NO CONCEPT LEFT BEHIND: TEST-TIME OPTIMIZATION FOR COMPOSITIONAL TEXT-TO-IMAGE GENERATION

Mohammad Hossein Sameti[*], *Amir M. Mansourian*[*], *Arash Marioriyad, Soheil Fadaee Oshyani,*
Mohammad Hossein Rohban, Mahdieh Soleymani Baghshah

Sharif University of Technology, Tehran, Iran

ABSTRACT

Despite recent advances in text-to-image (T2I) models, they often fail to faithfully render all elements of complex prompts, frequently omitting or misrepresenting specific objects and attributes. Test-time optimization has emerged as a promising approach to address this limitation by refining generation without the need for retraining. In this paper, we propose a fine-grained test-time optimization framework that enhances compositional faithfulness in T2I generation. Unlike most of prior approaches that rely solely on a global image–text similarity score, our method decomposes the input prompt into semantic concepts and evaluates alignment at both the global and concept levels. A fine-grained variant of CLIP is used to compute concept-level correspondence, producing detailed feedback on missing or inaccurate concepts. This feedback is fed into an iterative prompt refinement loop, enabling the large language model to propose improved prompts. Experiments on DrawBench and CompBench prompts demonstrate that our method significantly improves concept coverage and human-judged faithfulness over both standard test-time optimization and the base T2I model. Code is available at: https://github.com/AmirMansurian/NoConceptLeftBehind

***Index Terms*—** Text-to-Image Generation, Test-time Optimization, Compositionality

1. INTRODUCTION

Text-to-image (T2I) generation has seen rapid progress with models such as DALL·E [1], Stable Diffusion [2], and FLUX [3, 4], which can synthesize realistic images from natural language descriptions. Despite impressive zero-shot capabilities, these models often struggle with compositional faithfulness: faithfully representing all objects, attributes, and relations described in the prompt [5, 6].

A promising direction to mitigate such failures is test-time optimization. Instead of retraining large models, one can refine the input prompt or generation process iteratively, guided by a scoring function. Recent frameworks demonstrate that large language models (LLMs) can be used at inference time to generate and refine candidate prompts, while an external scorer evaluates image–text alignment [7, 8, 9, 10]. This loop can progressively improve results without additional training data. One such framework is MILS [10] (Multimodal Iterative LLM Solver), which leverages an LLM to propose prompts and a global CLIP-based [11] scorer to evaluate generated images. While MILS demonstrates the potential of test-time iterative optimization, it is limited by its reliance on a single, coarse similarity score. A global CLIP similarity may assign a high score even if some prompt concepts are missing, since it measures overall match rather than concept-level fidelity.

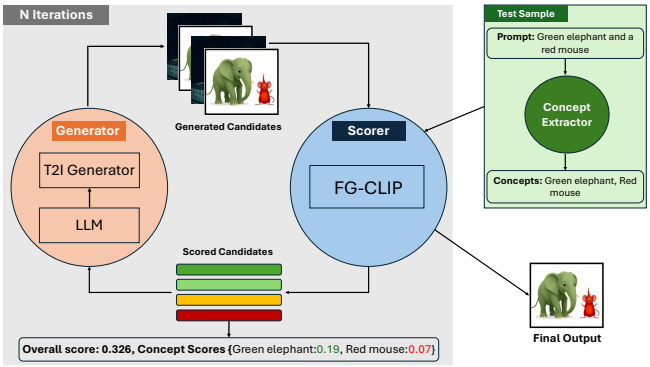

Fig. 1. **Overall diagram of the proposed framework.** Input prompt is first processed by the **Concept Extractor** module, which extracts key concepts from the initial prompt. Then, in each step, the **Generator** module uses an LLM to rewrite diverse prompts based on both overall and concept-level scores, and a T2I model generates candidate images. The **Scorer** module evaluates these candidates using a fine-grained variant of CLIP, and the scores are fed back to the generator. After a certain number of iterations, the final image is produced. Concept-level scores help the LLM address missing concepts from previous steps and rewrite improved prompts.

Contributions. We propose a fine-grained test-time optimization framework for text-to-image generation that improves compositional fidelity (Figure 1). Our approach extends iterative prompt refinement by introducing concept extraction and a fine-grained CLIP scorer that evaluates both

[*] Equal contribution.

global prompt–image alignment and per-concept correspondence. By feeding these detailed scores back to the LLM, the system is guided to explicitly recover missing objects, attributes, and relations, resulting in images that more faithfully capture all aspects of the input prompt while maintaining overall semantic alignment.

2. RELATED WORK

T2I generation has rapidly advanced from early GAN-based approaches to transformer-based and diffusion models that achieve striking visual quality [12, 13, 1, 14, 2]. Recent large-scale diffusion models, including DALL·E 2 [1], Stable Diffusion [2], and FLUX [3] demonstrate impressive zero-shot generation capabilities across diverse prompts. Despite these advances, even state-of-the-art models remain prone to compositional errors, often omitting or misrepresenting specific objects, attributes, or spatial relations when prompts become complex or multi-object in nature.

One promising direction for addressing this challenge is test-time optimization, which adapts or refines model behavior during inference without additional training, often improving robustness and alignment with task-specific requirements. Several recent studies have demonstrated that prompt optimization at test time can significantly enhance performance. For instance, [7] iteratively rewrites prompts by obtaining feedback from a visual question answering model, another variant refines prompts using a multimodal large language model [9], and MILS [10] introduces an iterative framework driven by CLIP-based feedback. To reduce the inference-time burden of such methods, [8] also proposes a fast, single-iteration prompt alignment.

3. METHOD

3.1. Overview

Given an input prompt P, we first extract its key semantic concepts (e.g., objects, attributes, relations), and then evaluate generated images with both a global similarity score and per-concept scores computed via a fine-grained CLIP model. This feedback is used in an iterative loop to refine candidate prompts proposed by an LLM, leading to images that more faithfully capture all prompt elements. Figure 1 illustrates the overall diagram of the proposed method.

3.2. Concept Extraction

Let P be the input prompt, decomposed into a set of k concepts: $\mathcal{C}(P) = \{c_1, c_2, \ldots, c_k\}$. Concepts include objects, attributes, and relations. We obtain $\mathcal{C}(P)$ using a syntactic parser or an LLM-based semantic extractor.

3.3. Fine-Grained Scoring

Given the input prompt P and the generated image I from the T2I model, we compute two types of similarity scores:

$$S_{\text{global}}(I, P) = \text{CLIP}(I, P), \tag{1}$$

$$s_i(I, c_i) = \text{CLIP}(I, c_i), \quad i = 1, \ldots, k, \tag{2}$$

where $\text{CLIP}(\cdot, \cdot)$ denotes cosine similarity in the joint embedding space of a fine-grained CLIP variant. The global score measures alignment with the entire prompt, while s_i evaluates the presence of each concept individually.

3.4. Optimization Formulation

The standard MILS [10] framework optimizes only the global score:

$$\max_I \; S_{\text{global}}(I, P). \tag{3}$$

In contrast, we formulate the problem as a multi-objective optimization:

$$\max_I \; S_{\text{global}}(I, P) + \frac{1}{k} \sum_{i=1}^{k} s_i(I, c_i), \tag{4}$$

where this objective explicitly encourages generated images to satisfy all extracted concepts.

3.5. Iterative Refinement

At each iteration, the LLM proposes a batch of candidate prompts. For each candidate $\tilde{P}$, images are generated and scored using the above objective. The top-ranked candidates, along with detailed per-concept scores, are fed back into the LLM to guide subsequent generations. This loop continues until convergence or a fixed number of iterations.

3.6. Signal Processing Viewpoint

Our method can be interpreted through the lens of signal processing. The prompt P is analogous to a composite signal with semantic "frequency components" $\{c_i\}$. A global similarity score reflects total signal energy, which may remain high even if some bands are missing. By projecting the generated image I onto each concept c_i via $s_i(I, c_i)$, we perform a filter-bank–like decomposition. Iteratively refining prompts with these projections resembles adaptive equalization, ensuring all components are preserved. This perspective grounds our method: it enforces both global alignment and the faithful preservation of individual concepts, akin to maintaining overall energy and spectral detail in classical signal reconstruction.

Table 1. Quantitative results comparison on T2I CompBench and DrawBench prompts.

Method	T2I CompBench				DrawBench			
	VQA	CLIP (L)	Captioning (L)	GPT4-o	VQA	CLIP (L)	Captioning (L)	GPT4-o
FLUX	0.865	0.272	0.687	0.717	0.620	0.279	0.645	0.719
MILS	0.925	0.287	0.694	0.744	0.665	0.299	0.671	0.765
Our	**0.955**	**0.295**	**0.701**	**0.810**	**0.715**	**0.304**	**0.677**	**0.827**

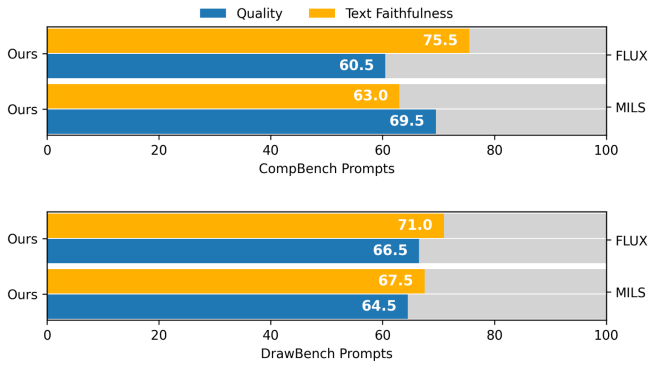

Fig. 2. Win rate comparison judged by human evaluation.

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

We evaluate our method using two widely adopted benchmarks. First, DrawBench [14], a set of 200 prompts designed to test text-image alignment, across 11 categories. Second, a curated subset of T2I-CompBench [15], where we select 200 prompts distributed across 8 compositional categories.

For evaluation, we employ both human and automated metrics. Human evaluation follows a win-rate protocol against the baseline under clear guidelines to reduce subjectivity, with three evaluators whose judgments are aggregated by majority vote. Automated evaluation consists of several complementary metrics: Visual Question Answering (VQA) using BLIP [16], CLIP Score [17] measuring text–image similarity in the CLIP embedding space, and Captioning Score, which computes the similarity between the BLIP-generated caption and the original prompt in the CLIP text embedding space. We further introduce a GPT Score, where GPT-4o rates alignment between prompt and image on a $[0, 1]$ scale, using category-specific prompts (e.g., attributes, spatial relations, numeracy) to capture different types of compositional alignment.

4.2. Implementation Details

We extend the MILS [10] codebase by integrating several key components. As the scorer, we employ FG-CLIP [18], a fine-grained variant of CLIP trained for region-to-word alignment. For text-to-image generation, we adopt FLUX.1 [schnell] [3], and for re-writing prompts and extracting concepts from the input, we utilize LLaMA-3.1-8B-Instruct [19].

During optimization, 50 new prompts are generated at each iteration, while the top 20 prompts from the previous step are retained. The process runs for a total of 10 iterations.

4.3. Quantitative and Qualitative Results

Table 1 presents the evaluation metrics for the base generator method (FLUX), the MILS baseline, and our proposed method on two datasets. As shown in the table, our method consistently outperforms the baselines across all metrics, demonstrating the effectiveness of incorporating fine-grained views to enhance compositional faithfulness in T2I generation. In particular, our method yields significant improvements over the FLUX baseline and achieves a clear margin of superiority over the MILS framework across nearly all metrics.

Furthermore, Figure 2 illustrates the win rate of our method compared with FLUX and MILS in terms of both image quality and text faithfulness. The results confirm that images generated by our method are more frequently preferred in human evaluations.

In addition, Figure 3 provides a qualitative comparison between our method and the baselines. It can be observed that the proposed method enhances the quality of generated images across various categories, including counting, spatial relationships, text rendering, and conflicting prompts.

4.4. Ablation Study

Figure 4 presents the results of our method across different iterations and evaluation metrics on both datasets. It can be observed that even in the first iteration, our method achieves strong performance compared to the FLUX baseline and almost consistently outperforms MILS at every iteration. As noted earlier, test-time optimization methods typically introduce additional computational overhead; however, our method surpasses the baselines even at the first step. This flexibility allows practitioners to obtain strong results with minimal overhead, while further improvements can be achieved by increasing the number of iterations if additional computation is acceptable.

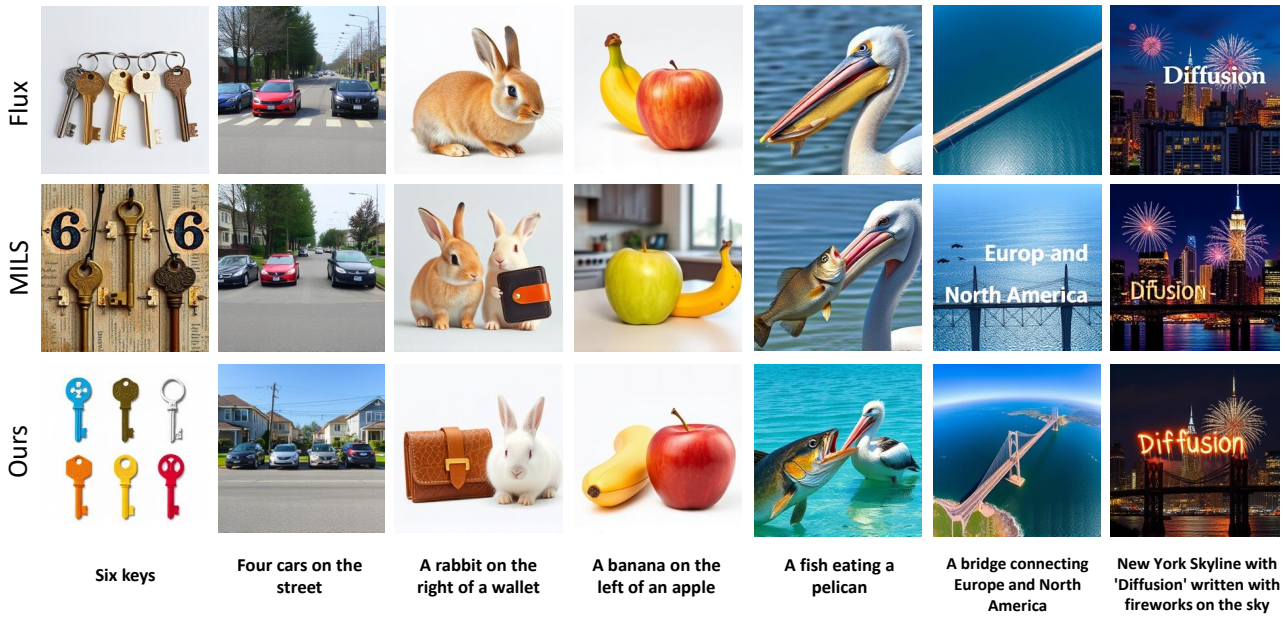

Fig. 3. Qualitative comparison with FLUX and MILS baselines on sample prompts from DrawBench and CompBench across different categories.

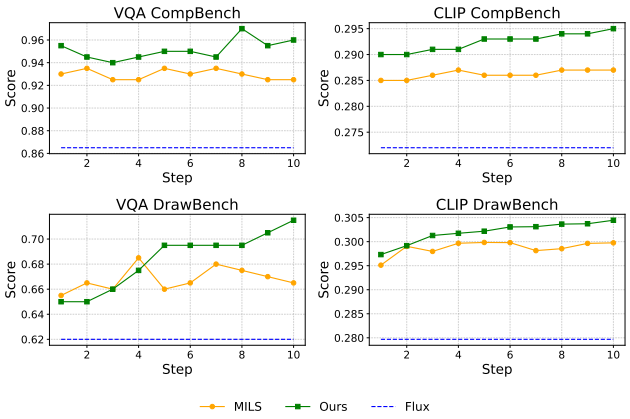

Fig. 4. Per-step comparison of our method with Flux and MILS baselines across two metrics on the DrawBench and CompBench prompts.

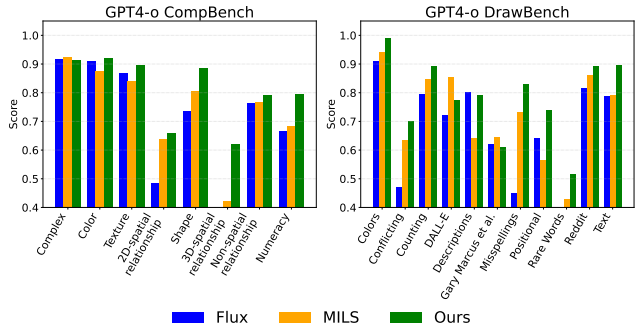

Fig. 5. Comparison of GPT4-o scores for the proposed method against FLUX and MILS baselines across different prompt categories on DrawBench and CompBench.

In addition, Figure 5 presents an ablation study of our method's performance across different prompt categories in comparison to the FLUX and MILS baselines. The results show that our method achieves performance that is consistently better than or on par with MILS across all categories. Notably, it performs marginally better in challenging categories such as counting, spatial relationships, and color prompts, which is consistent with the qualitative comparisons presented earlier in Figure 3. Thanks to the fine-grained view of our method, compositional improvements are achieved throughout the process, leading to marginally better results in certain categories.

5. CONCLUSION

We presented a fine-grained test-time optimization method for text-to-image generation. By decomposing prompts into concepts and scoring them with a fine-grained CLIP model, our approach supplies detailed feedback to an iterative refinement loop. This yields images that more faithfully capture all aspects of the input prompt, outperforming both MILS and raw T2I generation baselines. Our method highlights the promise of concept-aware test-time optimization for improving the compositionality of generative models.

6. REFERENCES

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, pp. 3, 2022.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[3] Black Forest Labs, "Flux," `https://github.com/black-forest-labs/flux`, 2024.

[4] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith, "Flux.1 kontext: Flow matching for in-context image generation and editing in latent space," 2025.

[5] Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua, "Silmm: Self-improving large multimodal models for compositional text-to-image generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18497–18508.

[6] Yixin Wan and Kai-Wei Chang, "Compalign: Improving compositional text-to-image generation with a complex benchmark and fine-grained feedback," *arXiv preprint arXiv:2505.11178*, 2025.

[7] Jaskirat Singh and Liang Zheng, "Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 70799–70811, 2023.

[8] Khalil Mrini, Hanlin Lu, Linjie Yang, Weilin Huang, and Heng Wang, "Fast prompt alignment for text-to-image generation," *arXiv preprint arXiv:2412.08639*, 2024.

[9] Mohammad Abdul Hafeez Khan, Yash Jain, Siddhartha Bhattacharyya, and Vibhav Vineet, "Test-time prompt refinement for text-to-image models," *arXiv preprint arXiv:2507.22076*, 2025.

[10] Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar, "Llms can see and hear without any training," *arXiv preprint arXiv:2501.18096*, 2025.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[12] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[13] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5802–5810.

[14] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[15] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78723–78747, 2023.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*. 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, PMLR.

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021.

[18] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin, "Fg-clip: Fine-grained visual and textual alignment," *arXiv preprint arXiv:2505.05071*, 2025.

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al., "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.