Image Generation from Contextually Contradictory Prompts

Supplementary Material

1. Ethics Statement

Our work contributes to improving the semantic alignment of text-to-image models under contradictory or biased prompts. As a consequence, our method enhances users' ability to control generative models and faithfully render contradictory concepts. While this provides positive benefits, such as reducing unintended biases and enabling more inclusive image generation, it also increases the potential for misuse, including the creation of harmful, misleading, or inappropriate content. As with any advance in generative modeling, these dual-use concerns highlight the importance of responsible deployment, safeguards, and continued ethical oversight to ensure that such improvements contribute positively to society.

2. Additional Results

Robustness across LLMs. Since our framework hinges on LLM-driven prompt decomposition, we further examined its robustness under different language models. We evaluated both a proprietary model (GPT-4o) and a comparatively lightweight open-source alternative (LLaMA-3.1-8B-Instruct). While GPT-4o delivers the strongest performance, the smaller LLaMA-3.1-8B-Instruct still yields consistent improvements over the baseline (see Table 1).

Improved realism. SAP generates photorealistic and semantically coherent images for prompts with atypical attribute combinations (Figure 1). In contrast, FLUX often defaults to cartoon-like renderings, even when photorealism is explicitly requested, revealing a contextual contradiction between fantastical content and realistic style. By using contradiction-free proxy prompts, SAP avoids these biases and produces realistic outputs regardless of whether photorealism is explicitly required in the prompt.

Non-contradictory prompts. To ensure applicability in general text-to-image scenarios, we verify that our method does not negatively affect prompts without contextual contradictions. We find that including even a single non-contradictory in-context example is sufficient for the LLM to default to using the full prompt in such cases. We evaluate this behavior using GPT-4o alignment scores on the PartiPrompts-Simple benchmark, which contains simple, non-contradictory prompts (Table 2).

Additional qualitative comparisons. Figures 3 and 4 present additional qualitative comparisons of our method, while Figure 5 shows results across multiple seeds.

Table 1. *Performance of SAP when combined with different LLMs, comparing GPT-4o and Llama-3.1-8B-Instruct.*

Models	Benchmarks		
	Whoops	Whoops-Hard	Contra-Bench
FLUX	78.85	44.3	57.16
$\text{SAP}_{\text{GPT4o}}$	85.10	62.13	66.16
$\text{SAP}_{\text{Llama3.1}}$	80.52	59.53	61.16

Table 2. *Alignment performance on the PartiPrompts-simple benchmark, which contains simple, non-contradictory prompts. Scores are computed using GPT-4o vision-language model. Our method achieves comparable performance to the base model, indicating no degradation on regular prompts.*

Models	PartiPrompts-simple
FLUX	93.46
SAP	93.06

Figure 1. *FLUX tends to generate realistic images by default. However, when given unrealistic prompts, it often produces cartoon-like samples. In contrast, our method, which gradually resolves such prompts through coherent proxy stages, consistently generates realistic and semantically aligned images.*

Failure Mode Analysis. We analyze failure cases of SAP and group them into three categories, summarized in Table 3. The majority of failures arise from incorrect proxy prompt assignment, where the selected proxy does not adequately preserve the intended structure or semantics of the target concept. In many such cases, manually correcting the proxy resolves the contradiction, as illustrated in Figure 2 (left), indicating that these failures stem from limitations

Table 3. *Breakdown of SAP failure cases by root cause.*

	Interval assignment	Proxy assignment	Model limitation
Failure cases	3.3%	63.3%	33.3%

Figure 2. *Failure cases of SAP. Left: a proxy assignment failure, where an unsuitable proxy leads to incorrect generation; manually correcting the proxy (right image in pair) resolves the contradiction. Right: failures due to inherent limitations of the underlying diffusion model, such as difficulty generating an empty pool or rendering unusual symbolic structures (e.g., a compass with incorrect directions).*

of the LLM rather than the proposed framework, and can potentially be mitigated. A second group of failures stems from inherent limitations of the underlying diffusion model, such as difficulty generating an empty pool or rendering unusual symbolic structures (e.g., compass directions), which persist even with correct prompt decomposition (Figure 2, right). Finally, only a small fraction of failures is attributed to suboptimal interval assignment, consistent with our findings that SAP is robust to moderate variations in timestep scheduling.

Computational Overhead. SAP introduces modest overhead due to a single LLM inference for prompt decomposition, performed once per prompt prior to diffusion, and a lightweight additional embedding pass. As this step occurs only once and is independent of the denoising process, it does not significantly impact overall generation time.

3. LLM Instruction for Prompt Decomposition

Tables 4 and 5 detail the full LLM instruction used for our method's decomposition, along with the corresponding in-context examples. In a single inference pass, our method detects contextual contradictions, generates proxy prompts, and assigns timestep intervals.

Table 6 presents examples of our LLM input prompts, along with the corresponding output explanations and the decomposition into proxy prompts and timestep intervals.

4. Provided Benchmarks

We describe the construction of *ContraBench* and *Whoops-Hard* in the main text. Here, we provide the full lists of prompts for these benchmarks in Table 7 and Table 8, respectively.

5. VLM Evaluation

We utilize GPT-4o to assess alignment between prompts and their generated images. The instruction prompt provided to the VLM is shown in Table 9.

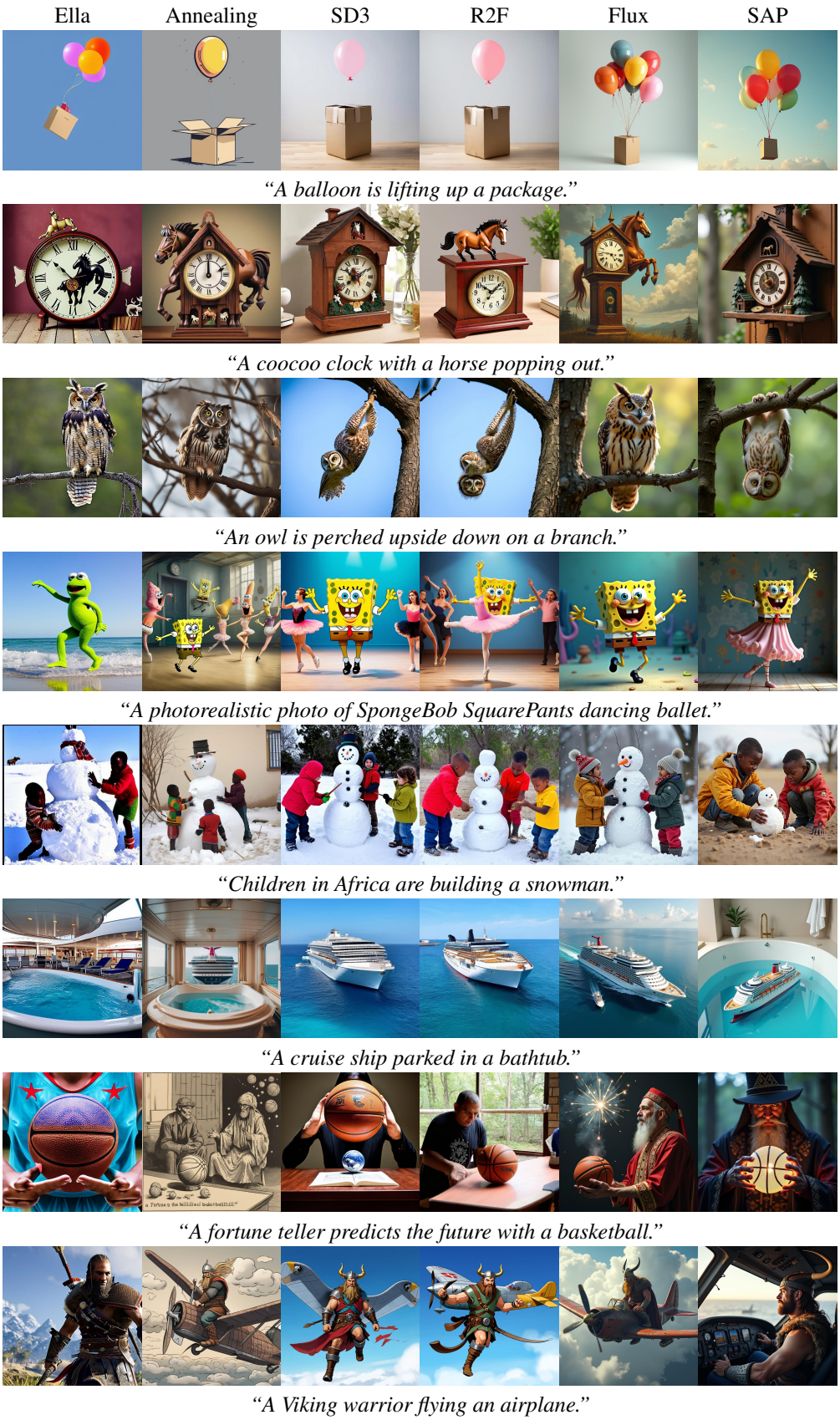

Figure 3. *Qualitative comparison. Our method consistently generates text-aligned images for contextually contradicting prompts.*

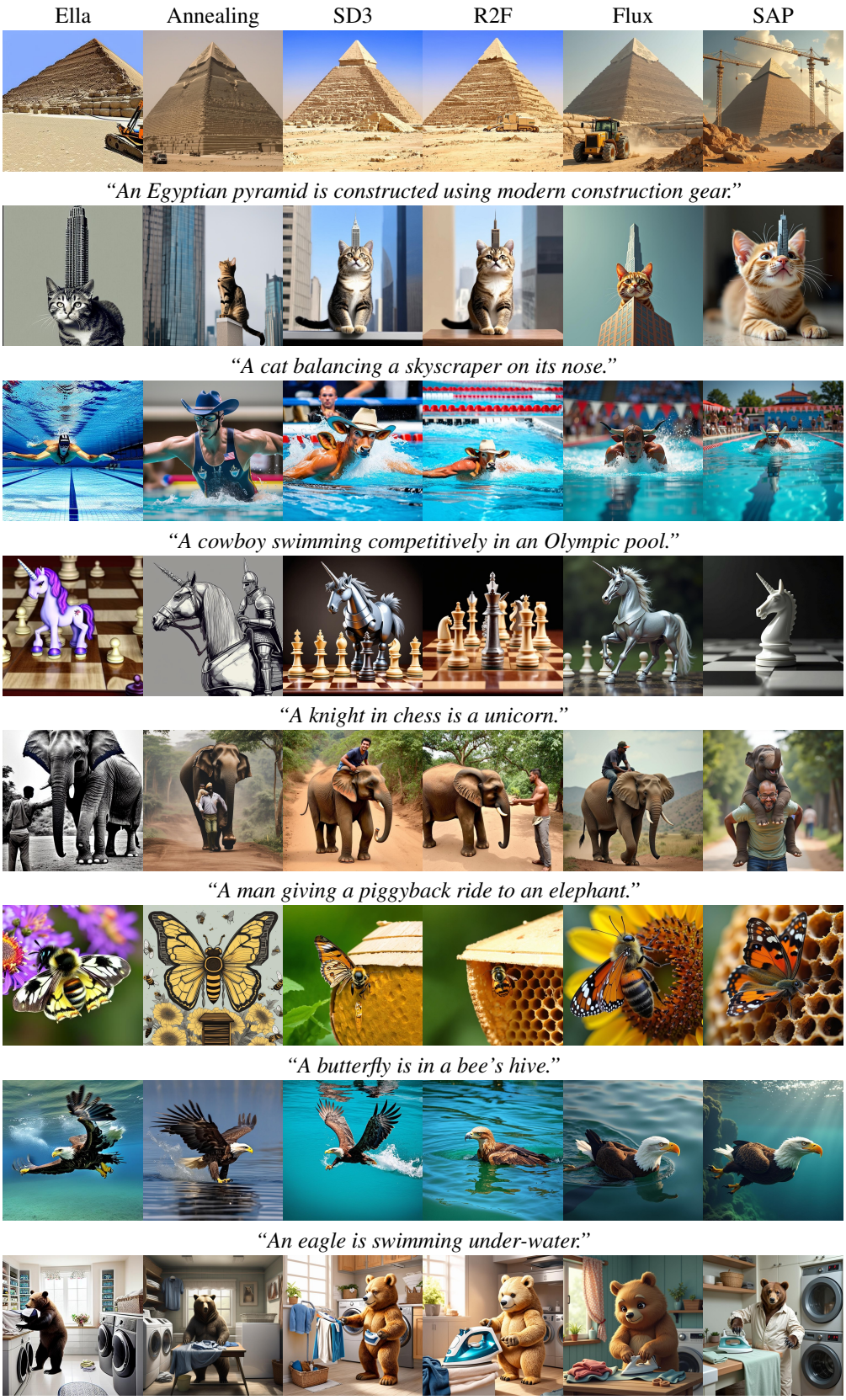

Figure 4. *Qualitative comparison. Our method consistently generates text-aligned images for contextually contradicting prompts.*

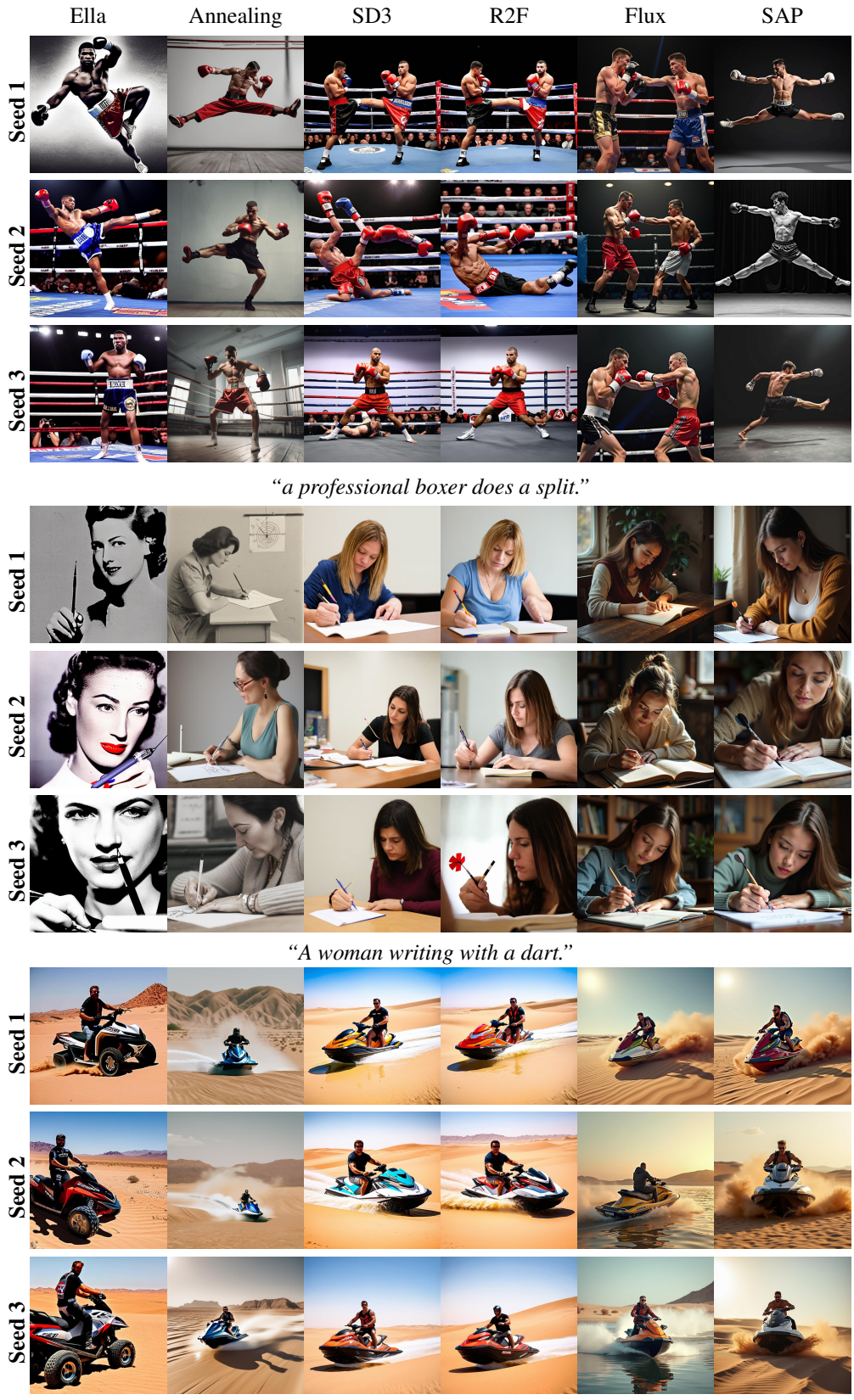

Figure 5. *Qualitative comparison across multiple seeds. Our method consistently generates text-aligned images for contextually contradicting prompts.*

Table 4. *Full LLM prompt instruction SAP, used to decompose prompts by denoising stages.*

<System Prompt>
You are an expert assistant in time step dependent prompt conditioning for diffusion models.
Your task is to decompose a complex or contextually contradictory prompt into up to **three** intermediate prompts that align with the model's denoising stages — from background layout to object identity to fine detail. Only introduce prompt transitions when needed.

Diffusion Semantics (Low → High Frequency Progression):
Steps 0–2: Scene layout and dominant color regions (e.g., sky, forest, sand tone)
Steps 3–6: Object shape, size, pose, and position
Steps 7–10: Object identity, material, and surface type (e.g., glass vs. rubber)
Steps 11–13+: Fine features and local details (e.g., tattoos, insects, facial detail)
Since denoising progresses from coarse to fine, it is crucial to stabilize large-scale visual structures (such as body shape, pose, and background) before introducing small or semantically charged elements (such as facial details, objects in hand, or surreal features).

Substitution Strategy:
1. Begin with high-level layout (background, geometry).
2. Use **placeholder concepts** if needed to stabilize layout before detailed insertions.
3. Substitutes must match in shape, size, and visual function.
4. Replace placeholders as soon as fidelity permits.
5. Do not maintain substitutions longer than needed.
6. If the prompt is visually coherent, return a **single prompt** with no decomposition.

Output Format:
a. Explanation: A short sentence explaining why decomposition is needed.
b. Final dictionary:

```
{
  "prompts_list": [
    "<prompt1>",
    "<prompt2>",
    "...",
    "<target prompt>"
  ],
  "switch_prompts_steps": [<step1>, <step2>, ...]
}
```

- The length of switch_ prompts_steps should be one less than prompts_list. - Do not include any text outside this structure.

Table 5. *In-context examples SAP, illustrating reasoning steps and prompt decomposition across denoising stages.*

Input: A polar bear in a desert
Output:
a. Explanation: A polar bear is common in snowy scenes, not deserts. Since no suitable object proxy exists, the prompt starts with the desert alone before introducing the unlikely animal.
b. Final dictionary:

```
{
  "prompts_list": [
    "A desert",
    "A polar bear in a desert"
  ],
  "switch_prompts_steps": [2]
}
```

Input: A fox in a nursery
Output:
a. Explanation: A fox is uncommon in indoor scenes. Starting with a dog, then a visually similar breed (Shiba Inu), provides a natural proxy before introducing the fox in a childlike setting.
b. Final dictionary:

```
{
  "prompts_list": [
    "A dog in a nursery",
    "A Shiba Inu dog in a nursery",
    "A fox in a baby room"
  ],
  "switch_prompts_steps": [4, 7]
}
```

Input: A shoes tree in the meadow
Output:
a. Explanation: Treating a “shoes tree” like an “apple tree” allows the model to build structure and object layout using familiar patterns. Introducing apples first grounds the scene in a biologically plausible layout before shifting to the surreal concept of shoes as fruit-like objects.
b. Final dictionary:

```
{
  "prompts_list": [
    "tree in the meadow",
    A tree full of apples, in the meadow",
    "A tree full of shoes in the meadow"
  ],
  "switch_prompts_steps": [4,8]
}
```

Input: A red sports car parked on a mountain road
Output:
a. Explanation: This is a realistic and visually coherent scene, so no decomposition is needed.
b. Final dictionary:

```
{
  "prompts_list": [
    "A red sports car parked on a mountain road"
  ],
  "switch_prompts_steps": []
}
```

Table 6. *Prompt decompositions of contradictory prompts using SAP.*

Original Prompt	Explanation	Prompt Sequence	Switch Step(s)
A grown man has a baby's pacifier in his mouth.	The grown man is established first for structure, while the pacifier, being a small but semantically significant detail, is added once the main identity is resolved.	1. A grown man with a small object in his mouth 2. A grown man has a baby's pacifier in his mouth	4
A dragon is blowing water.	Dragons are more commonly depicted blowing fire. A proxy of white smoke is visually similar to water mist in texture, stabilizing the emission process before resolving the surreal water emission.	1. A dragon blowing white smoke 2. A dragon blowing water	3
A pizza with grape toppings.	Pizza with traditional toppings stabilizes the geometry and color before introducing the visually similar yet unusual grape topping.	1. A pizza with pepperoni toppings 2. A pizza with grape toppings	3
A coin floats on the surface of the water.	Coins typically sink in water, not float. Starting with a leaf—an object that naturally floats—ensures that this behavior within the scene is handled correctly before introducing the coin.	1. A leaf floats on the surface of the water 2. A coin floats on the surface of the water	4
A cockatoo parrot swimming in the ocean.	Cockatoos are birds and naturally do not swim; starting with a simple bird on water stabilizes position and motion. Progressing to a duck, before introducing the cockatoo parrot, eases the transition into the final surreal visual.	1. A duck swimming in the ocean 2. A parrot swimming in the ocean 3. A cockatoo parrot swimming in the ocean	3, 6
Shrek is blue.	Shrek is a distinct character with a recognizable green color. Using a simple b̈lue ogreïnitially sets the stage for a color change before fully introducing Shrek to ensure visual coherence.	1. A blue ogre 2. Shrek is blue	3
A professional boxer does a split.	Professional boxers are typically shown in athletic stances related to fighting, not performing a split. Starting with a gymnast performing a split supports the action, introducing a boxer in similar attire balances identity shift without disrupting the pose.	1. A gymnast performing a split 2. A boxer performing a split 3. A professional boxer doing a split	3, 6

Table 7. *ContraBench. A curated benchmark of 40 contradictory prompts for evaluating text-to-image models.*

ID	Prompt	ID	Prompt
1	A professional boxer does a split	21	A mosquito pulling a royal carriage through Times Square
2	A bear performing a handstand in the park	22	A grandma is ice skating on the roof
3	A bodybuilder balancing on point shoes	23	A baseball player backswing a yellow ball with a golf club
4	A chicken is smiling	24	A house with a circular door
5	A cruise ship parked in a bathtub	25	A photorealistic image of a bear ironing clothes in a laundry room
6	A man giving a piggyback ride to an elephant	26	A pizza being used as an umbrella in the rain
7	A zebra climbing a tree	27	A cubist lion hiding in a photorealistic jungle
8	A coffee machine dispensing glitter	28	A cowboy swimming competitively in an Olympic pool
9	A vending machine in a human running posture	29	A realistic photo of an elephant wearing slippers
10	A ballerina aggressively flipping a table	30	A computer mouse eating a piece of cheese
11	A bathtub floating above a desert in a tornado	31	A horse taking a selfie with a smartphone
12	A monkey juggles tiny elephants	32	A sheep practicing yoga on a mat
13	A woman has a marine haircut	33	A snake eating a small golden guitar
14	A tower with two hands	34	A soccer field painted on a grain of rice
15	An archer is shooting flowers with a bow	35	A snake with feet
16	A muscular ferret in the woods	36	A woman brushing her teeth with a paintbrush
17	A barn built atop a skyscraper rooftop	37	A horse with a hump
18	A cat balancing a skyscraper on its nose	38	A hyperrealistic unicorn made of origami
19	A cow grazing on a city rooftop	39	A library printed on a butterfly's wings
20	A fireplace burning inside an igloo	40	A photorealistic photo of SpongeBob SquarePants dancing ballet

Table 8. *Whoops-Hard. A curated subset of 100 challenging prompts from the Whoops!* benchmark.

ID	Prompt	ID	Prompt
1	A bouquet of flowers is upside down in a vase	51	A Japanese tea ceremony uses coffee instead of tea
2	A man is welding without a mask	52	A wagon is being pushed from behind by two opposite facing horses
3	A man eats hamburgers in a baby chair	53	The Girl with a Pearl Earring wears a golden hoop earring
4	A turn right street sign with a left turn arrow	54	A chandelier is hanging low to the ground
5	Goldilocks sleeps with four bears	55	The portrait of the Mona Lisa depicts a stern male face
6	A cake wishes a happy 202nd birthday	56	A car with the steering wheel right in the middle of the dashboard
7	Children are unhappy at Disneyland	57	A pagoda sits in front of the Eiffel Tower
8	An orange carved as a Jack O'Lantern	58	A man without protection next to a swarm of bees
9	A pen is being sharpened in a pencil sharpener	59	A kiwi bird in a green bamboo forest
10	Steve Jobs demonstrating a Microsoft tablet	60	The Sphinx is decorated like a sarcophagus outside a Mayan temple
11	Shrek is blue	61	A butterfly is in a bee's hive
12	A MacBook with a pear logo on it	62	A rainbow colored tank
13	A woman hits an eight ball with a racket	63	Movie goers nibble on vegetables instead of popcorn
14	Vikings ride on public transportation	64	A grown man has a baby's pacifier in his mouth
15	A gift wrapped junked car	65	A full pepper shaker turned upside down with nothing coming out
16	A rainbow is filling the stormy sky at night	66	The Tiger King, Joe Exotic, poses with an adult saber-tooth tiger
17	John Lennon using a MacBook	67	A scale is balanced with one side full and the other empty
18	Michelangelo's David is covered by a fig leaf	68	A pizza box is full of sushi
19	Chuck Norris struggles to lift weights	69	A man wearing a dog recovery cone collar while staring at his dog
20	Paratroopers deploy out of hot air balloons	70	A woman's mirror reflection is wearing different clothes
21	A train on asphalt	71	A woman using an umbrella made of fishnet in the rain
22	Lionel Messi playing tennis	72	A field of sunflowers with pink petals
23	A man jumping into an empty swimming pool	73	An eagle swimming under water
24	An airplane inside a small car garage	74	A woman stands in front of a reversed reflection in a mirror
25	An upside down knife about to slice a tomato	75	Stars visible in the sky with a bright afternoon sun
26	Dirty dishes in a bathroom sink	76	A car with an upside down Mercedes-Benz logo
27	A roulette wheel used as a dart board	77	An owl perched upside down on a branch
28	A smartphone plugged into a typewriter	78	A man in a wheelchair ascends steps
29	A passenger plane parked in a parking lot	79	Bach using sound mixing equipment
30	Guests are laughing at a funeral	80	A steam train on a track twisted like a roller coaster
31	A cat chasing a dog down the street	81	Roman centurions fire a cannon
32	The Statue of Liberty is holding a sword	82	A crab with four claws
33	A Rubik's cube with ten purple squares	83	Elon Musk wearing a shirt with a Meta logo
34	A girl roller skating on an ice rink	84	A compass with North South South West points
35	A butterfly swimming under the ocean	85	A glass carafe upside down with contents not pouring
36	Lightning striking a shack on a sunny day	86	Princess Diana standing in front of her grown son, Prince Harry
37	The Cookie Monster is eating apples	87	A children's playground set in the color black
38	A man is given a purple blood transfusion	88	A mug of hot tea with a plastic straw
39	An unpeeled banana in a blender	89	A whole pomegranate inside a corked glass bottle
40	A square apple	90	Belle from Beauty and the Beast about to kiss the Frog Prince
41	A place setting has two knives	91	A person's feet facing opposite directions
42	A koala in an Asian landscape	92	A bowl of cereal in water
43	A mouse eats a snake	93	A boy playing frisbee with a porcelain disk
44	A field of carrots growing above ground	94	A chef prepares a painting
45	A pregnant woman eating raw salmon	95	A dragon blowing water
46	A tiger staring at zebras in the savanna	96	The lip of a pitcher on the same side as the handle
47	Albert Einstein driving a drag racing car	97	Greta Thunberg holding a disposable plastic cup
48	A soccer player about to kick a bowling ball	98	A fortune teller predicting the future with a basketball
49	An old man riding a unicycle	99	A balloon lifting up a package
50	A hockey player drives a golf ball down the ice	100	Bruce Lee in a yellow leotard and tutu practicing ballet

Table 9. *VLM instruction for evaluation. Used by GPT-4o to score semantic alignment of generated images.*

You are an assistant evaluating an image on how well it aligns with the meaning of a given text prompt.

The text prompt is: "`{prompt}`"

PROMPT ALIGNMENT (Semantic Fidelity)
Evaluate only the *meaning* conveyed by the image — ignore visual artifacts.
Focus on:

- Are the correct objects present and depicted in a way that clearly demonstrates their intended roles and actions from the prompt?
- Does the scene illustrate the intended situation or use-case in a concrete and functional way, rather than through symbolic, metaphorical, or hybrid representation?
- If the described usage or interaction is missing or unclear, alignment should be penalized.
- Focus strictly on the presence, roles, and relationships of the described elements — not on rendering quality.

Score from 1 to 5:

- 5: Fully conveys the prompt's meaning with correct elements
- 4: Mostly accurate — main elements are correct, with minor conceptual or contextual issues
- 3: Main subjects are present but important attributes or actions are missing or wrong
- 2: Some relevant components are present, but key elements or intent are significantly misrepresented
- 1: Does not reflect the prompt at all

Respond using this format:

```
### ALIGNMENT SCORE: <score>
### ALIGNMENT EXPLANATION: <explanation>
```