comparison.**

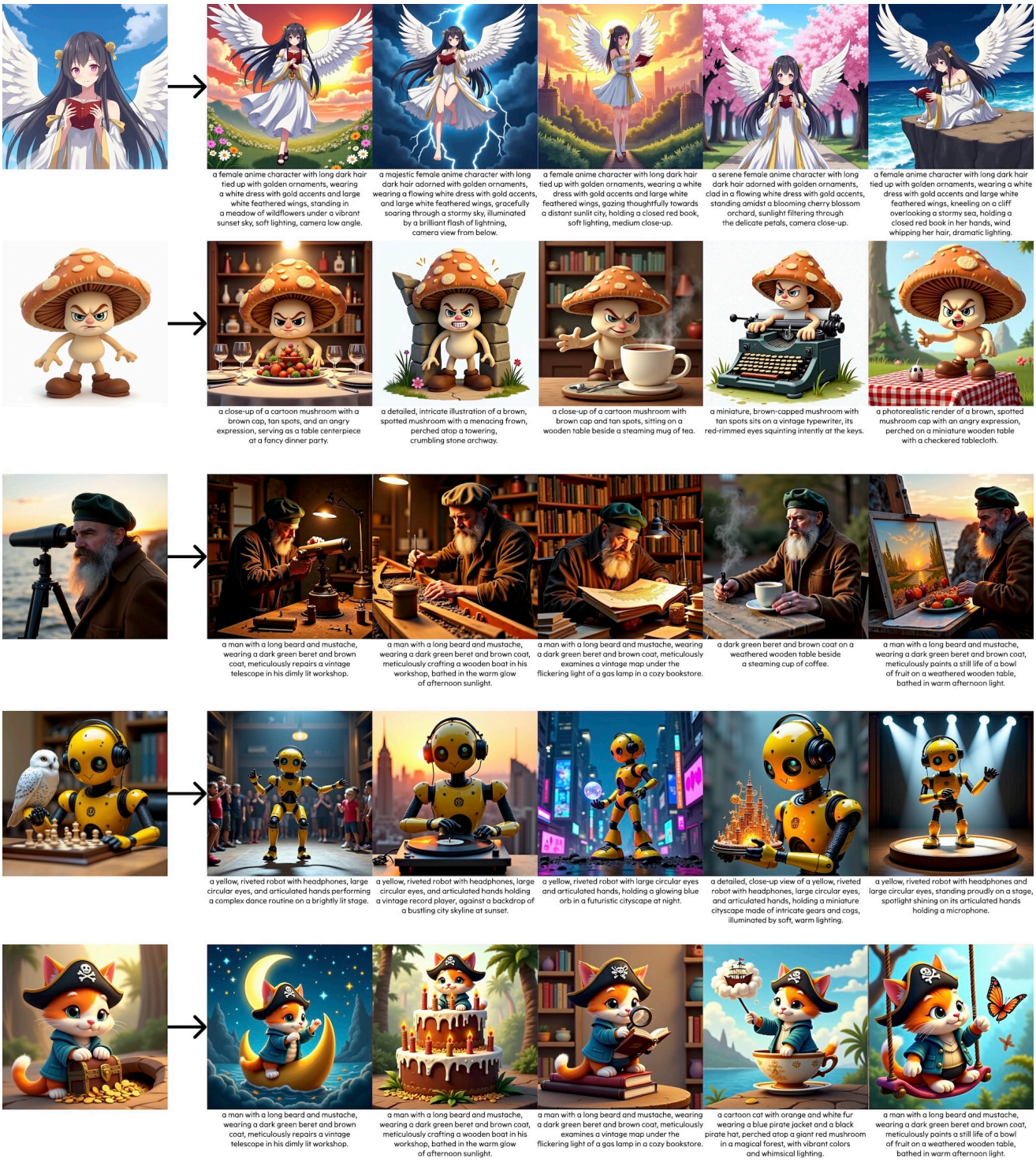


Figure 9. **Additional character identity preserving results.**

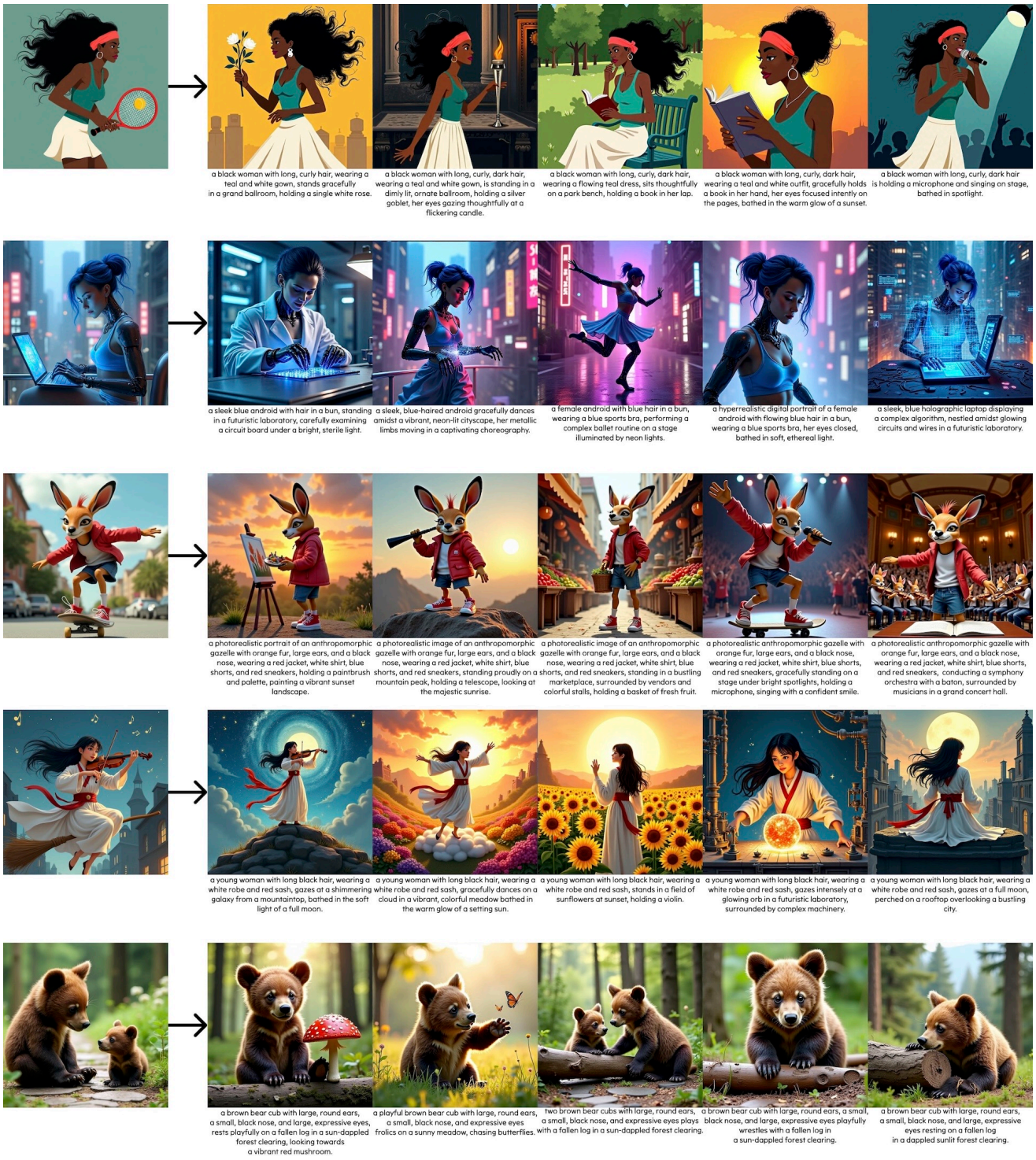


Figure 10. **Additional character identity preserving results.**

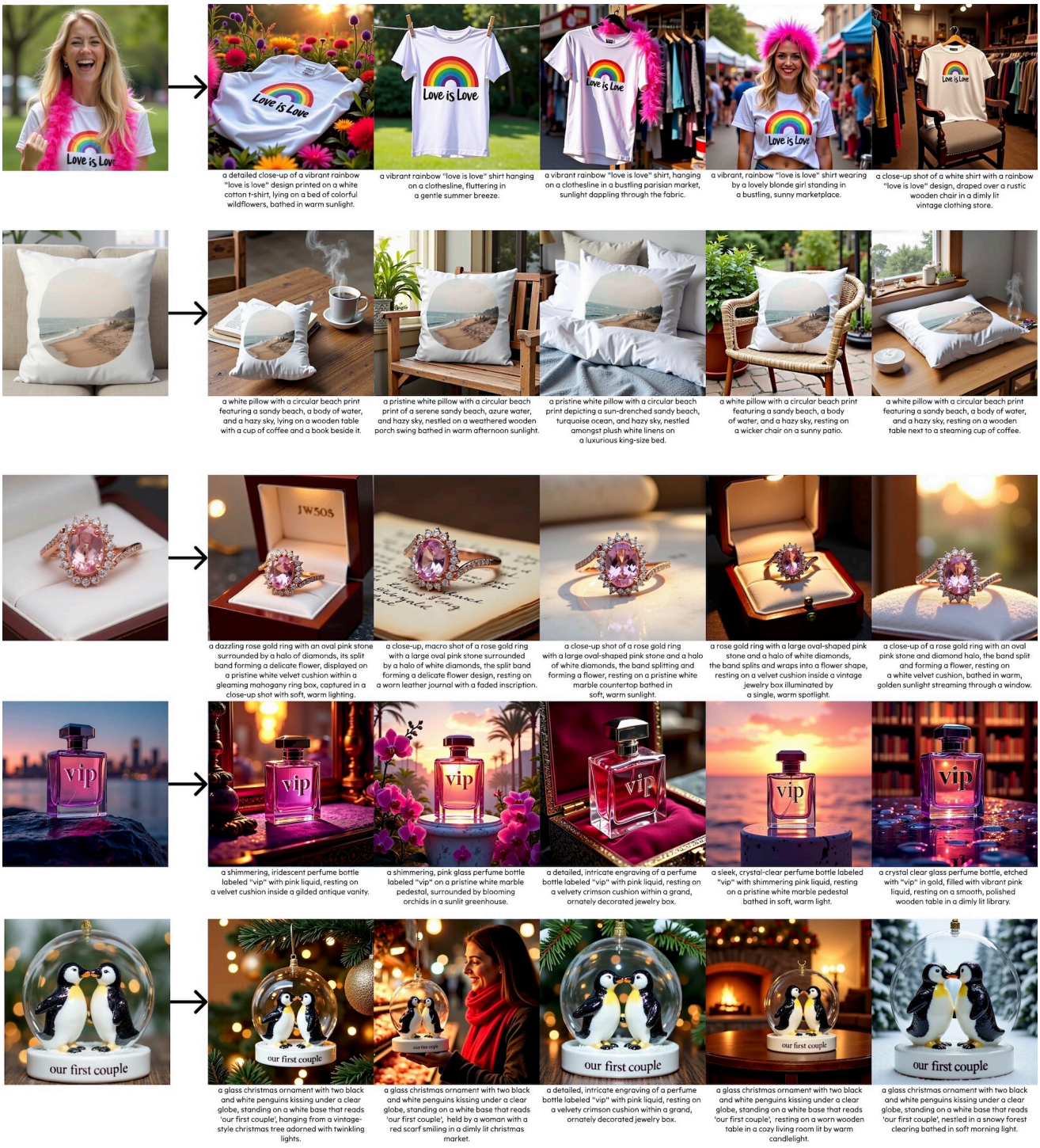


Figure 11. **Additional object/item identity preserving results.**

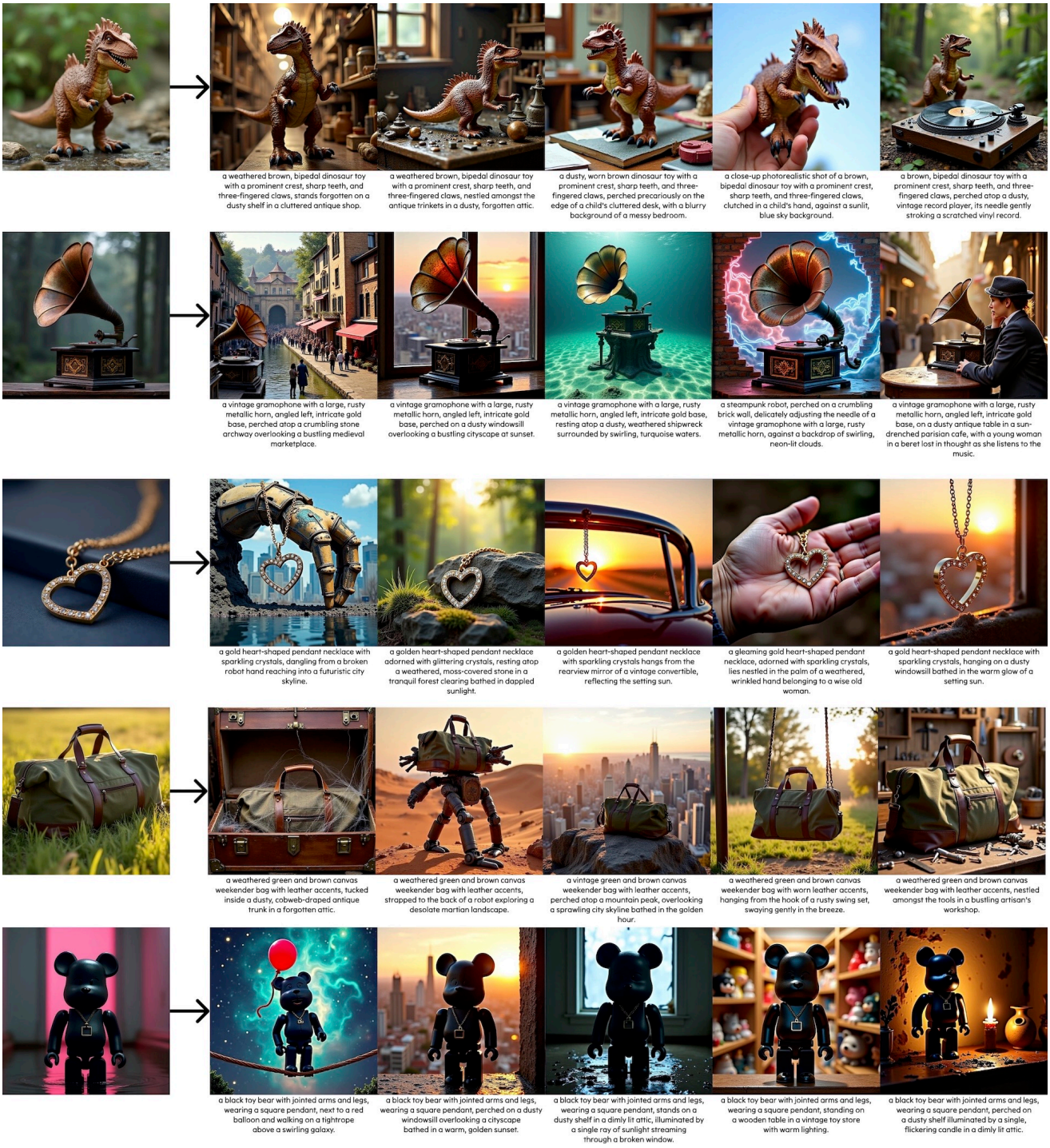


Figure 12. **Additional object/item identity preserving results.**

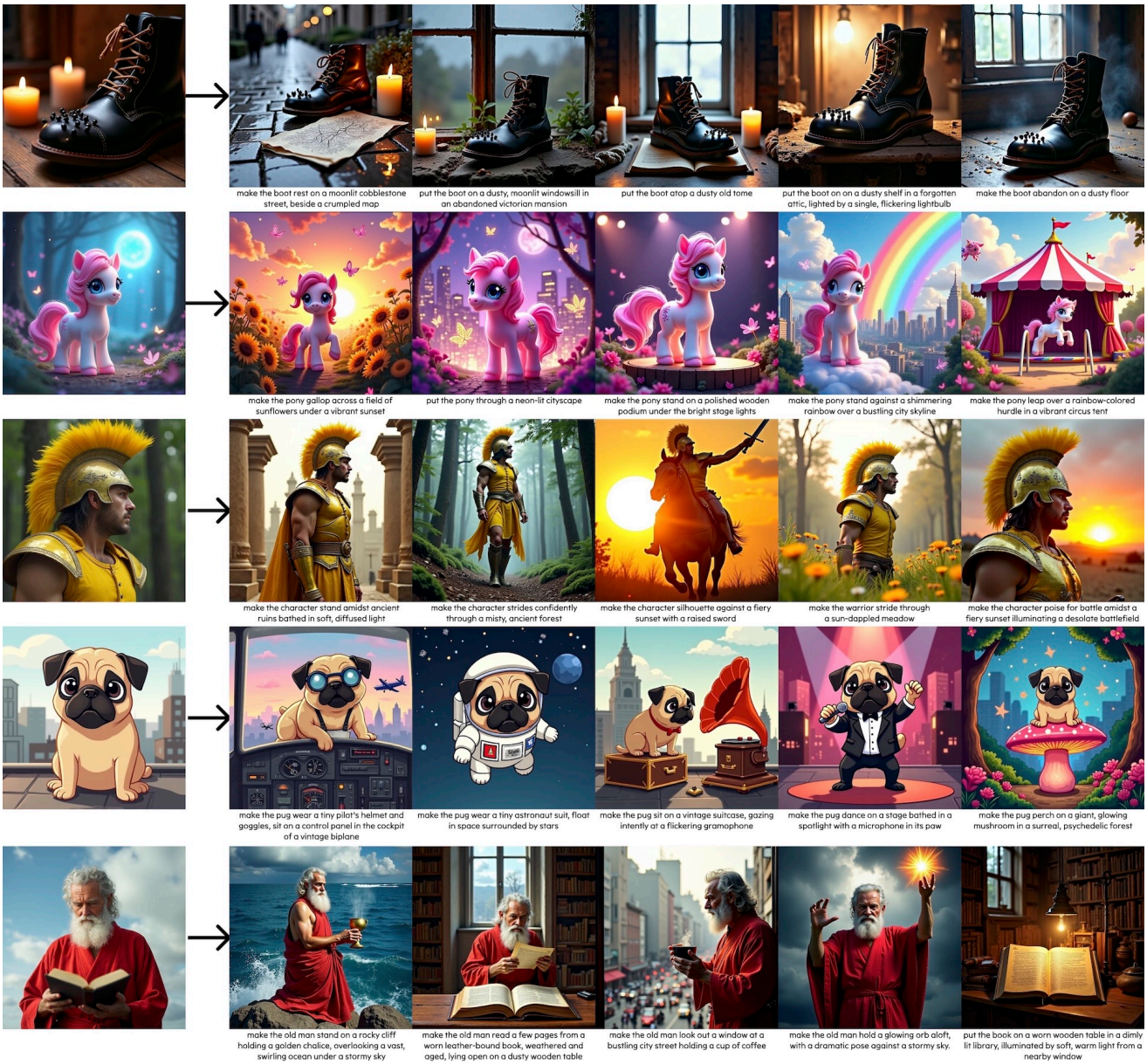


Figure 13. **Additional instruction prompting results.**

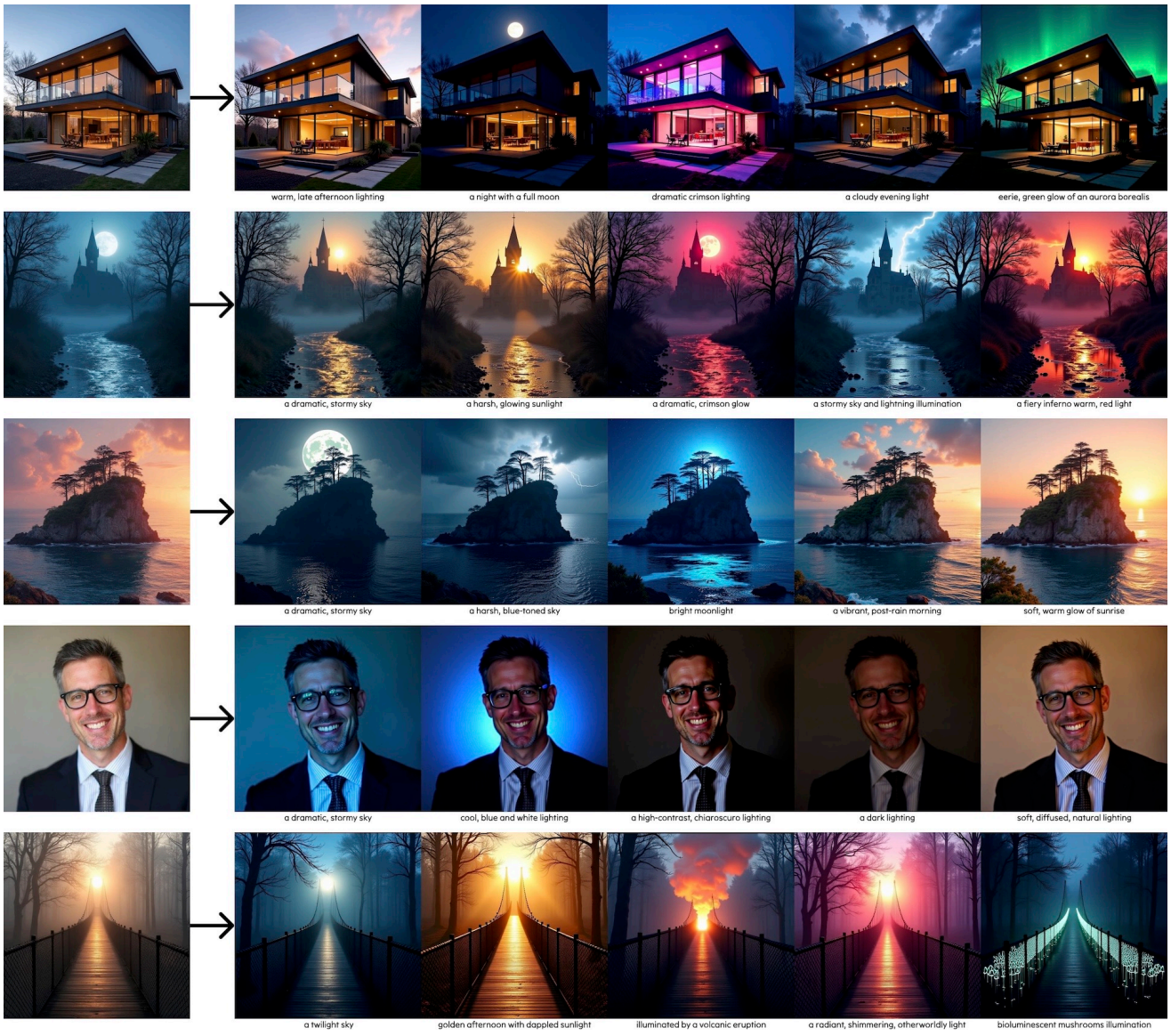
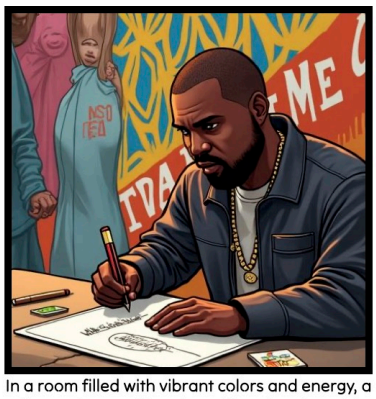


Figure 14. **Additional relighting results.**

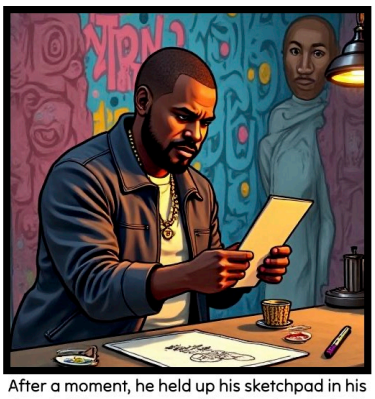
Figure 15. **Comic generation example 1.** The conditioned image is the first panel.

1



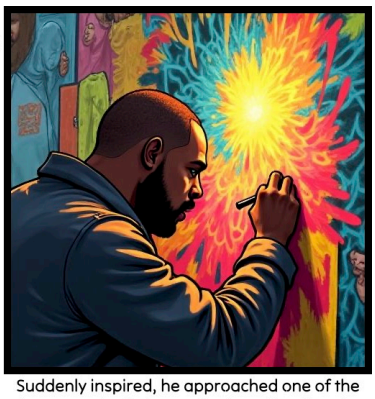
In a room filled with vibrant colors and energy, a focused man with a shaved head and a gold chain sat at a table. He was deeply engrossed in drawing on a sheet of paper, his pencil moving with purpose, while warm light spilled over graffiti-like murals painted on the walls around him. His expression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

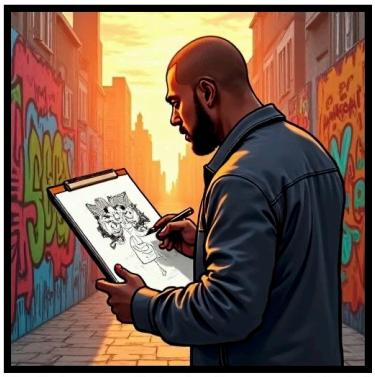
7

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