Test-time Prompt Refinement for Text-to-Image Models

Supplementary Material

1. Dataset and Evaluation Details

1.1. Benchmark Datasets

We use three benchmark datasets to assess compositional fidelity, prompt comprehension, and generalization:

1.1.1. GENEVAL: Compositional Accuracy

GENEVAL [14] consists of 553 prompts testing object presence, count, color accuracy, spatial positioning, and attribute binding. It provides structured evaluation for finegrained correctness in T2I models.

1.1.2. LLM-Grounded Diffusion Benchmark: Prompt Comprehension

Lian et al. [26] designed a benchmark to assess how well T2I models interpret prompts. Based on their given templates, we created 320 structured prompts, using 20 most common COCO objects, covering:

- Negation (A realistic photo of a scene without [object namel)
- Numerical reasoning (A realistic photo of a scene with [number] [object name])
- Attribute binding (A realistic photo of a scene with [modifier 1] [object name 1] and [modifier 2] [object name 2])
- Spatial reasoning (A realistic photo of a scene with [object name 1] on the [location] and [modifier 2] [object name2] on the [opposite location], where the location is chosen from left, right, top, and bottom.)

1.1.3. DrawBench: Generalization to Open-Ended Prompts

DrawBench [42] evaluates generative adaptability across ambiguous descriptions, numerical constraints, spatial relations, and rare words. It helps measures perceptual plausibility rather than strict correctness.

2. Qualitative Results

Figs. 1 and 2 showcase qualitative results of our Test-time Iterative Refinement (TIR) method on Drawbench prompts, using GPT-40 as the MLLM for prompt refinement and DALL-E 3 as the text-to-image (T2I) generator. Each row illustrates the initial prompt and image on the left, followed by three iterations of refinement. For each iteration, the left column presents the progressively refined prompts, while the right column shows the corresponding images generated from these refined prompts. We observe progressive improvements in visual-semantic alignment, demonstrating TIR's ability to iteratively correct misalignments and converge toward more faithful generations with respect to original user intent.



Figure 1. Qualitative results on Drawbench using DALL-E 3 as the T2I model and GPT-40 as the MLLM. Each row shows the prompt refinement trajectory, beginning with the initial prompt and generation on the left, followed by three rounds of GPT-40-guided refinements. The results demonstrate how TIR progressively enhances alignment between user intent and visual output.

A laptop on top of a teddy bear.

A banana on the left of an apple.

A pizza on the right of a suitcase.

A storefront with 'Diffusion' written on it.



A teddy bear with a laptop placed directly on top of its lapy placed directly on top of its lapy resulting the laptop is clearly visible and resting naturally or the bear's body without any blankets, pillows, or bed obscuring the view. The background should be simple and unobtrusive to emphasize the teddy bear and laptop.



A teddy bear with a laptop placed directly on top of its lap, ensuring the laptop is clearly visible and resting naturally on the bear's body. Remove all pillows, blankets, or bed elements from the scene, and place the teddy bear on a plain, flat surface. The background should remain simple and unobtrusive, emphaszing the teddy bear and the laptop.



A banana on the left of an apple, both placed on a wooden surface with soft, warm lighting in the background.

A teddy bear with a laptop placed directly on top of it, clearly showing the laptop resting on the bear's body, without blankets or other

objects obscuring the view



A banana positioned clearly to the left of an apple, both placed on a wooden surface, with soft, warm lighting in the background. Ensure the banana's placement is distinctly to the left of the apple and not overlapping or behind it.



A banana positioned clearly to the left of an apple, both placed on a wooden surface. Ensure the banana is upright and distinctly separated from the apple, with no overlapple. Use soft, warm lighting in the background to highlight both fruits.



A pizza placed entirely to the right of a suitcase, with the suitcase on the left and the pizza fully visible. Ensure the two objects are distinct and no overlapping, with clear separation between them.



A pizza placed entirely to the right of a suitcase, with the suitcase on the left and the pizza fully visible. Ensure the two objects are distinct, non-overlapping, and separated by a clear gap. The suitcase and pizza should not merge or blend into each other, maintaining clear boundaries.



A pizza placed entirely to the right of a suitcase, with the suitcase on the left and the pizza fully visible as a whole circular object. Ensure the two objects are distinct, nonoverlapping, and separated by a clear gap. The pizza should not appear cut off or divided, and both items should be clearly and fully represented.



A modern storefront with 'Diffusion' prominently written on it, featuring clean, minimalis design elements, large glass windows, and a stylish interior showcasing clothing and decor items. The storefront is located on a street corner with a contemporary urban vibe.



A modern storefront with 'Diffusion' prominently written on it, featuring clean, minimalist design elements, large glass windows, and a stylish interior showcasing clothing and decor tiems. The storefront is located on a street corner with a contemporary urban vibe, emphasizing a sleek and inviting aesthetic.



A modern storefront with Olfkusion' prominently written on it, featuring clean, minimalist design elements, large glass windows, and a stylish interior showcasing clothing and decor items. The storefront is located on a street corner with a contemporary urban vibs contemporary urban vibs contemporary urban vibs inviting aesthetic, with clear branding and an airy, open atmosphere.



A storefront with 'Google Research Pizza Cafe'



A storefront with 'Google Research Pizza Cafe' prominently written on the signboard. Ensure the text 'Google Research' is clearly visible and in a professional font above Pizza Cafe'. The design should emphasize a modern, research-oriented theme, with no unnecessary additions like 'Google' or extr neon elements. The storefront should have a clean, sleek appearance with a foous on the branding 'Google' or sext neon Pizza Cafe'.



A storefront with 'Google Research Pizza Cafe prominently written on the signboard Ensure the text 'Google Research' is clearly visible and written in a professional, modern front above 'Pizza Cafe'. Remove the misspelling Research' and the singular through through the singular through the singular through the singular thr



A storefront with 'Google Research Pizza Cafe' prominently written on the signboard. Ensure the text with the signboard. Ensure the text wishle and written in a professional, modern font above Pizza Cafe'. Correct the current misspelling of Research'. The storefront should avoid excessive neen lighting and entitle storefront should avoid excessive neen lighting and entitle storefront should avoid excessive neen lighting and entitle storefront should avoid proposed to store the storefront should research should avoid excessive neen lighting and minimalist aesthetic that slights with Google's branding, incorporate Google's color palette subthy into the design, such as hints of blue, red, yellow, and green, to reinforce the brand identity. The interior should remain visible, showcashing a modern and showcashing a modern and store should be showned to show the showcashing and modern and showcashing and modern and yellow the showned and show the showned and shown



New York Skyline with 'Hello World' written with fireworks on the sky.



New York Skyline with 'Hello World' written with colorful fireworks in the sky, showcasing the city's iconic landmarks at night with vibrant lighting and a festive atmosphere.



New York Skyline with 'Helio World' clearly written with colorful and legible freeworks in the sky, ensuring accurate spelling and proper alignment of the text. Showcase the city's iconic landmarks a ringit with vibrant lighting, a festive atmosphere, and no additional elements like planets or unrelated objects in the sky.



New York Skyline with 'Hello World' clearly written with colorful and legible freworks in the sky, ensuring accurate spelling and proper alignment of the text, Showcase the cary's iconic landmarks at night with written of the text, Showcase the arsive when the skyling and a resive additional elements such as the planet or unrelated objects in the sky, focusing solely on the fireworks and skyline.



New York Skyline with 'Google Research Pizza Cafe' written with fireworks on the sky.



New York Skyline at night with 'Google Research Pizza Cafe' written in vibrant fireworks across the sky, ensuring the text is clearly legible and centered above the skyline, with the city's iconic landmarks like the Empire State Building prominently visible.



New York Skyline at night with 'Google Research Pizza Cafe' written in large, wibrant, and colorful fireworks letters across the sky, ensuring the text is clearly legible, evenly spaced, and centered above the skyline. Include prominent city landmarks like the Empire State Building and ensure the Empire State Building and ensure the revorks are dynamic and varied in shape and color to emphasize the celebratory atmosphere.



New York Skyline at night with Google Research Pizza Cafe' written in large, wibrant, and colorful fireworks letters across the sky. Ensure the text is clearly legible, evenly spaced, explained to the sky. Ensure the text is clearly legible, evenly spaced, explained to the sky. Ensurement Strondway Cafe' text with 'Google Research Pizza Cafe' and make the letters appear dynamically formed by fireworks. Include prominent city landmarks like the Empire State Bullding and ensure the fireworks are valered in shape, and color to enhance the celebratory almosphere.



Figure 2. Qualitative results on Drawbench using DALL-E 3 as the T2I model and GPT-40 as the MLLM. Each row shows the prompt refinement trajectory, beginning with the initial prompt and generation on the left, followed by three rounds of GPT-40-guided refinements. The results demonstrate how TIR progressively enhances alignment between user intent and visual output.