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 fine-grained 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 name])
- 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-4o 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-4o 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-4o-guided refinements. The results demonstrate how TIR progressively enhances alignment between user intent and visual output.

Figure 2. **Qualitative results on Drawbench using DALL-E 3 as the T2I model and GPT-4o 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-4o-guided refinements. The results demonstrate how TIR progressively enhances alignment between user intent and visual output.