

# Rethinking Prompt Design for Inference-time Scaling in Text-to-Visual Generation

Subin Kim<sup>1</sup>

Yiran Xu<sup>3</sup>

<sup>1</sup>KAIST

Sangwoo Mo<sup>2</sup>

Difan Liu<sup>3</sup>

<sup>2</sup>POSTECH

Mamshad Nayeem Rizve<sup>3</sup>

Jinwoo Shin<sup>1</sup>

<sup>3</sup>Adobe

Tobias Hinz<sup>4</sup>

<sup>4</sup>Meta

subin-kim@kaist.ac.kr

## Abstract

Achieving precise alignment between user intent and generated visuals remains a central challenge in text-to-visual generation, as a single attempt often fails to produce the desired output. To handle this, prior approaches mainly scale the visual generation process (e.g., increasing sampling steps or seeds), but this quickly leads to a quality plateau. This limitation arises because the prompt, crucial for guiding generation, is kept fixed. To address this, we propose *Prompt Redesign for Inference-time Scaling*, coined PRIS, a framework that adaptively revises the prompt during inference in response to the scaled visual generations. The core idea of PRIS is to review the generated visuals, identify recurring failure patterns across visuals, and redesign the prompt accordingly before regenerating the visuals with the revised prompt. To provide precise alignment feedback for prompt revision, we introduce a new verifier, element-level factual correction, which evaluates the alignment between prompt attributes and generated visuals at a fine-grained level, achieving more accurate and interpretable assessments than holistic measures. Extensive experiments on both text-to-image and text-to-video benchmarks demonstrate the effectiveness of our approach, including a 15% gain on VBench 2.0. These results highlight that jointly scaling prompts and visuals is key to fully leveraging scaling laws at inference-time. Visualizations are available at the website: <https://subin-kim-cv.github.io/PRIS>.

## 1. Introduction

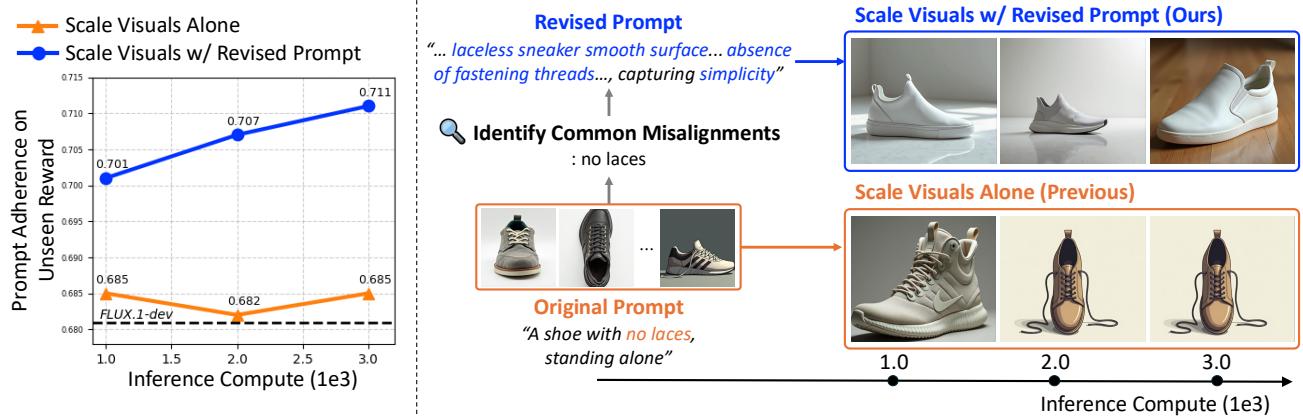
Generative models [5, 21, 35] have achieved remarkable progress across various domains, including language, image, and video, demonstrating strong capabilities in modeling complex data distributions. In the visual domain, denoising models [16, 27] conditioned on textual prompts now allow users to generate high-quality images and videos directly from natural language. However, as prompts become more intricate, e.g., requiring compositional structures in images

or complex motion, camera movements, and causal orders in videos, it becomes increasingly challenging to obtain outputs that fully align with the prompt in a single attempt.

Recent work addresses this shortfall in text-visual alignment by allocating additional compute at inference time (i.e., inference-time scaling). These approaches typically scale the visual generation either by increasing the compute budget for decoding a single candidate from a prompt [30], or by generating multiple candidates for the same prompt to produce a diverse pool of visual outputs [12, 18, 19]. However, they primarily focus on scaling visual parts while keeping the input prompt fixed. This creates a key bottleneck because many generation errors arise from ambiguous or incomplete prompts, and scaling visuals conditioned on a suboptimal prompt offers limited benefit since the prompt provides essential guidance for conditional generation.

More importantly, visual scaling consistently reveals recurring generative failures. For example, in Figure 1, when scaling with the intent “a shoe with no laces, standing alone,” the element “a shoe” is consistently achieved, yet laces still appear in every output. These failure patterns become even more pronounced as prompts grow more complex, such as in text-to-video generation, where producing a high-fidelity sample becomes substantially harder. However, prior prompt-refinement approaches [4, 6, 11, 40] are confined to individual samples, focusing on sample-specific deviations. As a result, they fail to correct the recurring failure modes that consistently appear across samples when scaling, missing the opportunity to jointly improve both the conditioning text and the generated outputs.

To address these limitations, we extend inference-time scaling beyond the visual domain to the text prompts, proposing *Prompt Redesign for Inference-time Scaling* (PRIS). Instead of passively waiting for a high-scoring sample when scaling visuals, PRIS identifies recurring failure modes across scaled visuals and adaptively revises the prompt to reinforce commonly under-addressed aspects while preserving the user’s original intent. Consequently, whereas fixed-prompt inference-time scaling quickly plateaus in prompt



**Figure 1. Our prompt redesign scales with compute, while fixed-prompts plateau.** Given a user-provided complex text prompt, scaling visuals alone with a fixed prompt at inference time often leads to early performance plateaus, especially for unseen rewards (see orange line and boxes). It also repeatedly produces outputs that exhibit common failures and cover only parts of the prompt, even as compute increases to sample more visuals. In contrast, scaling visuals alongside our redesigned prompts yields progressively improved generations and substantially higher prompt-adherence scores as compute increases for both given and unseen rewards (see blue line and boxes).

adherence even as compute increases due to repeated failures (see orange line in Figure 1), PRIS leverages compute more effectively by jointly scaling prompts with the scaled the visuals, enabling sustained improvements in text-visual alignment under the scaling law (see blue line in Figure 1).

To identify failure patterns for prompt revision, we develop Element-level Factual Correction (EFC), a fine-grained descriptive verifier to examining the generated visuals, built on a multimodal large language model (MLLM) (see Figure 2). When assessing the alignment between the visuals and the prompt, EFC first decomposes the prompt into disjoint semantic elements and verifies each against a caption of the generated visual, framing every element as a textual hypothesis. This text-to-text comparison mitigates the affirmative bias common in MLLM-based text-visual question answering [3, 7, 10], thus improving verification accuracy. We further introduce a benchmark pairing each prompt with multiple generated visuals, some fully aligned, others only partially so. On this dataset, EFC consistently distinguishes ground-truth visuals from plausible but misaligned distractors, significantly outperforming existing verifiers.

**Contributions.** Our contributions are as follows:

- We propose PRIS, Prompt Redesign for Inference-time Scaling, which identifies recurring failure patterns during visual scaling and revises prompts accordingly.
- We introduce a new verifier that assesses fine-grained alignments with text-based comparisons for prompt redesign.
- PRIS consistently enhances text-visual alignment without compromising visual fidelity, yielding a 7% improvement on GenAI-Bench for text-to-image and a 15% improvement on VBench2.0 for text-to-video generation.
- We present the first benchmark for evaluating verifiers in inference-time scaling.

## 2. Related Work

**Scaling inference-time compute in visual generation.** Despite recent progress driven by powerful denoising architectures [16, 27], producing faithful outputs in text-to-visual generation remains challenging, particularly for complex prompts. Since outputs are shaped jointly by the initial noise, the sampling trajectory, and the prompt, existing inference-time scaling methods allocate additional compute to exploring better noise seeds and trajectories, either by increasing sampling steps or by generating multiple candidates [30], often with advanced search algorithms [18, 19]. Reward models [28, 42] then score these candidates and select the best one, either at the final output stage (Best-of- $N$ ; BoN) [30] or during sampling through Search-over-Paths [12], which propagates high-reward trajectories. However, all prior methods share a key limitation: they expand only the visual search space while keeping the prompt fixed. In contrast, we treat the prompt as an equally critical and previously underexplored axis of inference-time scaling. Rather than discarding low-scoring generations, we analyze their recurrent failure patterns and redesign the prompt jointly with visual scaling, enabling subsequent generations to receive progressively stronger and more targeted guidance.

**Prompt design in text-to-visual generations.** Prompt design is a critical component of text-conditioned generation that serves not merely as a pre-processing step [43] but as a means to improve model comprehension, output quality, and adherence to the input description. Since the prompt itself guides the generation, even different phrasings of the same user intent can produce markedly different outputs. Yet, crafting effective prompts remains challenging, often requiring tedious trial-and-error. To address this, recent approaches [4, 6, 11, 14, 40] propose systems that interac-

tively help users explore alternative phrasings or automatically rewrite prompts, reducing reliance on naïve iterations. However, these approaches typically require human involvement [4, 40], and more importantly, they operate independently of inference-time scaling [6, 11, 14]: their per-sample rewrites react to individual noisy failures rather than addressing the recurring failure patterns that are revealed through visual scaling, limiting their ability to improve adherence. To fill this gap, we demonstrate that redesigning prompts based on *shared failures* across samples—rather than isolated per-sample corrections—achieves substantially higher adherence as compute scales. Our method applies to both text-to-image (T2I) and text-to-video (T2V) generation, whereas prior prompt-refinement efforts largely focus on T2I generations.

**Chain-of-thought and reasoning.** Incorporating chain-of-thought (CoT) reasoning into visual generation has emerged as a promising paradigm for improving image quality through iterative reflection and guidance [9, 17, 25, 38, 45]. Recent works pursue this direction via unified models [34] that combine visual understanding and generation and jointly optimize large language models with multimodal objectives and generation-specific losses. Our approach differs in two key aspects. First, we use the off-the-shelf MLLM [2], without any additional training, to scale prompts for arbitrary text-conditioned generators. Thus, our design supports both T2I and T2V settings and remains compatible with unified models for prompt refinement. Second, existing CoT-based approaches typically apply reasoning at the per-sample level, refining each output independently [45] or deciding whether a particular sample should be retained or discarded [9]. In contrast, our method aggregates information across the samples generated during scaling and updates the prompt based on cross-sample trends rather than isolated reflections.

### 3. Prompt Redesign for Inference-time Scaling

To enable common-failure-aware visual feedback in inference-time scaling, we introduce a new verifier, EFC, which provides fine-grained assessments of generated visuals (Section 3.1). Building on these assessments, we then present our prompt redesign strategy, PRIS, which extracts common failures and revises the prompt accordingly to improve fidelity as compute scales (Section 3.2).

#### 3.1. EFC: Element-level Factual Correction

Our goal is to identify recurring misalignments between the original prompt and the generated visuals, focusing on elements that the generator repeatedly fails to realize, including missing components, incorrect causal relations, and disordered temporal motions. Achieving this requires a fine-grained visual verifier capable of determining, for each generation, which prompt elements are satisfied or missing, as previous single-scalar alignment scores from previous ver-

ifiers [26, 28] cannot reveal such details. To this end, we introduce Element-level Factual Correction (EFC), a new verifier that provides interpretable, fine-grained text-visual assessments using an MLLM without additional training. See Figure 2 (a) for an overview; further illustrations of EFC are provided in Appendix B.

**Break down the prompt into fine-grained elements.** Holistic alignment scores often obscure which aspects of a generation are actually satisfied; as prompts become more complex, a visual can miss some of the required elements, and such omissions cannot be revealed by a single holistic score. To address this, EFC first decomposes the original prompt  $p$  into a set of verifiable atomic semantic elements  $p = \{p_1, p_2, \dots, p_s\}$ , where each element  $p_i$  corresponds to a distinct element. Here, atomic facts are extracted according to predefined semantic categories, such as image-level elements covering object presence, properties, and spatial arrangement, and motion-level elements covering object motion, camera movement, scene transitions, and temporal ordering. Then, EFC classifies each  $p_i$  as either {core, extra}. The core elements are objective, factual, and essential to the intent of the prompt, while the extra elements are more subjective or stylistic, so they are often flexibly interpreted. This classification is then used to guide the prioritization of generated samples during final scoring.

**Assess each element through factual correction.** After decomposing the prompt into multiple elements, EFC performs factual correction on each element by evaluating it against each generated visual  $D$  to determine whether the element is accurately realized. Here, EFC assesses alignment using a text-text verification approach, rather than direct yes/no visual question answering (VQA), because it achieves higher accuracy by mitigating confirmation bias and consistently providing interpretable descriptions, whereas binary VQA often omit such information. To enable this, EFC first generates a natural-language caption for  $D$ , then infers the relationship between each element  $p_i$  and this caption. This step is formulated as a natural language inference (NLI) task: if the caption semantically supports the element, the relationship is labeled as *entailment*; if the caption contradicts the element, it is labeled as *contradiction*; and if the caption does not provide sufficient information to confirm the element, the label is *neutral*. For any element  $p_i$  initially classified as *neutral*, such as when the caption omits or ambiguously describes it, EFC generates an open-ended question  $q_i$  whose expected answer corresponds to  $p_i$ , without binary framing. Then, EFC queries  $q_i$  over the visual  $D$ , obtain a free-form response, evaluate it against  $p_i$  in a second NLI step, and relabels the element as either *entailment* or *contradiction*.

**Prioritize core elements in final scoring.** After factual correction, we obtain verification results  $C$  for each visual. EFC then assigns a score based on the number of elements labeled

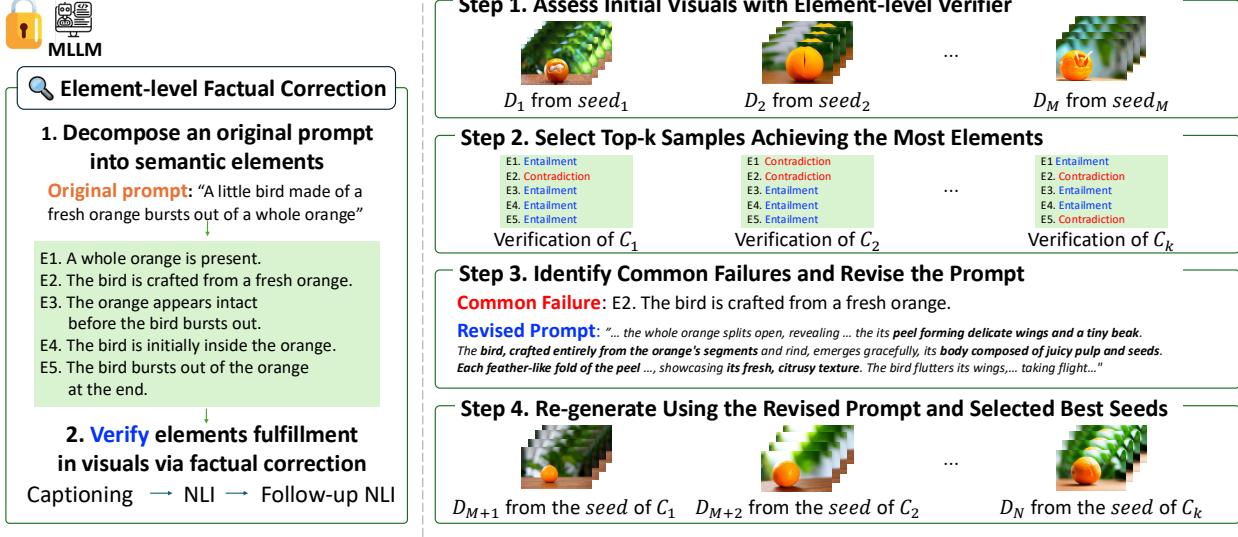


Figure 2. **Overview of Prompt Redesign for Inference-time Scaling (PRIS)**, which leverages diagnostic feedback from our verifier EFC to revise prompts during inference based on generated visuals. EFC decomposes prompts into semantic elements and verifies each element for fine-grained text-visual alignment (left). Guided by the EFC, PRIS proceeds as follows (right): Step 1 reviews initial generations with EFC; Step 2 selects top- $k$  successful samples and identifies recurring failures; Step 3 redesigns the prompt to emphasize common failures; and Step 4 regenerates visuals with the revised prompt and top- $k$  seeds. The process can be iterated by returning from Step 4 to Step 2.

as *entailment* in  $C$ . Core elements are prioritized because they are objective, factual, and less open to subjective interpretation, making them essential for faithfully capturing the prompt’s intent. When multiple candidates achieve the same core accuracy, EFC breaks ties using extra-element accuracy.

### 3.2. Common-failure-aware Prompt Redesign for Inference-time Scaling

We propose a prompt redesign framework, PRIS, which revises the prompt to address common failures across samples using the fine-grained text-visual alignment assessments produced by our verifier EFC. By pinpointing where alignment breaks down in earlier generations, PRIS incorporates these diagnostic signals into subsequent prompt updates, guiding later samples toward higher fidelity to the original intent. See Figure 2 (b) for an overview.

- **Step 1. Generation and verification.** PRIS first generates  $M$  candidate visual samples and evaluates the fulfillment of elements  $\{p_1, p_2, \dots, p_s\}$  for each sample using our verifier EFC (Section 3.1), obtaining verification results for each sample ( $C_1$  through  $C_M$ ).
- **Step 2. Select the top- $k$  best-performing samples.** PRIS then selects the top- $k$  samples that collectively cover the largest number of elements, with ties further resolved using the scalar score from a reward model [26, 28] trained on human-preference datasets. This ensures that the selected samples better reflect human-preferred ones.
- **Step 3. Identify misalignment patterns and revise the prompt.** Within the selected subset, PRIS identifies com-

mon failures, defined as elements whose success probability is below 50% within the top- $k$  samples, by aggregating the element-level assessments from EFC across these visuals. Based on these common failures, PRIS revises the *original prompt*  $p$  into a *revised prompt*  $p'$  that explicitly reinforces overlooked elements while preserving those already well addressed. This targeted refinement encourages subsequent generations to focus more effectively on under-represented elements. If no common failures are observed (i.e., every element has a success probability above 50%), PRIS instead treats the prompt itself as the refinement target to encourage exploration of prompt variations.

- **Step 4. Regenerate with the revised prompt and selected noise conditions.** Using the revised prompt  $p'$ , PRIS regenerates  $(N - M)$  samples by reusing the noise latents of the top- $k$  samples. Reusing these seeds preserves earlier successes more reliably than random initialization, as certain noise conditions are known to yield better alignment for specific prompt types [1, 32, 44]. After regeneration, PRIS verifies and ranks the samples using EFC.

This generation-prompt revision-regeneration loop can be repeated. In our main experiments, we apply it once, as this already provides substantial gains; further analysis is provided in Section 4.4. Within PRIS, EFC-guided prompt redesign leverages prior failures to refine prompts and exploit favorable noise configurations. By treating partially correct generations as informative feedback rather than discarding them, PRIS makes more effective use of the generator’s previously expended compute and improves output fidelity.

Table 1. **Quantitative results of T2I on GenAI-Bench.** \* denotes results with standard prompt expansion; BoN refers to ‘‘Best-of- $N$ ’’ selection using fixed prompts. **Bold** shows the best.

Method	VQA-Score (Given)	DA-Score w. BLIP2-VQA (Unseen)	Aesthetic Quality (Unseen)
FLUX.1-dev [21]	0.718	0.681	5.764
+BoN	0.783	0.682	5.761
<b>+PRIS</b>	<b>0.854</b>	<b>0.707</b>	<b>5.765</b>
FLUX.1-dev* [21]	0.769	0.695	5.824
+BoN*	0.829	0.710	5.820
<b>+PRIS*</b>	<b>0.853</b>	<b>0.713</b>	<b>5.841</b>

## 4. Experiments

We comprehensively evaluate PRIS and EFC for inference-time scaling. First, we study the effect of prompt redesign under a fixed compute budget (Section 4.1). Next, we analyze the scaling behavior of PRIS by expanding the generator’s compute budget or iteratively revising prompts, and examine its integration with visual scaling algorithms originally designed for fixed prompts (Section 4.2). Finally, we assess the ability of EFC and existing verifiers to select the best-quality sample from mid-quality candidates (Section 4.3) and conduct ablations on both PRIS and EFC (Section 4.4).

### 4.1. Effect of PRIS on Inference-time Alignment and Visual Quality

We study the effect of prompt redesign on output quality under a fixed compute budget, defined as the number of function evaluations (NFE). For detailed experimental setups, please refer to Appendix C.

**Experimental setup.** For EFC, our MLLM-based verifier, we use Qwen2.5-VL [2] with the process-specific prompt instructions described in Section 4.3, without any additional training. We compare with Best-of- $N$  (BoN) [30], which generates  $N$  samples at once and selects the best, while our method generates half of them (setting  $M = \lfloor N/2 \rfloor$ ), revises the prompt with feedback, and regenerates the rest using the revised prompt and top- $k$  seeds. We set  $k = \lceil N/4 \rceil$ , thereby producing two revised prompt variants for the remaining  $N - M$  samples. We also include standard prompt expansion [35] (denoted as \*) to contrast with our failure-aware revisions. We evaluate on challenging benchmarks [24, 43] where generations frequently exhibit partial failures. Thus, for base generator selection, we first measure each generator’s base prompt fidelity using embedding similarity [36] between the generated caption and the original prompt, and retain only those with sufficiently high adherence to exclude substantially weak models.

For T2I generation, we use FLUX.1-dev [22] on GenAI-Bench [24] sampling 320 prompts (20% of the full set) to avoid redundancy. For the guidance reward, we utilize VQA-Score [26]. For held-out evaluation, we use



Figure 3. **Qualitative comparisons of T2I generation.** \* denotes results with standard prompt expansion.

DA-Score [33] for fine-grained prompt adherence and an aesthetic predictor [23] for image quality. NFE is set to 2000 ( $N = 20, 50$  denoising steps, classifier-free guidance [15], cfg=3.5). For T2V generation, we use Wan2.1-1.3B/14B [35] with VideoAlign [28] as guidance, and evaluate on VBench2.0 [43] across four dimensions: controllability, creativity, commonsense, and physical plausibility. NFE is set to 2000 ( $N = 20, 50$  steps, cfg=6) for Wan2.1-1.3B and 1000 ( $N = 10, 50$  steps, cfg=6) for Wan2.1-14B.

**Experimental results on T2I generation.** We present results in Table 1, Figure 3, and Appendix D. As shown in Table 1, our approach PRIS consistently outperforms all baselines across metrics. Notably, it yields substantial gains in prompt adherence while maintaining comparable aesthetic quality. Even against the standard prompt expansion variant (denoted as \*), our method achieves significantly higher scores. These results suggest that prompt expansion is most effective when guided by visual feedback, rather than by adding arbitrary details as in standard prompt expansion. The qualitative results in Figure 3 further support this claim, showing that PRIS exhibits a stronger ability to handle complex, compositional prompts compared to BoN. For the top row in Figure 3, after identifying layout specification as a challenge in the initial outputs, our method revises the prompt to emphasize layout-related details. Likewise, for the prompt ‘fork not made of wood’ (bottom row in Figure 3), where BoN still produces wooden forks due to the negation, our method explicitly instructs the model to generate ‘silver forks,’ thereby resolving the misunderstanding.

**Experimental results on T2V generation.** Our method delivers substantial improvements in prompt alignment for T2V generation, as shown in Table 2, Figure 4, and further examples in Appendix D. PRIS achieves gains of +13.88% and +15.19% in the Controllability and Creativity categories for the small and large models, respectively. This significantly surpasses BoN\*, which applies standard prompt expansion at initialization without visual feedback on where to focus. Specifically, the largest gains appear in Dynamic Attribute and Motion Order Understanding, which require sequential reasoning (e.g., ‘‘A then B,’’ ‘‘A transitioned to B’’). Here, PRIS identifies failures in the initial outputs and revises

Table 2. **Quantitative comparisons of T2V generation on VBench-2.0.** \* denotes results obtained using the standard prompt expansion, and **bold** indicates the best results. We use  $N = 20$  samples for Wan2.1-1.3B (small) and  $N = 10$  for Wan2.1-14B (large), which can lead to the smaller model achieving higher scores due to the larger number of samples. BoN refers to “Best-of- $N$ ” selection using fixed prompts.

Category	Method	Dynamic Spatial Relationship	Dynamic Attribute	Motion Order Understanding	Human Interaction	Composition	Average
Controllability & Creativity	Wan2.1-1.3B* [35]	35.56	46.67	52.87	74.44	48.33	51.57
	+BoN* ( $N = 20$ )	<b>43.33</b>	53.33	51.72	<b>90.00</b>	50.2	57.73 $\uparrow +6.16$
	<b>+PRIS*</b> ( $N = 20$ )	<b>43.33</b>	<b>73.33</b>	<b>68.97</b>	<b>90.00</b>	<b>51.6</b>	<b>65.45 <math>\uparrow +13.88</math></b>
	Wan2.1-14B* [35]	50.00	48.89	43.33	78.89	47.18	53.66
	+BoN* ( $N = 10$ )	46.67	56.67	60.00	80.00	49.23	58.51 $\uparrow +4.85$
	<b>+PRIS*</b> ( $N = 10$ )	<b>60.00</b>	<b>73.33</b>	<b>66.67</b>	<b>90.00</b>	<b>54.23</b>	<b>68.85 <math>\uparrow +15.19</math></b>
Category	Method	Camera Motion	Motion Rationality	Mechanics	Material	Thermotics	Average
Commonsense & Physics	Wan2.1-1.3B* [35]	41.38	38.10	75.00	75.38	<b>86.25</b>	63.22
	+BoN* ( $N = 20$ )	37.93	35.71	84.00	73.91	81.48	62.61 $\downarrow -0.61$
	<b>+PRIS*</b> ( $N = 20$ )	<b>51.72</b>	<b>50.00</b>	<b>80.00</b>	<b>78.26</b>	70.37	<b>66.07 <math>\uparrow +3.46</math></b>
	Wan2.1-14B* [35]	36.67	40.00	83.33	77.78	<b>79.49</b>	63.45
	+BoN* ( $N = 10$ )	<b>43.33</b>	43.33	<b>86.36</b>	80.77	76.92	66.14 $\uparrow +2.69$
	<b>+PRIS*</b> ( $N = 10$ )	<b>43.33</b>	<b>53.33</b>	<b>86.36</b>	<b>88.46</b>	76.92	<b>69.98 <math>\uparrow +6.53</math></b>

Original short prompt: “A person is turning on the desk lamp.”

Initially expanded prompt: “A person is... gently turning on a desk lamp... twist the lamp’s switch... newly lit lamp casts warm light...”

Revised prompt: “A young... hand resting gently on the base of a desk lamp... the person twists the switch, and the warm glow gradually illuminates the space... as the lamp turns on... background remains softly blurred... the transition from darkness to light emphasizes the calming effect... warmly lit room...”



Figure 4. **Qualitative comparisons on T2V generation.** Our revised prompt elaborates on previous failures by emphasizing causal order, ensuring the lamp turns on immediately when touched.

prompts to clarify how sequences should unfold, emphasizing the parts that previously failed. Qualitative examples, including revised prompt in Figure 4, illustrate these improvements. Beyond these categories, PRIS also improves Commonsense and Physics by +3.46% and +6.53%, respectively. A notable exception is Thermotics, where performance drops slightly due to the reward model overfitting to exact numeric values rather than broader physical plausibility. Finally, while Zheng et al. [43] suggests that camera motion is largely determined by base model capacity, our results show that refining prompts to specify how camera motion should unfold in conjunction with other scene elements can still yield measurable improvements.

## 4.2. Scaling Behaviors of PRIS

**PRIS scales with increasing NFEs.** We provide both quantitative and qualitative evidence that PRIS scales with increasing NFEs, increasing the number of samples  $N$ , whereas a fixed prompt quickly saturates and fails to scale (Figures 1 and 5). In Figure 1, BoN, which relies on fixed prompts, plateaus on held-out evaluation beyond 1e3 NFEs, while PRIS continues to improve, both quantitatively in reward

score and qualitatively in visual alignment. Similarly, Figure 5 further illustrates this trend qualitatively: PRIS produces progressively taller trees while satisfying all prompt elements, even at smaller budgets, whereas BoN repeatedly fails, generating a boy wearing a helmet.

**Effectiveness of iterative prompt revisions with PRIS.** Table 3 evaluates whether iterative revisions, which update prompts based on newly identified failures, provide benefits beyond a single iteration of revision. As shown, iterative revision yields consistent improvements across both given and held-out metrics for prompt adherence while maintaining comparable aesthetic quality. The first update brings a substantial gain, and the second adds further improvements, suggesting that iterative revision progressively strengthens alignment. While multiple revisions yield cumulative gains, the first update already offers a substantial improvement; therefore, we adopt a single refinement step in the main experiments. Moreover, such gains do not appear without PRIS. Simply generating more samples with increased compute budget leads to saturated performance on unseen rewards. This highlights that targeted prompt correction is more effective than brute-force visual scaling.



Figure 5. **Qualitative examples with increasing inference-time compute.** PRIS generates progressively taller trees while satisfying all attributes, whereas BoN consistently misses some.

**Integration with visual scaling methods beyond BoN.** PRIS is complementary to existing visual scaling methods that expand the sampling space with fixed prompts. While these approaches modify noise or sampling dynamics, PRIS targets prompt-level failures and can be combined with them by enabling prompt revision. To validate this complementarity, we integrate PRIS with two established T2I methods, DAS [19] and RBF [18], following their original protocols and evaluating on GenAI-Bench consistent with our main setup. Table 4 and Figure 6 show that integration yields superior alignment on both given and unseen rewards. Notably, whereas RBF often sacrifices aesthetics to improve alignment, our approach improves both simultaneously. Qualitatively, Figure 6 further shows that although DAS and RBF alone struggle on difficult prompts, their integration with PRIS resolves these cases, producing outputs that are both prompt-aligned and visually coherent. Full experimental details, additional examples, and T2V results are provided in Appendix A.

### 4.3. Evaluating the Verification Accuracy of EFC

We introduce EFC to address the lack of fine-grained evaluation in text-visual alignment, enabling element-level verification for precise and interpretable assessment. As prompts become more complex, it is critical to check whether all attributes are satisfied rather than relying on a single opaque score. However, widely used human preference datasets [13, 28, 29, 37, 41] are insufficient as they provide only pairwise judgments and do not capture whether a single video fully satisfies the prompt. Moreover, they do not reflect inference-time needs, where a verifier must pick the best-aligned sample from a diverse pool of mid-quality outputs. To fill this gap, we construct a benchmark that pairs prompts with both fully aligned (ground-truth) and partially aligned (distractors) visuals, covering varying degrees of completeness and fidelity. Additional dataset details and analyses are provided in Appendix B.

**Constructing the benchmark.** Each prompt is paired with

Table 3. **Quantitative results for iterative prompt refinement with increasing inference-time compute.** Iteratively revision prompts consistently improves reward scores by addressing common failures, and the gains even generalize to unseen rewards. In contrast, fixed prompts often saturate and fail to transfer.

Method	NFEs (1e3)	VQA-Score (Given)	DA-Score w. BLIP2-VQA (Unseen)	Aesthetic Quality (Unseen)
Initial (w.o. revision)	0.5	0.736	0.679	5.756
Initial (w.o. revision)	1.0	0.764	0.684	5.766
1 <sup>st</sup> revision	1.0	0.834	0.703	5.755
Initial (w.o. revision)	1.5	0.776	0.683	5.751
2 <sup>nd</sup> revision	1.5	0.849	0.705	5.740

multiple aligned and partially aligned videos, along with tags indicating reasons for misalignment for the partially aligned cases. Prompts are drawn from widely used video model demos and from VBBench 2.0 [43], yielding a total of 410 prompts. Candidate videos are generated using diverse state-of-the-art closed- and open-source models [8, 20, 35]. Several human annotators mark a video as aligned if it fully satisfies the prompt and provide explanations when they label it as misaligned. The final label for each video is determined by majority vote.

**Evaluation setup and baselines.** We evaluate EFC and existing verifiers on our benchmark which simulates the inference-time scaling setting. As baselines, we consider widely used learned reward models [28, 39, 41], trained on preference datasets to output a scalar score per video. We then evaluate EFC itself, which performs zero-shot prompt-adherence verification using MLLMs [2], along with an ablation that removes its factual correction component. In this ablation, verification is reduced to decomposed visual QA, where each element is judged independently via QA, and the final score is determined by the number of elements marked aligned. For all methods, we select the top-scoring video and evaluate against human annotations.

Table 5. **Quantitative results on verifier accuracy** in selecting GT visual outputs. **Bold** indicates the best results.

Verifier	Accuracy
VisionReward [41]	0.571
UnifiedReward [39]	0.498
VideoAlign [28]	0.693
Decomposed binary VQA	0.700
PRIS (Ours)	<b>0.763</b>

**Evaluation results.** Table 5 shows that EFC achieves the highest accuracy, significantly surpassing even VideoAlign, the strongest reward model. Moreover, unlike learned reward models, EFC provides fine-grained, interpretable reasoning even without training. It also outperforms decomposed VQA,

Table 4. Quantitative results of integrating PRIS with T2I visual scaling methods on GenAI-Bench. BoN refers to “Best-of-N” selection using fixed prompts. **Bold** shows the best.

Method	VQA-Score (Given)	DA-Score w. BLIP2-VQA (Unseen)	Aesthetic Quality (Unseen)
SDXL [31]	0.639	0.652	5.759
+BoN	0.649	0.663	5.810
+DAS [19]	0.657	0.671	5.819
<b>+DAS w/ PRIS</b>	<b>0.700</b>	<b>0.688</b>	<b>5.897</b>
FLUX.1-schnell [21]	0.672	0.676	5.519
+BoN	0.869	0.704	5.497
+RBF [18]	0.922	0.706	5.426
<b>+RBF w/ PRIS</b>	<b>0.936</b>	<b>0.723</b>	<b>5.528</b>

supporting our design choice of factual correction with text-based verification, consistent with recent findings that text-based measures are more reliable than direct VQA [3, 7].

#### 4.4. Ablation Study and Analysis

**Impact of common-failure-aware prompt redesign.** We conduct an ablation comparing our redesign strategy, which revises prompts based on failure patterns consistently shared across the top- $k$  best-aligned samples, against a per-sample prompt revision baseline that attempts to fix failures independently for each sample. Our method focuses revisions on attributes that are systematically difficult for the model to generate, rather than dispersing corrections across noisy, sample-specific failures. As shown in Table 6, common-failure revision consistently outperforms per-sample revision in both T2I and T2V.

Table 6. Ablation study of PRIS. # d.e. and # c.f. denote the numbers of decomposed elements and common failures, respectively; **bold** indicates the best.

Task	Prompt Revision	Avg. # d.e.	Avg. # c.f.	Score
T2I	w/o revision			0.783
	Per-sample	3.5	-	0.853 $\uparrow +0.070$
T2V	Common-failure		0.72	<b>0.854 <math>\uparrow +0.071</math></b>
	w/o revision			0.711
T2V	Per-sample	7.3	-	0.619 $\downarrow -0.092$
	Common-failure		1.46	<b>0.754 <math>\uparrow +0.043</math></b>

Notably, in T2V generation, where fully aligning the generated visuals with the prompt is considerably more challenging than in T2I, per-sample revision performs worse than applying no revision at all. Attempting to correct every observed error dilutes the update signal, spreads compute on low-probability or seed-specific artifacts, and ultimately fails to address the truly persistent misalignments. In contrast, our method anchors prompt updates to high-lielihood failures that repeatedly surface across strong samples, making revisions both focused and efficient. Finally, the number of common failures (c.f. in Table 6) observed among top-

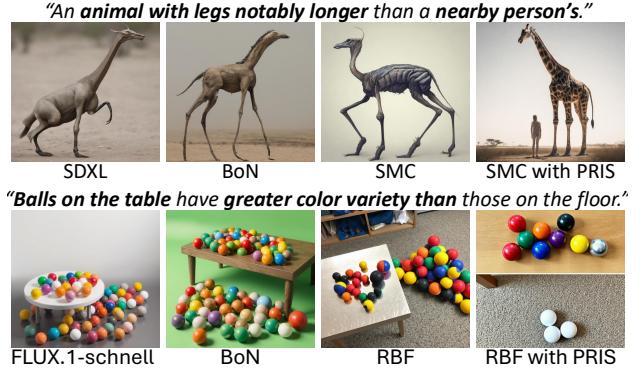


Figure 6. Qualitative examples of integrating PRIS with T2I noise-scaling baselines.

performing seeds empirically supports our key motivation: different seeds do share recurring failure modes, and exploiting these shared patterns becomes increasingly essential for effective prompt redesign as prompt complexity grows.

**Compute time analysis.** In our experiments, we follow the standard practice of comparing methods under the same NFEs [18, 30]. We also evaluate under matched total wall-clock time in Table 7, including verifiers. Even under this setting, allocating wall-clock time to our framework is more effective than simply increasing NFEs (i.e., increasing  $N$ ) for the generator. Although EFC introduces a modest overhead, mainly from captioning, PRIS achieves substantially larger gains in prompt adherence. These results indicate that directing wall-clock time toward verifier-guided prompt revision is more beneficial than spending the same time on brute-force generation. Please refer to Appendix A for a detailed breakdown of the compute time for verification.

Table 7. Quantitative evaluation with matched compute.

Task	Method	NFEs (1e3)	Score
T2I	BoN	4.0	0.790
	PRIS	1.0	<b>0.834</b>
T2V	BoN	4.0	0.935
	PRIS	2.0	<b>0.964</b>

## 5. Conclusion

We address an overlooked gap in inference-time scaling by redesigning prompts from common failures in samples, rather than relying on visual-only scaling or isolated per-sample prompt updates as in prior work. We introduce EFC, a verifier that provides fine-grained text-visual alignment assessments, and PRIS, a EFC-guided framework that revises prompts based on common failure patterns observed across generated outputs. By reviewing generated visuals to identify recurring misalignments, PRIS adaptively updates the prompt in response to the generator’s failures. Across both T2I and T2V settings, our approach achieves substantial gains in prompt adherence, demonstrating that jointly scaling

prompts and visuals yields stronger scaling behavior than visual-only methods. Furthermore, our findings show that leveraging recurring failure patterns for prompt redesign is crucial to fully realize the benefits of inference-time scaling.

## References

- [1] Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024. 4
- [2] Shuai Bai, Kefin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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-v1 technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 5, 7
- [3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 2, 8
- [4] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14, 2023. 1, 2, 3
- [5] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1
- [6] Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. Prompt expansion for adaptive text-to-image generation. In *ACL*, 2024. 1, 2, 3
- [7] Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *CoLM*, 2025. 2, 8
- [8] Google DeepMind. Veo: Google deepmind’s text-to-video generation model. <https://deepmind.google/models/veo/>, 2025. Version 3. Accessed: 2025-09-22. 7, 10
- [9] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let’s verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 3
- [10] Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. The instinctive bias: Spurious images lead to illusion in mllms. In *EMNLP*, 2024. 2
- [11] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *NeurIPS*, 2023. 1, 2, 3
- [12] Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. *arXiv preprint arXiv:2505.17618*, 2025. 1, 2, 7
- [13] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narasupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhui Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *EMNLP*, 2024. 7
- [14] Nailei Hei, Qianyu Guo, Zihao Wang, Yan Wang, Haofen Wang, and Wenqiang Zhang. A user-friendly framework for generating model-preferred prompts in text-to-image synthesis. In *AAAI*. AAAI Press, 2024. 2, 3
- [15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1, 2
- [17] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. In *NeurIPS*, 2025. 3
- [18] Jaihoon Kim, Taehoon Yoon, Jisung Hwang, and Minhyuk Sung. Inference-time scaling for flow models via stochastic generation and rollover budget forcing. In *NeurIPS*, 2025. 1, 2, 7, 8
- [19] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *ICLR*, 2025. 1, 2, 7, 8, 4
- [20] Kuaishou Technology. Kling: Kuaishou’s video generation model. <https://help.scenario.com/en/articles/kling-video-models-the-essentials/>, 2025. Version 2.1. Accessed: 2025-09-22. 7, 10
- [21] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 1, 5, 8
- [22] 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, 4
- [23] LAION. LAION-Aesthetics V2 — 600M subset (aesthetic  $\geq 5$ ). <https://laion.ai/blog/laion-aesthetics/>, 2024. Accessed: 2025-08-14. 5
- [24] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 5
- [25] Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. *arXiv preprint arXiv:2503.19312*, 2025. 3
- [26] 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 *ECCV*, 2024. 3, 4, 5

- [27] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 1, 2
- [28] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 2, 3, 4, 5, 7, 10
- [29] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *CVPR*, pages 8009–8019, 2025. 7
- [30] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Scaling inference time compute for diffusion models. In *CVPR*, pages 2523–2534, 2025. 1, 2, 5, 8
- [31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 8, 4
- [32] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. 4
- [33] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. In *NeurIPS*, 2023. 5
- [34] Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfai Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025. 3
- [35] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 5, 6, 7, 10
- [36] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Ranjan Majumder, and Furu Wei. Improving text embeddings with large language models. In *ACL*, 2023. 5
- [37] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024. 7
- [38] Yi Wang, Mushui Liu, Wanggui He, Longxiang Zhang, Ziwei Huang, Guanghao Zhang, Fangxun Shu, Zhong Tao, Dong She, Zhelun Yu, et al. Mint: Multi-modal chain of thought in unified generative models for enhanced image generation. *arXiv preprint arXiv:2503.01298*, 2025. 3
- [39] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 7, 10
- [40] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024. 1, 2, 3
- [41] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. 7, 10
- [42] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2
- [43] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 2, 5, 6, 7
- [44] Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *ICCV*, 2025. 4
- [45] Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *ICCV*, 2025. 3, 8

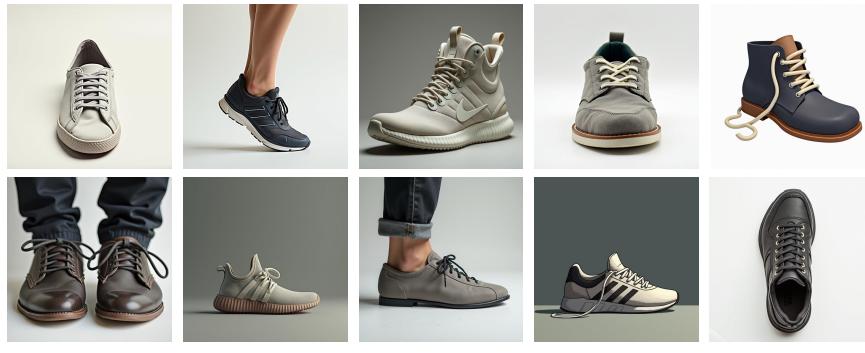
# Rethinking Prompt Design for Inference-time Scaling in Text-to-Visual Generation

## Supplementary Material

### A. Additional Analysis

#### A.1. Qualitative Examples of Common Failures

We present qualitative examples of the identified common failure patterns for text-to-image generation in Figure 7 and for text-to-video generation in Figure 8.



**Original prompt**  
: "A shoe with no laces, standing alone"

**Decomposed elements**  
1. There is a shoe (core)  
2. The shoe has no laces (core)  
3. The shoe is standing alone (core)

**Common failure patterns**  
: 2. The shoe has no laces



**Original prompt**  
: "On a wooden table, both the spoons and plates are made of wood, only the fork is not made of wood."

**Decomposed elements**  
1. There is wooden dining table (core)  
2. There are spoons present (core)  
3. There are plates present (core)  
4. There is a fork present (core)  
5. The spoons are made of wood (core)  
6. The plates are made of wood (core)  
7. The fork is not made of wood (core)  
8. The spoons, plates, and fork are located on the wooden dining table

**Common failure patterns**  
: 7. The fork is not made of wood

Figure 7. **Qualitative examples of recurring misalignments** when generating multiple images from a fixed prompt, with decomposed elements and common failures identified by EFC.



#### Original prompt

: “The glass car window changed into a wooden car window”

#### Decomposed elements

1. A car window is present in the scene (core)
2. The car window initially appears to be made of glass (core)
3. The car window later appears to be made of wood (core)
4. The transformation of the car window occurs sequentially over time (core)
5. The glass car window transforms into a wooden car window (core)

#### Common failure patterns

- : 3. The car window later appears to be made of wood,
- 4. The transformation of the car window occurs sequentially over time,
- 5. The glass car window transforms into a wooden car window



#### Original prompt

: “A person is turning on the desk lamp”

#### Decomposed elements

1. There is a person in the scene (core)
2. There is a desk lamp on the desk (core)
3. The person is positioned near the desk lamp (extra)
4. The desk lamp is initially turned off (core)
5. The desk lamp has a switch that can be activated (core)
6. The person's hand moves towards the desk lamp (core)
7. The person's fingers interact with the lamp switch (core)
8. The desk lamp transitions from being off to on after the person interacts with the switch (core)

#### Common failure patterns

- : 4. The desk lamp is initially turned off,
- 8. The desk lamp transitions from being off to on after the person interacts with the switch

Figure 8. **Qualitative examples of recurring misalignments** when generating multiple videos from a fixed prompt, with decomposed elements and common failures identified by EFC. We illustrate the first and the last frame for each generated video.

## A.2. Details of Element-level Factual Correction (EFC)

We present a detailed overview of the visual verification process in our verifier, Element-level Factual Correction (EFC) in Figure 9.

**Element-level factual correction.** The goal of this process is to provide fine-grained and interpretable feedback on whether each part of a prompt is faithfully realized in the generated visuals. Given a prompt and its corresponding outputs (images or videos), EFC first decomposes the prompt into multiple disjoint semantic elements using a system prompt. For each element, it also constructs an open-ended question, where the element itself serves as the expected answer. Next, EFC verifies the fulfillment of these elements in the generated visuals through factual correction. Instead of relying on visual question answering, our key idea is to perform text-based comparison between the semantic elements and the visuals. To enable this, EFC first extracts captions from the generated visuals and then applies natural language inference (NLI) to classify each element as entailment, neutral, or contradiction.

**Open-Ended visual probing.** For elements classified as neutral, where captions are missing or ambiguous, EFC reuses the previously generated open-ended questions, queries the visual input again, and applies a second NLI step to the corresponding free-form answers, assigning a final label of either entailment or contradiction. Unlike direct QA, this procedure asks open-ended questions and compares their answers with the target elements to determine whether the expected element is present, rather than relying on yes/no responses. This removes affirmation bias inherent in binary QA and avoids providing contextual cues that may cause the verifier to rely on textual hints instead of extracting information directly from the visuals. Through this process, EFC pinpoints which parts of the prompt are faithfully represented and which are contradicted, thereby enabling accurate and interpretable fine-grained feedback for generated visuals.

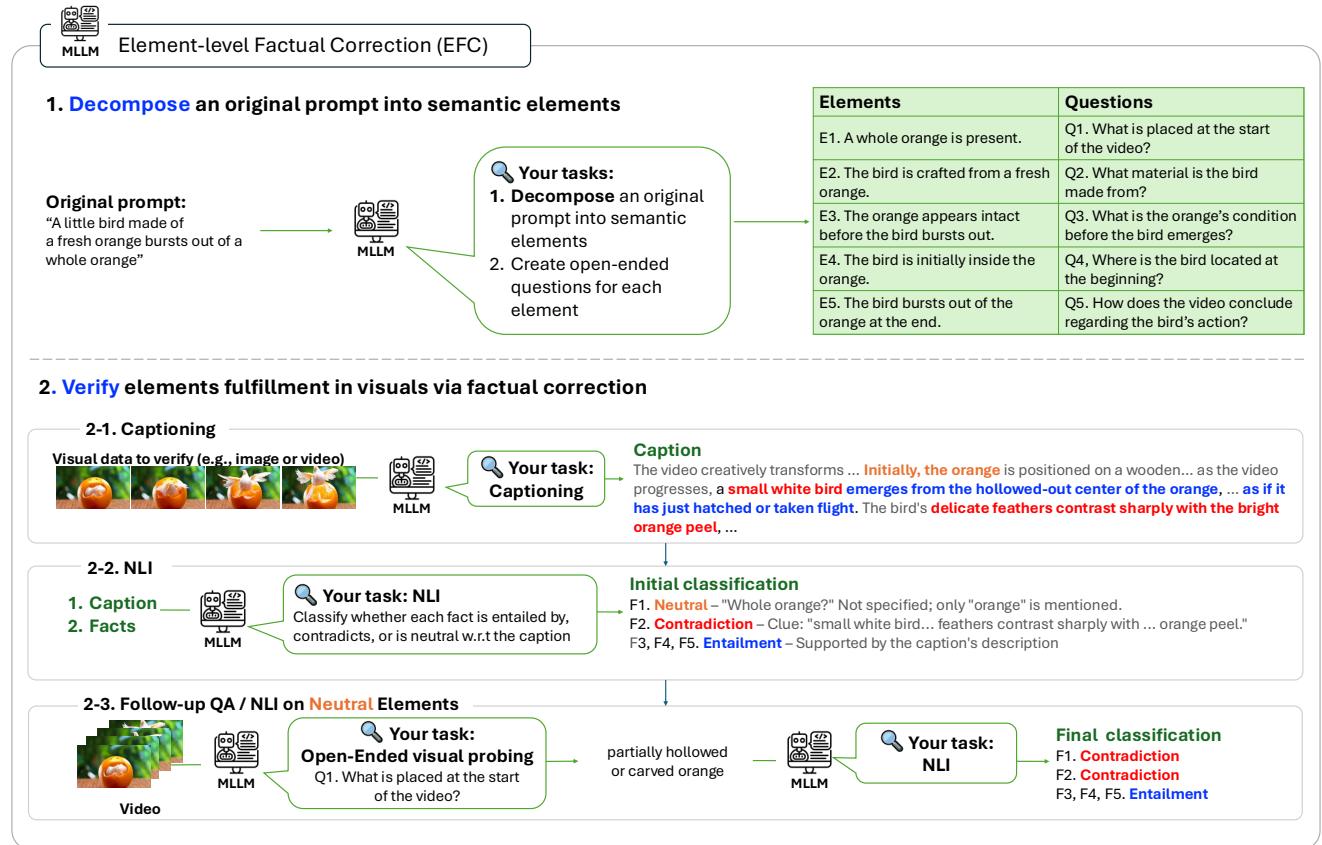


Figure 9. **Illustration of EFC.** The figure illustrates how EFC provides fine-grained, interpretable verification of prompt adherence. It first decomposes the prompt into semantic elements, then generates captions from the visuals, and applies factual correction to classify each element as entailment, neutral, or contradiction. Elements initially labeled neutral (due to missing mentions in the caption) are reevaluated to decide between entailment and contradiction. This design avoids direct QA, leading to more accurate verification.

### A.3. Details on Integration Beyond BoN

This section provides additional details on the integration of our framework with visual scaling methods, complementing Section 4.2.



Figure 10. **Qualitative artifact results with RBF.** RBF alone often generates visuals where the prompt text is directly rendered on the image due to reward over-optimization, whereas combining RBF with our method substantially alleviates this issue.

**Experiments on text-to-image generations.** We integrate our approach with two inference-time scaling methods focused on visuals: DAS [19] and RBF [19]. Following their original experimental protocols, we use SDXL [31] for DAS and Flux.1-schnell [22] for RBF. In both settings, we generate a total of 8 samples, divided into two batches of 4. When combined with our method, the first batch of 4 samples is generated, the prompt is revised, and another 4 samples are generated, ensuring that the total number of function evaluations remains equivalent.

In addition to Table 4 and Figure 4 in the main manuscript, Figures 11 and 12 demonstrate that our integrated results achieve substantially better prompt adherence than visual scaling alone, for DAS and RBF, respectively. This indicates that advanced visual scaling methods can be further enhanced when combined with scaled prompts. It is also noteworthy that scaling visuals alone often leads to undesired outcomes caused by reward over-optimization (see Figure 10). In such cases, the model may even render the textual prompt itself, since these images achieve artificially high reward scores. For example, Figure 10 shows that RBF frequently generates images where the prompt text is printed directly. By contrast, our method mitigates this issue: the revised prompt guides the generator, while the original prompt is used only for the reward signal. This separation effectively reduces over-optimization artifacts and yields more faithful generations, even when PRIS is combined with RBF.

*"A kitchen with a **larger quantity** of milk than juice."*



*"A tissue pack shows two cartoon characters: **one in a red dress** on the left, **one without** on the right."*



*"**Four cupcakes** with sprinkles on a plate with **two forks**."*



*"In an early morning park, **a man** in a grey and white tracksuit is **not running**."*



*"A **child** not building a sandcastle at the beach."*



*"A **woman in a wheelchair** is taller than the **boy** next to her."*



SDXL

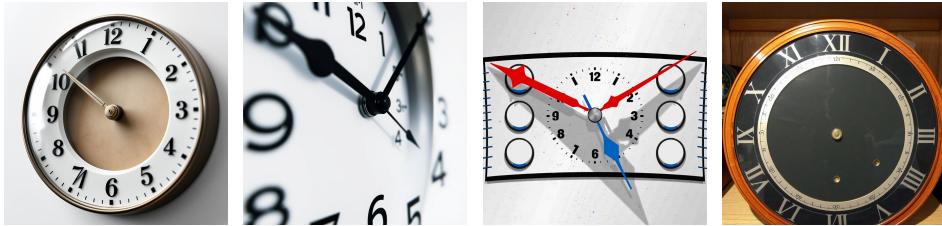
BoN

SMC

SMC with PRIS

Figure 11. Qualitative comparisons when integrating our method with SMC under the same compute budget. Our approach more faithfully follows the prompt, effectively enabling SMC to scale visuals.

*"A clock with no hands to tell the time."*



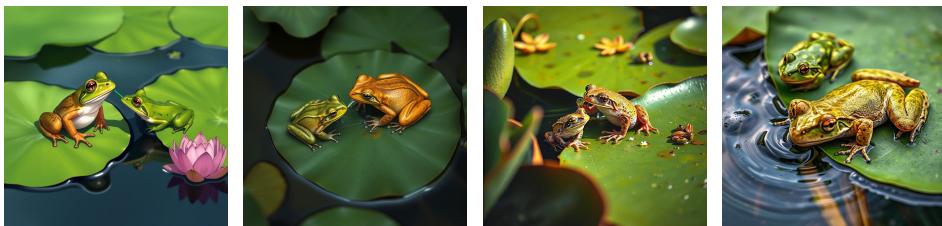
*"A shoe rack without any red pairs of shoes on it."*



*"There is a large fish aquarium in the center of the luxurious living room, but there are no fish in it."*



*"Two frogs on a lotus leaf in a pond, and the one who is drinking is in front of the one who is not."*



*"Four roses in a clear glass vase, all of which are red, and all of which are not open."*



*"A teddy dog and a Persian cat watch a burning table, with the teddy dog at a farther distance."*



FLUX.1-schnell

BoN

RBF

RBF with PRIS

Figure 12. Qualitative comparisons of RBF integrated with our method under the same compute budget. Our method adheres more closely to the prompt and further improves RBF's visual scaling.

**Experiments on text-to-video generations.** We integrate our approach with EvoSearch [12], following its original setup on Wan2.1-1.3B. EvoSearch uses an evolution schedule of [5, 20, 30, 45] and a population schedule of [10, 5, 5, 5], totaling 2,000 NFEs. For integration, we first generate 10 samples with 50 steps (1,000 NFEs), then allocate the remaining 940 NFEs with [5, 30] for the evolution schedule and [5, 4] for the population schedule, resulting in 60 fewer NFEs than EvoSearch. As in the main manuscript, we evaluate on VBench2.0 with VideoAlign as the guiding reward.

Table 8 and Figure 13 present the quantitative and qualitative results. Unlike EvoSearch, which was evaluated on relatively simple prompts, our experiments employ more complex ones. In this setting, EvoSearch scores degrade after scaling, suggesting limited generalization to the unseen reward of VBench2.0. By contrast, when integrated with our method, it achieves improved average scores on VBench2.0.

Table 8. **Quantitative T2V results on VBench2.0, comparing EvoSearch alone with EvoSearch integrated with PRIS.** EvoSearch fails to generalize to unseen rewards, whereas integration with PRIS improves performance.

Method	Motion Rationality	Motion Order Understanding	Dynamic Attribute	Average
Wan2.1-1.3B	38.10	<b>52.87</b>	46.67	45.88
EvoSearch	32.14	51.72	43.33	42.20 ↓ -3.68
EvoSearch + PRIS	<b>53.57</b>	48.28	<b>60.00</b>	<b>53.95 ↑ +8.07</b>

**Dynamic attributes :** “A butterfly’s wing change from yellow to white.”



**Motion rationality:** “A person is opening the window.”



Figure 13. **Qualitative examples comparing EvoSearch and EvoSearch+PRIS.** In the first case, EvoSearch fails to change the butterfly’s wing color despite scaling, whereas our method succeeds. In the second case, EvoSearch depicts the window as already open before the person attempts to open it, while our method correctly shows the window opening as the person reaches out.

#### A.4. Detailed Computational Time Analysis

In this section, we provide a detailed breakdown of verification and generation time, complementing Section 4.4. All measurements are conducted on a single NVIDIA H100 80GB GPU. For images, generating a single sample resolution (1024, 1024) with Flux.1-dev takes on average 13 seconds, while verification with our verifier, EFC, requires 41 seconds. This implies that each verification is computationally equivalent to generating approximately three additional images. To balance this overhead, we set the number of function evaluations (NFE) to 4000 for BoN and 1000 for our method, corresponding to 40 and 10 images, respectively (with 50 sampling steps and classifier-free guidance). For videos, generating an 81-frame sequence at resolution (480, 832) with Wan2.1-1.3B requires 105 seconds on average, while verification takes 100 seconds, approximately equivalent to one additional video generation. Accordingly, we set the NFE to 4000 for BoN (40 videos) and 2000 for our method (20 videos), again under 50 sampling steps with classifier-free guidance.

Our verifier is intentionally built on a pretrained MLLM without task-specific optimization, demonstrating that strong results can be achieved without additional training. Nonetheless, fine-tuning the base MLLM remains a promising direction for reducing verification time and improving efficiency.

## A.5. Comparison with ReflectionFlow

We compare our method with ReflectionFlow [45], which relies on a trained reflection model to iteratively edit each generated sample. Our approach differs in three fundamental ways. First, we revise the prompt itself based on common failure patterns across samples, rather than reacting to individual errors. Second, our method is entirely training-free and does not rely on any auxiliary editing models. Third, it is applicable to any text-conditioned generator, whereas ReflectionFlow requires model-specific training. For a favorable comparison, we allocate 3840 NFEs ( $N = 64$ ) to ReflectionFlow, following its default configuration, while ours uses only 2000 NFEs ( $N = 20$ ). Even under this compute-advantaged setup for ReflectionFlow, our method consistently produces more accurate and semantically aligned results, as shown in Figure 14. PRIS outperforms ReflectionFlow across diverse categories, including comparison, counting, attributes, and negation, highlighting the advantage of correcting prompts based on shared failure modes rather than relying on per-sample post-hoc edits.



Figure 14. **Qualitative comparisons with ReflectionFlow.** Despite being training-free, our method markedly outperforms the learned approach, underscoring the effectiveness of *correcting prompts using shared failure patterns across samples*.

## A.6. Prompt Transferability and Future Work



Original Prompt



Revised Prompt

**Original Prompt:** In a classroom, the clock is not on the wall

**Revised Prompt:** In a classroom, **the clock is placed on a polished wooden desk**, its round face softly illuminated, while the walls remain unadorned, free of any other timepieces.



Original Prompt



Revised Prompt

**Original Prompt:** A little boy with a ping pong paddle looks more excited than a little girl without one.

**Revised Prompt:** A young boy holding a bright yellow ping pong paddle beams with **wide eyes and an open smile**, while nearby, a **calm little girl** gazes at him with a curious expression, **her hands resting by her side**.

Figure 15. **Qualitative example of prompt transferability.** Prompts revised for Flux1.dev are applied to Firefly Image 4 Ultra. By clarifying vague instructions, specifying object presence and absence, and reinforcing contextual cues, the revised prompts yield visuals with stronger adherence compared to those generated from the original prompts.

We observe that our revised prompts are not only effective for the original generator but also transferable to other models, demonstrating their generalizability. This stems from the fact that our revisions resolve ambiguities in the original prompts, making them more precise and robust. Although different generators may specialize in certain aspects, such as producing fine-grained details or maintaining object counts, they often exhibit overlapping weaknesses. Addressing these weaknesses through prompt revision thus benefits multiple models simultaneously.

Figure 15 illustrates this transferability. The prompts originally revised for Flux1.dev are successfully applied to Firefly Image 4 Ultra. For example, the revised prompts clarify vague or underspecified instructions (e.g., replacing “not on the wall” with “the clock is placed on a polished wooden desk”), making object presence and absence explicit (e.g., reformulating “the girl is without a ping pong paddle” into “her hands resting by her side”), and reinforcing contextual cues. These findings suggest a promising research direction: fine-tuning LLMs or other prompt-rewriting systems on pairs of naïve user-provided prompts and failure-focused revisions. By learning systematic transformations from short, underspecified, and loosely written prompts into precise, detailed, and effective ones, rather than relying on random expansions, such models could reduce verification costs and inference-time overhead, accelerating the discovery of high-quality prompts from the outset.

## A.7. Future work and limitations.

Our core idea—identifying shared failure patterns with high precision and addressing them through targeted prompt revision—is broadly applicable to any text-conditioned generative model. We believe extending this idea to other modalities and tasks is a compelling direction for future research, potentially challenging existing inference-time scaling laws. In addition, our benchmark provides a new avenue for evaluating verifiers at the attribute level. We also find that prompts refined on one model often generalize well to others (see Appendix A.6). This observation suggests a promising direction: fine-tuning LLMs or other prompt-rewriting models using paired data consisting of randomly expanded prompts and their failure-focused revisions. Such training resources could reduce verification overhead and inference-time costs, enabling more efficient discovery of high-quality prompts from the start.

## B. Benchmark Construction and Evaluations

### B.1. Benchmark Category

**Details about benchmark constructions.** Existing visual evaluation datasets are mostly limited to human-preference annotations. While useful for coarse quality assessment, such datasets are insufficient for our focus: selecting the best-aligned videos from among multiple misaligned candidates, which lies at the core of inference-time scaling. To address this limitation, we construct a new benchmark explicitly tailored for inference-time scaling and use it to evaluate our verifier, its ablations, and existing baselines. Beyond serving as a testbed for our study, this benchmark also provides a valuable resource for future research on visual prompt-adherence verification.

In our benchmark, each prompt is paired with multiple generated videos, with at least one ground-truth (GT) aligned reference and others containing slight misalignments, thereby forming a mid-quality candidate pool. In total, the benchmark comprises 410 prompts. We collect prompts showcased in demos of both popular open-source [35] and closed-source video models [8, 20], and categorize them into two broad groups: motion (120 prompts) and physics (144 prompts). To further enrich the evaluation, we also adopt prompts from VBench 2.0, spanning three fine-grained motion-related categories: dynamic attributes (47 prompts), motion order (68 prompts), and motion rationality (31 prompts). For each prompt, we generate videos using multiple text-to-video models [8, 20, 35] as well as image-to-video models [35], ensuring the inclusion of both GT-aligned and misaligned outputs. Each video is independently annotated by three human evaluators as GT or non-GT, and the final label is assigned by majority vote.

**Detailed analysis of verifiers on our benchmark per category.** In addition to the overall accuracy reported in Table 5 of the main manuscript, we present per-category accuracy in Table 9. As the results show, EFC consistently achieves the highest accuracy across all categories. Compared to the decomposed binary VQA baseline, which shares our decomposition strategy but replaces our text-to-text verification with binary VQA, EFC yields a substantial performance gain, underscoring the advantage of our text-based approach over visual QA methods. When compared to learned reward models (i.e., MLLM-based verifiers fine-tuned on human-preference datasets), including VideoAlign (the strongest among them and used as our tie-breaker), EFC still maintains a significant lead. Notably, it achieves this performance without any additional training on preference datasets, but rather through a systematic zero-shot verification process. Furthermore, we attribute this gap to the fact that reward models are typically trained on human-preference data, where subtle aspects such as frame quality, motion smoothness, or stylistic biases often dominate judgments, even when they are not directly related to prompt adherence. In contrast, EFC focuses explicitly on verifying semantic alignment with the prompt, making it both more accurate and interpretable.

Table 9. Quantitative results of verifier accuracy per prompt category on our constructed dataset. **Bold** indicates the best result.

Method	Motion	Physics	Dynamic Attributes	Motion Rationality	Motion Order Understanding	Average
VisionReward [41]	0.650	0.569	0.319	0.662	0.452	0.571
UnifiedReward [39]	0.492	0.507	0.298	0.588	0.581	0.498
VideoAlign [28]	<b>0.792</b>	0.660	0.511	0.794	0.516	0.693
Decomposed binary VQA	0.733	0.667	0.617	0.809	0.613	0.700
PRIS (Ours)	<b>0.792</b>	<b>0.764</b>	<b>0.638</b>	<b>0.838</b>	<b>0.677</b>	<b>0.763</b>

While our study focuses on prompt-adherence verification, we believe that our verification framework can be extended to other important axes of evaluation, such as motion quality, NSFW filtering, and bias detection, by replacing prompt decomposition with task-specific decomposition strategies. This flexibility offers promising directions for future research.

## C. Experiments Details

### C.1. Detailed Setup

For GenAI-Bench, since many prompts within the same categories (e.g., counting, differentiation, comparison, negation, universal) are similar but differ only in objects, we randomly subsample 20% to reduce redundancy. For selecting  $k$ , we set  $k = N//4$ , as  $N//2$  samples are first generated for review before prompt revision, and half of them are used as top-performing seeds.

### C.2. Base Model Selection

To ensure that our study focuses on the effect of prompt redesign in inference-time scaling, we first measure the degree of prompt adherence across candidate leading open-source video models such as Wan, LTX, and Hunyuan. This step is necessary because if a model fails to follow the prompt at all, there is little need to apply prompt redesign. Specifically, we compute the text embedding similarity between the original prompt and the generated video caption. We use Qwen-32B for captioning and employ the SentenceTransformer model (intfloat/e5-mistral-7b-instruct) to measure embedding similarity. We present the similarity score in Table 10.

Table 10. **Quantitative results of prompt adherence** across different text-to-video models, used to exclude base models with poor alignment and retain only those with acceptable adherence.

Metric	Method	Motion Rationality	Motion Order Understanding	Dynamic Attribute	Average
VideoAlign	LTX	0.764	-0.153	-0.977	-0.122
	Hunyuan	0.904	0.212	-0.775	0.114
	Wan	<b>1.475</b>	<b>0.940</b>	<b>-0.397</b>	<b>0.673</b>
Text Similarity	LTX	0.635	0.642	0.600	0.626
	Hunyuan	0.678	0.671	0.616	0.655
	Wan	<b>0.717</b>	<b>0.702</b>	<b>0.631</b>	<b>0.683</b>

Based on this analysis, we selected Wan as our primary video model, since it demonstrates a reasonable level of prompt adherence while leaving room for improvement through verification and redesign. In contrast, models such as LTX and Hunyuan were excluded, as their low adherence made them unsuitable for evaluating prompt redesign at inference-time scaling, particularly on complex prompts in VBench2.0 that involve status changes or multiple consecutive events within a single video.

## D. Additional Qualitative Experimental Results

### D.1. Text-to-Image Generation

We provide additional qualitative results beyond Figure 3, demonstrating that our prompt redesign improves coherence of the final visual outputs under the same NFE budget (2000, as in the main experiments). As shown in Figure 16, which compares the top-scoring outputs generated from the original GenAI-Bench prompts, our method performs particularly well on prompt sets containing ambiguous attributes, numerical specifications, or subtle constraints (e.g., “without,” “greater variety”), effectively elaborating them into more faithful visual realizations than baselines.

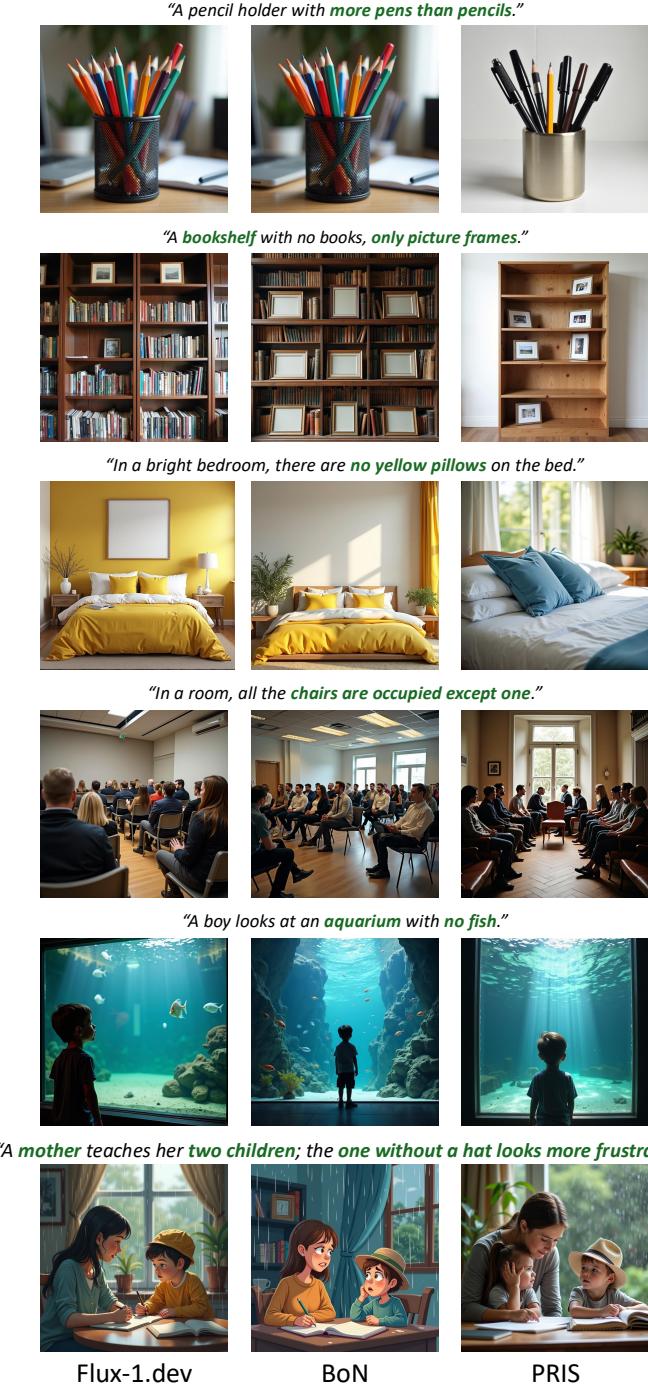


Figure 16. **Qualitative comparisons on T2I generation** where visual generation is (initially) conditioned on the original prompts.

We also compare with standard prompt expansion in Figure 17, where ours achieves substantially higher prompt fidelity compared to the baselines. Unlike standard prompt expansion, which cannot target or identify the most challenging semantic elements, our joint scaling of visuals and prompts more faithfully preserves the intended semantics by adaptively revising the prompt based on recurring failure modes.

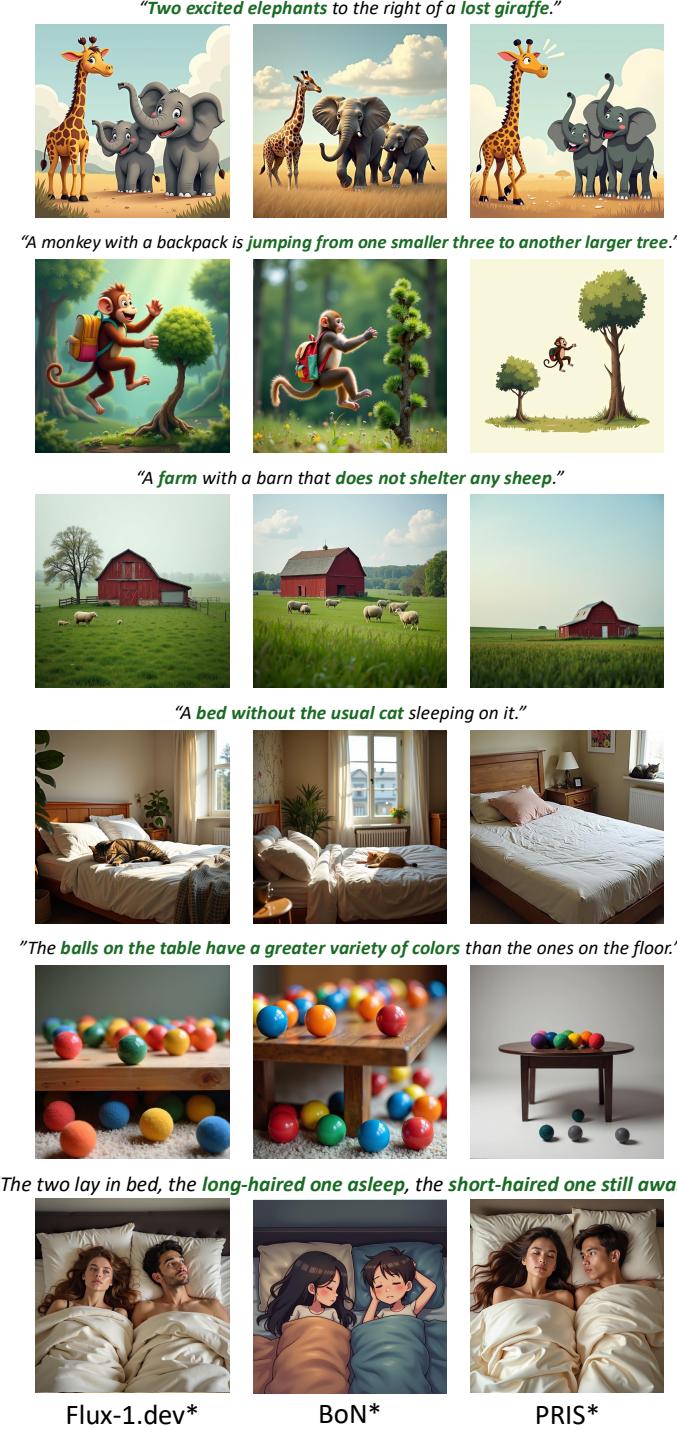


Figure 17. **Qualitative comparisons on T2I geneation** where visual generation is (initially) conditioned on standard prompt expansion.

## D.2. Text-to-Video Generation

In addition to Figure 4, we present additional qualitative top-scoring examples in Figure 18. As shown, our method more faithfully follows the intent of the original prompt. The final top-scoring visuals generated with our PRIS demonstrate significantly stronger prompt adherence compared to baselines. Specifically, BoN often misses key events or produces unnatural temporal order. For example, it may depict only a single motion (e.g., morphing without differentiating “cleaning the kitchen” in the 1st visual) or assign different motions to different people (in the 4th visual). BoN also frequently fails to capture dynamic changes, generating only static states (3rd and 6th visuals). Furthermore, BoN often does not correctly realize sequential actions, such as repeatedly attempting to break chocolate pieces, whereas our method generates coherent sequences where the person both attempts the action and displays the broken pieces (5th visual).

## D.3. More Visualizations

We include an HTML file to the attached zip file. To explore the generated visuals and comparisons with baselines alongside their corresponding prompts, please open `visuals/index.html` in a Chrome browser (This file is located in the `visuals` directory within the attached zip file). This visualizes the generated visuals, including images and videos, in the `visuals/resources` folder.

**Motion order understanding:** “A person is cleaning the kitchen, then they suddenly start mopping the floor.”



**Camera motion:** “Garden, pan-left”



**Dynamic attributes:** “The glass car window changed into a wooden car window.”



**Motion order understanding:** “A person is working on a project, and then suddenly starts cooking dinner.”



**Motion rationality:** “A person is breaking a chocolate bar into pieces.”



**Dynamic attributes:** “The moon changes from silver to yellow.”



Figure 18. **Qualitative comparisons on T2V generation** where visual generation is (initially) conditioned on standard prompt expansion, with Wan2.1-1.3B (top) and Wan2.1-14B (bottom).