

The Devil is in the Prompts: Retrieval-Augmented Prompt Optimization for Text-to-Video Generation

Supplementary Material

We provide additional materials to support our main paper. In Section 1, we introduce the specific inference prompt template for prompt optimization utilizing GPT-4 [2] and extraction for relation graph. In Section 2, we provide more qualitative comparisons of generated videos between user-provided prompts and different prompt optimization methods, and present more extension results to other T2V models utilizing our proposed RAPO. In Section 3, we conduct ablation experiments on different hyperparameters towards the scale of relation graph.

1. Specific Inference Prompt Template

In this section, we introduce the specific inference prompt template utilizing GPT-4 for prompt optimization and extracting related modifiers from prompts database of T2V training prompts.

1.1. Prompt Optimization utilizing GPT-4

We prompt GPT-4 to perform prompt optimization as comparison of RAPO, transferring simple user-provided prompts into longer prompts with detailed descriptions about scenes, actions and subjects descriptions. The general prompt instruction template is shown in Table 1. In instruction, for each input user-provided prompt x_i , we obtain corresponding optimized prompt y_i .

GPT-4 Instruction Template for Prompt Optimization.

Instruction: *Imagine you are a Text Optimizer tasked with enhancing user-provided prompts. Refine the input sentence by including a detailed subject description, specifying the action being performed, and vividly describing the scene. Optionally, incorporate elements of camera angles, lighting, shadows, and atmosphere to enrich the imagery. Consider adding complementary actions to make the sentence more dynamic and engaging, ensuring it flows naturally and avoids nonsensical phrasing.*

User-provided prompt: $\{x_i\}$.

Optimized prompt: $\{y_i\}$.

Table 1. GPT-4 instruction template for prompt optimization. In instruction, for each input user-provided prompt x_i , we obtain corresponding optimized prompt y_i .

1.2. Modifiers Extraction for Relation Graph

We prompt large language model to extract scenes and corresponding related modifiers (e.g., subject, action, atmo-

sphere descriptions) from training prompts of T2V models. The general prompt instruction template is shown in Table 2. We provide examples of training prompts with corresponding extracted modifiers $\{D_m = d_i|_{i=1}^{N^m}\}$, in which N^m is the number of examples and d_i contains a training prompt with corresponding extracted modifiers. For each input prompt p_i , we extract modifiers $\{m_i|_{i=1}^{N^k}\}$ utilizing LLM via instruction, in which N^k is the number of extracted modifiers.

LLM Instruction Template for Modifiers Extraction to Construct Relation Graph.

Instruction: *Imagine you are a Modifiers Extractor tasked with extracting modifiers from each prompt. Extract the actions of the subject's ongoing activity and the scene where the activity takes place from a sentence, and the corresponding subject and atmosphere descriptions. Only the results are output, no additional explanation is required.*

Examples: $\{D_m = d_i|_{i=1}^{N^m}\}$.

Input prompt: $\{p_i\}$.

Extracted modifiers: $\{m_i|_{i=1}^{N^k}\}$.

Table 2. LLM instruction template for modifiers extraction to construct relation graph. For each input prompt p_i , we extract modifiers m_i utilizing LLM via instruction.

2. Additional Qualitative Comparisons

We represent additional qualitative comparisons between simple user-provided prompts and different prompt optimized methods for static dimensions and dynamic dimensions as shown in Figure 1 and 2. The optimized prompts from RAPO can significantly enhance both the static and dynamic quality of the generated videos, enabling them to be more visually appealing. Moreover, we represent extension of applying RAPO on CogVideoX [6] and T2V-Turbo [3].

3. Ablation Experiments on Hyperparameters

We conduct ablation experiments on the scale of the constructed relation graph on T2V-CompBench [4]. Specifically, we represent the scale of relation graph by the valid number of sentences composing the relation graph. As shown in Table 3, the performance on different metrics improves with increasing the number of sentences at beginning then gradually stabilizes. At the same time, with the

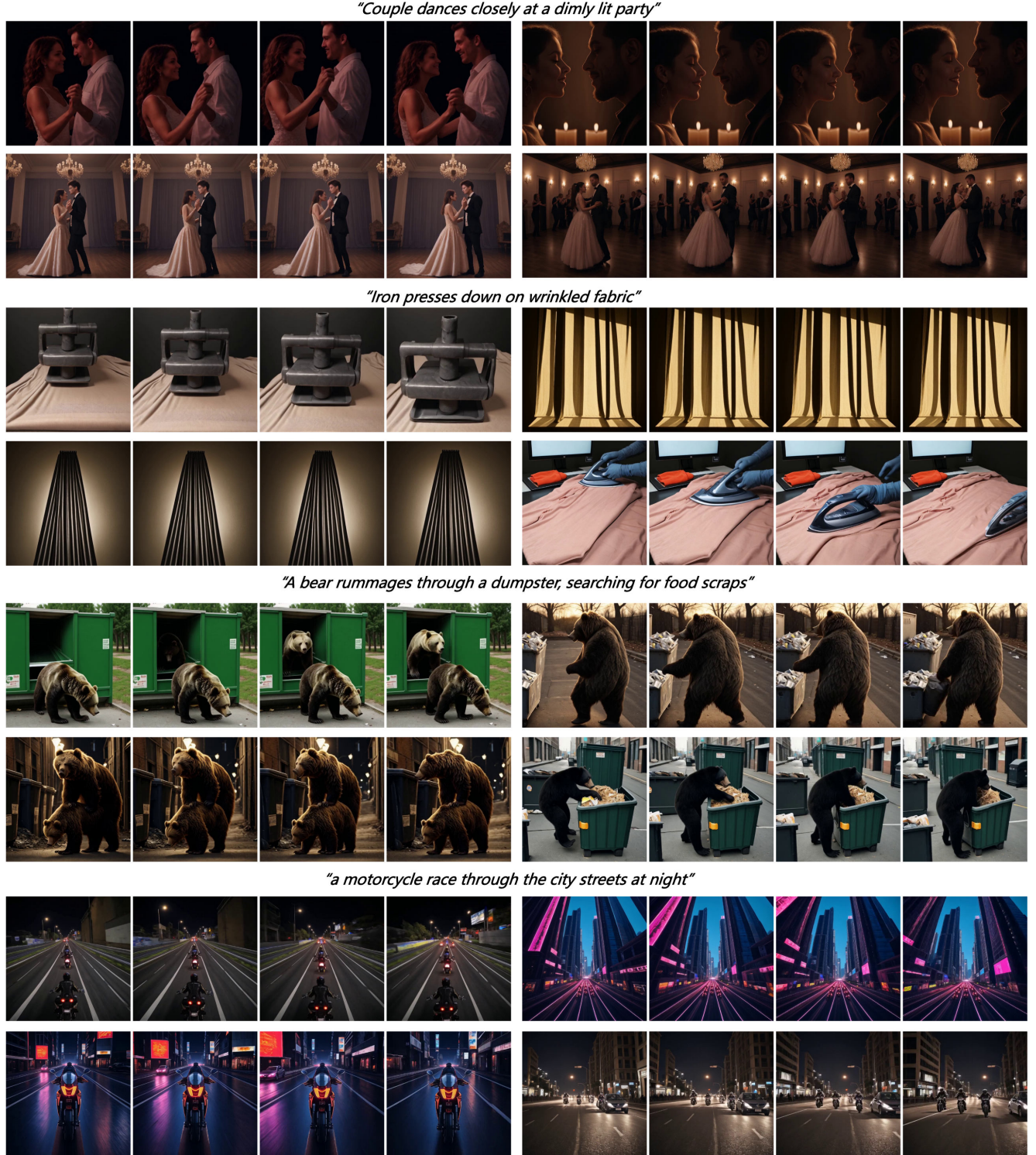


Figure 1. Generated videos for dynamic performance using LaVie [5] with short prompts, prompt optimization methods from GPT-4 [2], Open-sora [1], and our method. From left to right, and top to bottom: user-provided, GPT-4, Open-sora, RAPO.

054
055

increasing of the scale of relation graph, the reference speed also also decreases. Therefore, we need to choose a suitable

scale of relation graph to strike a balance between reference speed and performance.

056
057

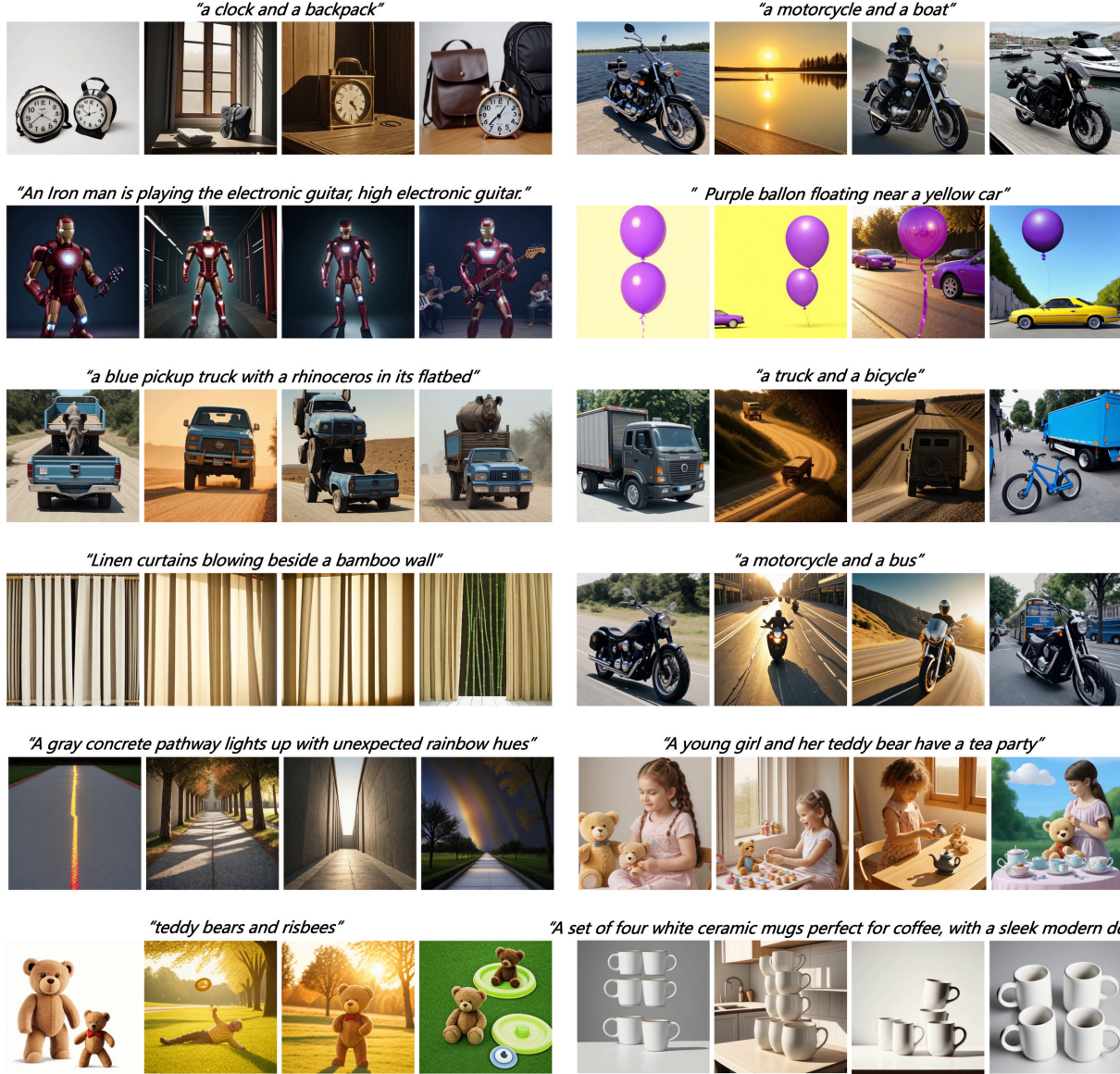


Figure 2. Generated videos for static performance using LaVie [5] with short prompts, prompt optimization methods from GPT-4 [2], Open-sora [1], and our method. From left to right: user-provided, GPT-4, Open-sora, RAPO.

Sentence number	consistent attribute binding	dynamic attribute binding	action binding	object interactions
100k	0.631	0.235	0.534	0.758
400k	0.648	0.229	0.562	0.760
800k	0.672	0.257	0.573	0.821
1200k	0.685	0.267	0.630	0.827
1600k	0.712	0.273	0.631	0.824
2000k	0.692	0.267	0.635	0.839

Table 3. Ablation studies on the scale of relation graph using LaVie. The performance on different metrics improves with increasing sentence numbers at beginning then gradually stabilizes.

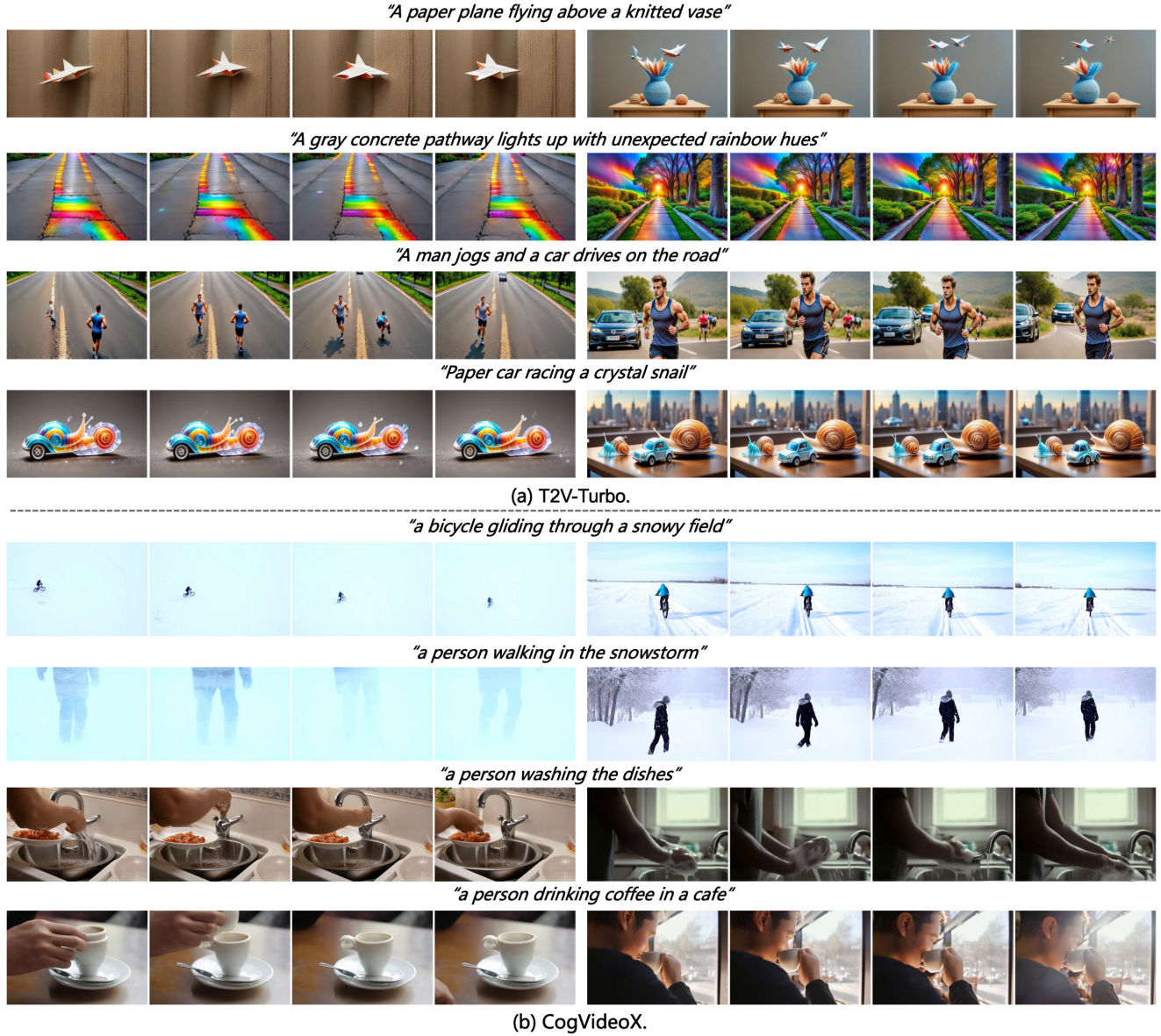


Figure 3. **Extension results of RAPO to CogVideoX and T2V-Turbo.** *Left:* results from user-provided prompts. *Right:* results from RAPO augmented prompts.

References

- [1] Open-sora: Democratizing efficient video production for all. 2024. URL: <https://github.com/hpcaitech/Open-Sora>. 2, 3
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 3
- [3] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhui Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024. 1
- [4] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024. 1
- [5] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2, 3
- [6] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiao

082 han Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video
083 diffusion models with an expert transformer. *arXiv preprint*
084 *arXiv:2408.06072*, 2024. [1](#)