# **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-40 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.



### 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, bottomleft, 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.



### 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).



### **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

### GPT Evaluation Prompts - Concept Preservation

#### System Prompt

Yes, Lunderstand the task. In involves evoluting the semantic consistency between a reference image and a generated image based on specific criteria. The evolution focuses on four moin species: shape to confi testures (if applicable). The goal is taket determine have closely the generated image matches the reference image in terms of these aspect the evoluation should result in a specific score ranging from 0 (no resemblance) to 4 (near-identical resemblance).

To evaluate the images, I plan to follow these steps:

1.\*\*Snope\*\* Assess the mon body outline, structure, and proprintion of the generate a mage are consistent with the reference mage. This includes looking at the generatic structure is and a structure is and an another the structure is and an another the generatic structure. The structure is and an another is an another is an another is an another includes a structure. The structure is another is an another is an another is an another includes a structure in the structure is another includes a structure. The structure is another is an anothe

aesthetic appeal. 4. \*\*Facial Features\*\*: If the subject includes a person or animal, closely compare facial features to judge visual similarity.

After analyzing these aspects, I will assign a score based on the overall performance of the generated image in relation to the reference

My output format should be Score: [0-4], and I don't need to write out the specific analysis

Please provide me with the samples I need to evaluate.

#### ### Task Defi

You will be provided with an image generated based on reference image. As an experienced evaluator, your task is to evaluate the semantic consistency between the subject of the generated image and the reference image, according

### Scoreg Chero The Schero compared whether two subjects and consistent baced on four bacic kined features: The Schero compared whether two subjects and consistency of the grantices of the grantices of the rest baced of the schero compared whether the schero compared with the schero compared with the schero of the rest compared whether the schero compared wheth

tou head to give a specific integer score based on the componentsive performance of the visual reduces dowe, ranging from 0 to 4: Very Poor (0): Morimal resemblance. The generated image's subject has no relation to the reference. - Poor (0): Morimal resemblance. The subject fulls within the same broad category but differs significantly. - Sin (0): Morimal resemblance. The subject fulls within the same broad category but differs significantly.

Goad (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
 Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the refere

### Input format

very nine you will receive two intoges, the institutige is a reference intoge, and the second intoge is the generated intoge. Nease corefully review each image of the subject.

### Output Format Score: [Your Score]

u must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.

# GPT Evaluation Prompts - Prompt Following

#### System Prompt

(s), (understand the task.) In invites evolution if the semantic consistency between on image and its accompanying teap prompt based on four key orients relevance, accuracy, monopheneses, and construct. The goal is to determine have well the valua content of the image and last constructions. The value is determined to accurate the semantic consistency between the image and text, where 0 indicates no correlation and 4 indicates a near-perfect andiation.

#### To evaluate the semantic consistency, I will

1 \* Halenoors\*\* Oecki film subjects and elements in the image and electry-related to the main topics and concepts mentioned in the text.
2 \* Accurges\*\* Lock for the presence of approximation of electric and elements in the inclusion (elements).
4 \* Complete \*\* Lock elements \*\* Assess whether the image captures of includ elements and elements and elements. Including elements elements including elements.
A \* Complete \*\* Lock elements \*\* Assess whether the image captures of includ elements and elements.
A \* Complete \*\* Lock elements elements in the image captures of including elements.
All elements \*\* Assess whether the image captures of including elements and elements.
All elements \*\* Assess whether the image captures of including elements and elements.
All elements \*\* Assess whether the image captures of including elements and elements.
All elements \*\* Assess whether the image captures of including elements.
All elements \*\* Assess whether the image captures of including elements.
All elements \*\* Assess whether the image captures of including elements.
All elements \*\* Assess whether the image captures of including elements.
All elements \*\* Assess whether the image captures of including elements.
All elements \*\* Assess whether the image captures of including elements.
All elements \*\* Captures \*\* Assess whether the image captures of including elements.
All elements \*\* Assess \*\* Assess elements in the image captures of including elements.
All elements \*\* Assess \*\* Assess \*\* Assess elements in the image captures of including elements.
All elements \*\* Assess elements in the image captures of including elements.
All elements \*\* Assess elements in the image captures of including elements.
All elements \*\* Assess elements \*\* Assess elements elements elements elements elements elements e

Heer	Drownet	

User Prompt
### Tosk Definition
You will be provided with an image and text prompt.
As an experienced evaluator, your task is to evaluate the semantic consistency between image and text prompt, according to the scoring criteria.
### Scoring Criteria
When assessing the semantic consistency between an image and its accompanying text, it is crucial to consider how well the visual content of the image aligns with the textual description. This evaluation can be based on several key aspects:
1. Relevance: Determine if the elements and subjects presented in the image directly relate to the core topics and concepts mentioned in the text. The image should reflect the main ideas or narratives described.
<ol> <li>Accuracy: Examine the image for the presence and correctness of specific details mentioned in the text. This includes the depiction of particular objects, settings, actions, or characteristics that the text describes.</li> </ol>
3. Completeness: Evaluate whether the image captures all the critical elements of the text. The image should not omit significant details that are necessary for the full understanding of the text's message.
4. Context: Consider the context in which the text places the subject and whether the image accurately represents this setting. This includes the portrayal of the appropriate environment, interactions, and background elements that align with the text.
### Scoring Range
Based on these criteria, a specific integer score from 0 to 4 can be assigned to determine the level of semantic consistency:
<ul> <li>Very Poor (0): No correlation. The image does not reflect any of the key points or details of the text.</li> </ul>
<ul> <li>Poor (1): Weak correlation. The image addresses the text in a very general sense but misses most details and nuances.</li> </ul>
<ul> <li>Fair (2): Moderate correlation. The image represents the text to an extent but locks several important details or contains some inaccuracies.</li> </ul>
<ul> <li>Good (3): Strong correlation. The image accurately depicts most of the information from the text with only minor omissions or inaccuracies.</li> </ul>
<ul> <li>Excellent (4): Near-perfect correlation. The image captures the text's content with high precision and detail, leaving out no significant information.</li> </ul>
### Input format
Every time you will receive a text prompt and an image.
Please carefully review image and text prompt. Before giving a score, please provide a brief analysis of the above evaluation criteria, which should be very concise and accurate.
### Output Format
Analysis: [Your analysis]
Score: [Your Score]

estand the task. It involves evaluating the semantic consistency where components are directly copied from the reference image focuses on four main aspects: shape, color, texture, and facial t

After analyzing these aspects, I will assign a score based on the overall performance of the generated image in relation to the reference image, assigning a score of 0 in any case where direct copying is wident without added elements or sufficient understanding. The score will reflect how similar the generated image is to the reference, strictly adhering to the with string with a score of 0 and 1 and 1

My output format should be Score: (0-4), and I don't need to write out the specific analysis process.

Please provide me with the samples I need to evaluate.

#### User Promp

## It is 1.\* int 2.1 wh 3.4

### Task Definition

is an experienced evoluator, your task is to evaluate the semantic consistency between the subject of the generated image and the reference image, according to the scoring criteria. Neare penalities any instances where components are directly copied from the reference image without adding new elements or demonstrating sufficient understanding by assigning accord 0.

coring Criteria	
ten compared whether two subjects are consistent based on four basic visual features:	
sope**: Evaluate whether the main body outline, structure, and proportions of the generated image match those of the reference image. This includes the geometric shape of the oddy, clarity of edges, relative sizes, and spatial relationships between various parts composing the main body. If the shape appears to be directly copied without creative relation or and/estimations.	
olor**: Compare the accuracy and consistency of the main colors in the generated image with those of the reference image. This includes saturation, hue, brightness, and er the distribution of colors is similar. Strongly penalize to 0 if colors are replicated without any creative variation or depth of understanding.	
exture <sup>4+</sup> : Focus on the local parts of the RGB image—whether the generated image effectively captures fine details without appearing blurry, and whether it possesses the ad realism, clarity, and oesthetic appeal. Unless specifically mentioned in the test prompt, excessive abstraction and formalization of testure are not necessary. Strongly reduce are to 0 if testures or directly cogies directly and provide the provide test prompt.	
acial Features**: If evaluating a person or animal, facial features greatly affect the judgment of image consistency. You need to focus on whether the facial area looks very similar y; However, if the facial features are duplicated without adding new elements or showing understanding, strongly adjust the score downward to 0.	

su need to give a specific integer score based on the comprehensive performance of the visual features above, ranging from \*\*0 to 4\*\*: \*\*Very Paor (0)\*\*: No resemblance. The generated image's subject has no relation to the reference, or it copied over a lot from the reference image.

\*\*\*Nor (U<sup>++</sup>) Minurel resemblance. The subject fails within the same broad category but differs significantly, Max, use this core if the image how cooled scration promovers without obled externols or understanding. \*\*\* Gar (U<sup>++</sup>) Single constructions and the subject shows likeness to the reference with no hind discretions. Provide participal without sufficient creative \*\* Gar (U<sup>++</sup>). Single constructions are subject shows likeness to the reference with no hind discretions. The mage hows consistence and creative understanding without sufficient creative \*\* Gar (U<sup>++</sup>). Single constructions are sufficient to the subject shows likeness to the reference with no hind discretions. The mage hows consistence and creative understanding without are shown (I).

wing. \*Excellent (4)\*\*: Near-identical. The subject of the generated image is virtually indistinguishable from the reference, achieved through understanding rather than direct copying.

### Input format Every time you will receive two images, the first image is a reference image, and th

Please carefully review each image of th

### Output Format Score: [Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis proce

ystem Prompt es, I understand the task. It involves evaluating ompleteness, and context. The goal is to determ aduation will result in a score ranging from 0 to

ation. I should parallate the score if the generated image (the first image) resembles a direct copy of the reference image (the second image). share the examine consistency, thit: elevance\*\*. Check if the subjects and elements in the image are directly related to the main topics and concepts mentioned in the text. scoresys\*: Lock of the subjects and elements in the image and directly related to the main topics and concepts mentioned in the text.

\*\*Context \*\* Loamier if the image accurately represent the starting and context described in the text, including appropriate environments, interactions, and background elements. The constraint place starts in all provide constraints and aspire access that reflects the overall semantic consistency between the image and text. The score will reflect how mistr the image is to the text prompt, strainty adhering to the evaluation orders provided.

### Jser Prompt

Note appriced with a generated image, a text group, and a reference image.
 See appriced workshop, where it is a workshop is the sense image and text group, succedary to the scoring orders, these pendets ary instances where concentration and early cost and text provide the sense image and text group. Succedary to the scoring orders, these pendets ary instances where concentration and early cost and text provide the sense image and text provide to the scoring orders, these pendets ary instances where concentration and early cost and text provide the sense image and text provide to the score order order order text sense image and the score order order order text sense image and the score order order order order text sense image and text provide to the score order order order order text sense image and text provide to the score order or

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-40, 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 FLUXbased 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 Control-Net 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)	1	0.355	0.869	0.308	0.491	0.848	0.416
IP-Adapter	1	0.380	0.717	0.272	0.500	0.756	0.378
IP-Adapter (our data)	1	0.414	0.712	0.295	0.512	0.723	0.370
Ours	1	0.631	0.726	<u>0.458</u>	<u>0.789</u>	<u>0.757</u>	<u>0.597</u>



### 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.





th long dark nents, wearing cents and large nding in "bran" ith aold ao



Iden ornaments, weari gold accents and large wings, kr









oushr miniature, brown-capped tan spots sits on a vintage med eyes squinting i







ith a long bea dark green be sly crafting a

, a man with a lo iat, a dark green be ie



a man with a long beard and must-earing a dark green beret and brow meticulously paints a still life of a l of fruit on a weathered wooden to













dripp ng, gold coins, o by lush tropico lite

a cartoon cat with orange and white fur wearing a blue jacket and a black pirate hat, holding a magnifying glass and reading an ancient book, sitting on stacked books atop a wooden table in a cozy library.

a cartoon ca wearing a blu rate hat, perch in a magical nd a black a giant red mushr

a cortoon cat with or unself wearing a blue pirate jacket and a black pirate hat, sitting on a wooden swing in mil motion, reaching toward a vibrant monard butterfly, surrounded by lush green forest trees under bright daylight

Figure 10. Additional character identity preserving results.









yes gazing thoughtfully at a





and, her eyes focused intently on ned in the warm glow of a sunset.





in a futuristic laboratory, carefully examining a circuit board under a bright, sterile light. amidst a vibrant, neon-lit cityscape, her metall limbs moving in a captivating choreography.





a complex algorithm, nestled amidst gl circuits and wires in a futuristic labore





ng in a b ting in a bustling vendors and at of fresh fruit.







a young woman with long black hair, wearing a white robe and red sash, gazes at a full moon, perched on a rooftop overlooking a bustling city. warran with long black hair, wearing a ayoung worran with long black hair, wearing a ayoung se and red sash, gazes at a shimmering white robe and red sash, gracefully dances on a white rom a mountaintop, bathed in the soft cloud in a vibrant, colorful meadow bathed in light of a full moon. the warran glow of a setting sun. ata obe and red so flowers at sun sh, stands in a field of be and red sash, gazes inter ring orb in a futuristic labore atory glo





Figure 11. Additional character identity preserving results.







by a lovely blonde girl standing a bustling, sunny marketplace

nt rainbow "love is love" shirt hanging on a clothesline, fluttering in a gentle summer breeze. a defailed close-up of a vibrant rainbow "love is love" design printed on a white cotton t-shirt, lying on a bed of colorful wildflowers, bathed in warm sunlight.

a clothesline in a bustling parisian ma sunlight dappling through the fabric.

up shot of a white shirt with a is love" design, draped over o wooden chair in a dimly lit vintage clothing store.





a white pillow with a circular beach print featuring a sandy beach, a body of water, and a hazy sky, lying on a wooden table with a cup of coffee and a book beside it.

print of a serene sandy beach, azure water, and hazy sky, nestled on a weathered wooden barch swing bathed in warm afternoon sunlight p

adopicting a sun-drenched sandy bea rquoise ocean, and hazy sky, nestled amongst plush white linens on a luxurious king-size bed.



a white pillow with a circular beach print eaturing a sandy beach, a body of water and a hazy sky, resting on a wooden table next to a steaming cup of coffee.







a close with a large o of white d form





a close-up of a rose gold ring with an ov pink stone and diamond halo, the band sy and forming a flower, resting on a white velvet cushion, bathed in warm golden sunlight streaming through a wind





a shimmering, iridescent perfume bottle labeled "vip" with pink liquid, resting on relvet cushion inside a gilded antique vanity.



a detailed, intricate engraving bottle labeled "vip" with pink liqu a velvety crimson cushion with ornately decorated jewel a grand











a glass christmas arnament with two black and white penguins kissing under a clear globe, standing on a white base that reads 'our first coupe', held by a woman with a red scarf smiling in a dimly lit christmas market.

a detailed, intricate engraving of a perfume bottle labeled "vip" with pink liquid, resting o a velvety crimson cushion within a grand, ornately decorated jewelry bo

a glass christmas ornament with two – and white penguins kissing under a clear globe, standing on a white base that reads 'our first couple', resting on a worn wooden table in a cozy living room lit by warm candlelight.

stal



a glass christmas ornament with two b and white penguins kissing under a cle globe, standing on a white base that re 'our first couple', nestled in a snowy fo clearing bathed in soft morning ligh rest









pror fin





metallic ha base, pe

ngled left, gold b

ofa ud

metallic base, on drenched parisian cafe, with a young woman in a beret lost in thought as she listens to the music.







ten heart-shaped pendant neckl d with glittering crystals, resting athered, most-country in a ga ador adorneu winn g. a weathered, moss-coverea sion c tranquil forest clearing bathed in da sunlight. ppled reflecting the setting sun



gold heart-shaped pendant necklace wit sparkling crystals, hanging on a dusty windowsill bathed in the warm glow of a setting sun.







# a weathered green and brown canvas eekender bag with leather accents, tucked inside a dusty, cobweb-draped antique trunk in a forgotten attic.

ack toy bear with jointed arms and legs aring a square pendant, next to a red balloon and walking on a tightrope above a swirling galaxy.

ba



n table in a vintage with warm lighting.

ge toy store

aw

th

Figure 13. Additional object/item identity preserving results.

windowsill overlooking a cityscape bathed in a warm, golden sunset.



aring a square penaant, stanas or r shelf in a dimly lit attic, illuminate a single ray of sunlight streaming through a broken window.





Figure 14. Additional instruction prompting results.



Figure 15. Additional relighting results.

1



In a room dimly lit by a single lamp, a serious man in a dark suit sat at a wooden table, reading an old book intently. Framed portraits adorned the green walls, and shadows shifted subtly under the soft, directional lighting. The man's expression was deeply focused, as if the secrets of the universe lay within the pages.



Suddenly, he realized something, his intense gaze locked onto a passage he had just read. The warm lamplight threw his shadow across the room, making it loom large against the walls and highlighting his furrowed brow. Something he had discovered in the book seemed urgent almost alarming.



He leaned forward over the table, urgently flipping through the pages. His hands trembled slightly, and the golden light from the lamp illuminated his tense features, deepening the lines of concentration etched into his face. Whatever he sought, he was desperate to find it.



Raising his head, the man's eyes fixed on the wall of portraits. Holding the old book open in one hand, he approached the paintings, his eyes narrowing with focus. The faces in the frames seemed to stare back at him, and he scanned each one carefully, as if hoping to find something—some connection—that only he could see.



With a sense of determination, he reached out to touch a specific portrait, the old book tucked under his arm. His fingers lightly brushed the frame, and his expression grew thoughtful, curious. The soft light emphasized his focused gaze, as if the touch itself might reveal a hidden truth.



Just then, a creaking sound broke the silence. A hidden door began to open in the wall, and the man stepped back in shock, his eyes wide. The old book clutched in his hands, he stared at the widening gap, where light from a secret passage spilled into the room, creating eerie, shifting shadows.



Steeling himself, he cautiously stepped into the narrow corridor. The passage was lined with dusty bookshelves, and the faint, flickering light barely illuminated the space. He held the old book close, his expression a mix of wariness and determination, ready to face whatever lay ahead.



Midway through the corridor, he stopped abruptly. Glowing symbols began to appear on the walls, casting an ethereal light that danced around him. The man's face was illuminated, a look of wonder mingling with his focused determination, as if he was on the brink of understanding a great mystery.



Finally, he reached the end of the passage, where a grand, ancient door loomed before him. It was adorned with intricate, glowing runes that pulsed with a life of their own. The man held the old book tightly against his chest, awe and anticipation mixing in his expression, while radiant light seeped through the cracks.

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



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.



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.



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.



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.



But he was not done vet. 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.



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.





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.



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.