

# T2I-Copilot: A Training-Free Multi-Agent Text-to-Image System for Enhanced Prompt Interpretation and Interactive Generation

Chieh-Yun Chen, Min Shi, Gong Zhang, Humphrey Shi

SHI Labs @ Georgia Tech

[github.com/SHI-Labs/T2I-Copilot](https://github.com/SHI-Labs/T2I-Copilot)

This supplement includes pseudocode (Supp. A), the effect of including more models (Supp. B), an ablation study of MLLMs (Supp. C), cost comparison (Supp. D) and more qualitative results (Supp. E), covering ambiguous terms, single-turn results, and multi-turn results in both automatic mode and human-in-the-loop settings.

## A. Pseudocode for T2I-Copilot

---

### Algorithm 1 : T2I-Copilot Multi-Agent System

---

**Input:** User input prompt  $P$ , Optional reference image  $I_{ref}$ , Creativity level  $C_{level}$ , Human-in-the-loop flag  $H_l$ , Improvement suggestions  $S_{imp}$

```
1:  $R_A \leftarrow A_{in}(P, I_{ref}, C_{level}, H_l)$   $\triangleright$  Generate Analysis Report
2:  $I_{gen}^n \leftarrow A_{gen}(R_A, P, I_{ref})$   $\triangleright$  Generate Initial Image
3:  $score, S_{imp} \leftarrow A_{eval}(I_{gen}^n, P, R_A)$   $\triangleright$  Evaluate Image Quality
4:  $n \leftarrow 0$ 
5:  $U_f \leftarrow None$   $\triangleright$  Initialize user feedback variable
6: while  $score < THRESHOLD$  and  $n < MAX\_regen\_count$  do
7:   if  $H_l$  then
8:      $U_f \leftarrow$  Get user feedback on  $I_{gen}^n$ 
9:   end if
10:   $I_{gen}^n \leftarrow A_{gen}(R_A, P, I_{ref}, S_{imp}, U_f)$   $\triangleright$  Regenerate Image
11:   $score, S_{imp} \leftarrow A_{eval}(I_{gen}^n, P, R_A)$   $\triangleright$  Re-evaluate Image Quality
12:   $n \leftarrow n + 1$ 
13: end while
return  $I_{gen}^n$   $\triangleright$  Final Output Image
```

---

Our system includes modular error handlers and fallbacks. Failures fall into two categories, both handled automatically. First, for region extraction errors, RES segments unwanted objects or MLLM generates masks from bounding boxes; if both fail, MLLM infers boxes from prompt-image context, and RES is retried with full prompts. Only after all options fail is an error raised in  $A_{gen}$ , triggering fallback to the previous image or notifying the user (0.1% of cases). Second, for format extraction errors, malformed outputs trigger an automatic MLLM retry, which typically succeeds (0.3% for GPT-4o-mini, 1% for Qwen2.5-VL-3B). These mechanisms localize errors and prevent cascading failures.

## B. The effect of including more models

Our experiments show that the effect varies depending on specific conditions. We initially incorporated five models for selection, including a position-aware T2I model, RAG-Diffusion [3], a reference-based IP-Adapter, and a reference-based style transfer model. However, we found that the last two models were rarely selected. Moreover, using RAG-Diffusion led to a 3.43% decrease in performance on VQAScore [6] on GenAI-Bench [5], and the performance of the position-aware model was inconsistent. This inconsistency stemmed from the dependency on LLMs for position separation and RAG-Diffusion’s effectiveness in following the designed positional relationships. Instead, we found that simple reprompting in a prompt-guided T2I model could achieve similar positional control without the added complexity.

Similarly, integrating IP-Adapter did not significantly improve adherence to reference images for reference-based generation, leading to a slight performance drop of 0.39%. Instead, we could directly use a reference-based editing model to incorporate this functionality more effectively.

Furthermore, including more models requires careful system prompt design for model selection. Without proper prompt tuning, many tools remain unused. For instance, GenArtist [10] includes 10 models for T2I generation and 8 for editing. The default super-resolution tool is excluded from the selection process because it is directly applied to every generated sample rather than being chosen dynamically. As a result, among the remaining models, only 3 generation models and 2 editing models were selected when generating 1,800 images in DrawBench [8] and GenAI-Bench [5], leaving 6 generation tools and 6 editing tools unused. This highlights the need to evaluate whether adding more models meaningfully contributes to performance improvements.

Method (MLLM backbone)	Performance (VQAScore)			Inference time (s)				Multi-turn Turns	Cost/image ( $10^{-3}$ USD)		
	Basic	Advanced	Overall	MLLM Latency	Generator	Editor	End-to-End		LLM	T2I	All
FLUX1.1-pro	0.884	0.666	0.766	-	-	-	3.38	1.0	-	40.0 <sup>API</sup>	40.0
GenArtist (GPT-4o-mini)	0.693	0.504	0.588	5.75	5.7	3.7	56.96	3.3	3.1 <sup>API</sup>	1.6	4.7
Ours (GPT-4o-mini)	0.892	0.747	0.813	23.7	17.4	5.9	45.72	1.5	5.0 <sup>API</sup>	1.6	6.6
Ours (Mistral Small 3.1-24B)	0.893	0.761	0.821	44.0	17.6	5.7	64.74	1.5	7.3	1.5	8.9
Ours (Qwen2.5-VL-7B)	0.893	0.743	0.811	26.9	17.9	6.2	57.24	1.6	2.2	1.7	4.0
Ours (Qwen2.5-VL-3B)	0.873	0.695	0.777	12.8	18.1	6.3	36.02	1.5	1.1	1.6	2.6

Table A. MLLM ablation with average performance and cost on GenAI-Bench [16]. Cost refers to GPU expenses unless otherwise noted.

Prompt: A **Mustang** galloping across a field, with a dog chasing joyfully behind.

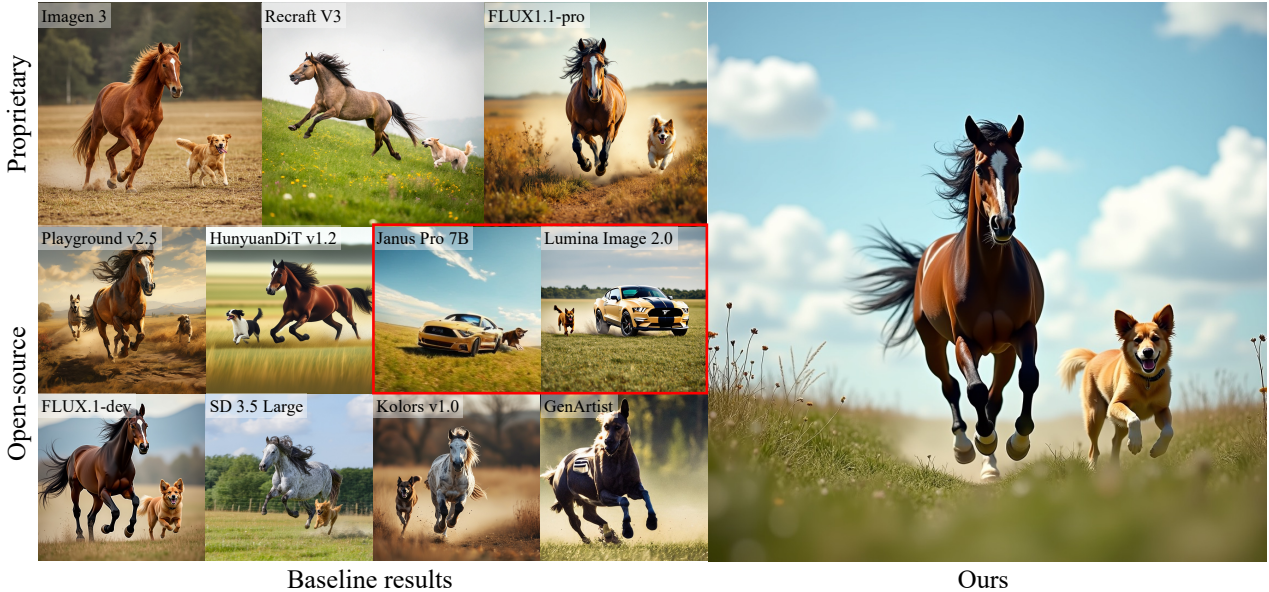


Figure 1. **Ambiguities sample.** The prompt “A Mustang galloping across a field, with a dog chasing joyfully behind” is ambiguous; “Mustang” could mean a car or a horse. While Janus Pro 7B [2] and Lumina Image 2.0 [9] depict a Ford Mustang, others show a horse. Our Input Interpreter Agent resolves this by recognizing that “galloping” applies to horses, ensuring correct subject interpretation.

### C. Ablation study of MLLM backbones

We evaluate open-sourced Mistral Small 3.1 24B [7], Qwen2.5-VL 7B and 3B [1] on L40S GPUs, with the 24B model using two GPUs and the others using one. Table A shows that performance is similar across model sizes: compared to GPT-4o-mini, 7B model scores  $-0.2\%$  VQAScore at  $-40\%$  cost.

### D. Cost comparison

In Table A, beyond the direct spending per image: \$0.005 for our method (using GPT-4o-mini) vs. \$0.04 for FLUX1.1-pro [4], we also account for self-hosted hardware costs. Specifically, using an L40S GPU priced at \$11,250 and depreciated over five years with 144 hours of weekly usage results in a rate of \$0.30 per GPU-hour.

Generating an average of 1,600 images, our method’s  $A_{gen}$  takes 19.72s/image, translating to \$0.0016/image. This amounts to only 16.59% of FLUX1.1-pro’s cost, while achieving a +6% improvement in VQAScore [6]. Compared to GenArtist [10], our method incurs 1.41x higher cost but delivers a +38% gain in VQAScore [6].

### E. More qualitative results

#### E.1. Ambiguous term

Ambiguous terms in text-to-image prompts can lead to unintended or inconsistent image generation. When a term has multiple possible interpretations, different models may generate vastly different images, reflecting the ambiguity inherent in natural language. Take Fig. 1 as an example.



Prompt: A cup set to the right of a newspaper.



Figure 2. **Qualitative result in single-turn:** Demonstrate generation performance on positional relationship of two objects.

Prompt: A Cardinal flying towards a bird feeder held by a person.



Figure 3. **Qualitative result in single-turn:** Demonstrate generation performance on action relationship of a bird and a human with given object (bird feeder).

E.2. More results in single-turn: Figs. 2, 3, 4, 5, 6, 7.

E.3. More results in multi-turn

E.3.1. Automatic: Figs. 8 and 9.

E.3.2. Human-in-the-loop: Figs. 10 and 11.

F. User study website screenshot: Fig. 12.



Prompt: A row of houses with chimneys, but **no** smoke coming out.



Figure 4. **Qualitative result in single-turn:** Demonstrate generation performance on logical negation of excluding smoke in the image.

Prompt: A glass with **no** water, only ice melting.

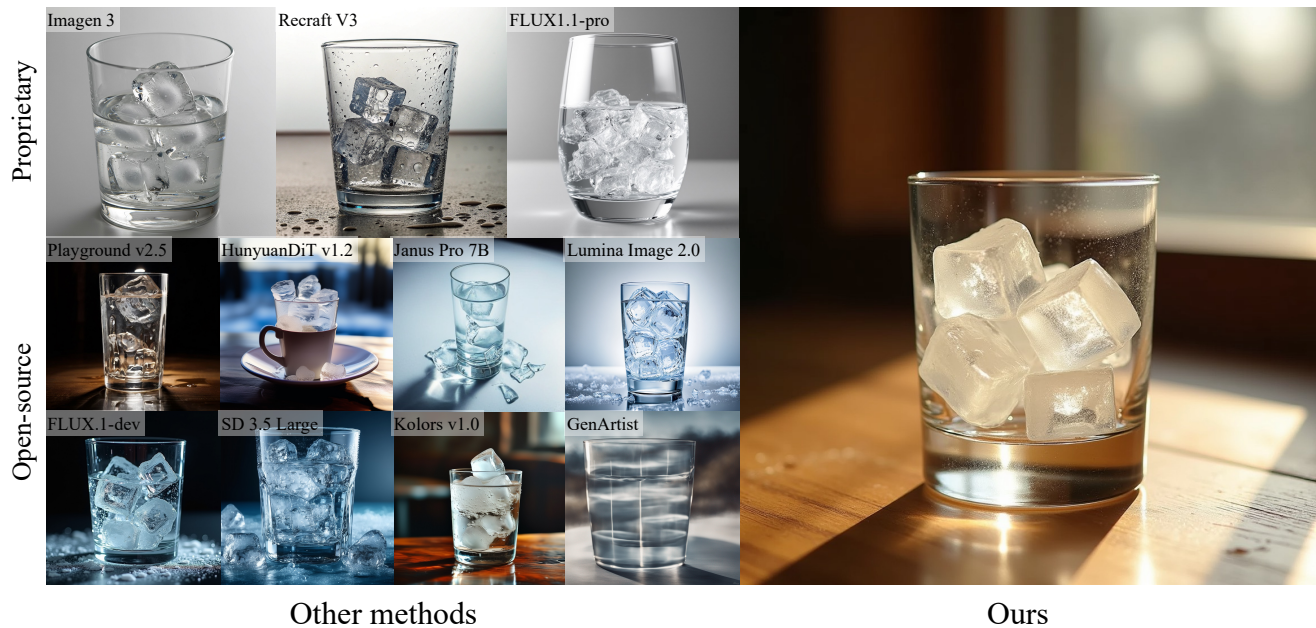


Figure 5. **Qualitative result in single-turn:** Demonstrate generation performance on logical negation of excluding water in the image.



Prompt: A person teaching another person how to ride a bicycle on a quiet street.



Figure 6. **Qualitative result in single-turn:** Demonstrate generation performance on action relationship of two persons in the image.

Prompt: A pirate ship sailing through the stars, 'Celestial Seas' written on the stern.



Figure 7. **Qualitative result in single-turn:** Demonstrate generation performance on including specific text in the specific region.

Prompt: Rainbow coloured penguin.

Improvement Suggestion: Incorporate distinct rainbow colors on the penguin's feathers to align with the prompt.



Ours ( $n=0$ )  
(Model: FLUX.1-dev)

Ours ( $n=1$ )  
(Model: Editing)

Figure 8. **Qualitative result in multi-turn:** Demonstrate enhancement performance on providing improvement suggestion and successfully modifying a specific region of the image automatically.

Prompt: A sphere made of kitchen tile.

Improvement Suggestion: Enhance the texture to more accurately reflect that of kitchen tiles



Ours ( $n=0$ )  
(Model: FLUX.1-dev)

Ours ( $n=1$ )  
(Model: FLUX.1-dev)

Figure 9. **Qualitative result in multi-turn:** Demonstrate enhancement performance on providing improvement suggestion and successfully modifying the texture of the image automatically.



Prompt: A gigantic dog that is taller than the tree next to it.

User Feedback: Enhance the size of the dog so that it appears taller than the tree, emphasizing its size contrast.



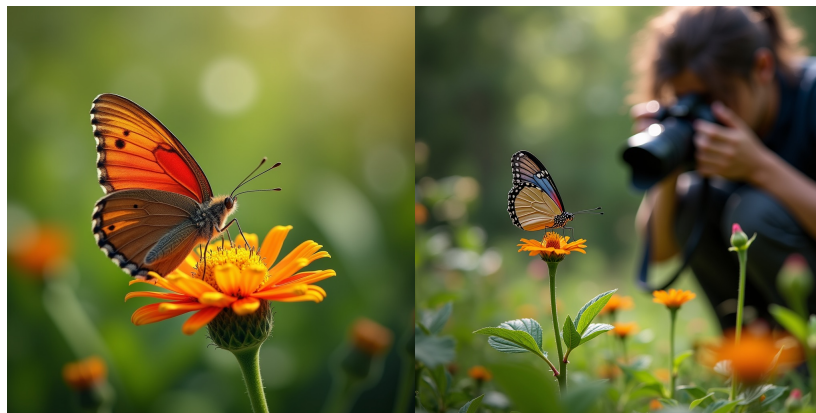
Ours ( $n=0$ )  
(Model: FLUX.1-dev)

Ours ( $n=1$ )  
(Model: FLUX.1-dev)

Figure 10. **Qualitative result in multi-turn:** Demonstrate enhancement performance on including user feedback and successfully modifying the size contrast of dog and tree.

Prompt: A photographer capturing a butterfly on a wildflower.

User Feedback: Incorporate the photographer into the scene



Ours ( $n=0$ )  
(Model: FLUX.1-dev)

Ours ( $n=1$ )  
(Model: FLUX.1-dev)


Figure 11. **Qualitative result in multi-turn:** Demonstrate enhancement performance on including user feedback and successfully including the photographer into the scene.

”


Prompt:  
One cat and three dogs sitting on the grass.  
一貓和三狗坐在草地上

2 Text-Image Alignment \*

Select the image that best represents the content described in the prompt.




☐ Image\_A




☐ Image\_B

3 Text-Image Alignment & Aesthetic Quality \*

Beyond alignment, choose the image that looks more visually appealing and meets your expectations.



☐ Image\_A



☐ Image\_B

Figure 12. The screenshot of user study website.



## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. [2](#)
- [2] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. [2](#)
- [3] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024. [1](#)
- [4] Black Forest Labs. FLUX, 2024. [2](#)
- [5] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. GenAI-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop @ CVPR*, 2024. [1](#)
- [6] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–384, 2024. [1](#), [2](#)
- [7] Mistral AI. Mistral Small 3.1 24B, 2025. [2](#)
- [8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [1](#)
- [9] Lumina Team. Lumina-image 2.0 : A unified and efficient image generative model, 2025. [2](#)
- [10] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [1](#), [2](#)