Seeing is Believing: Faithful Text-to-Image Synthesis via Cross-Modal Cognitive Prompt Rewriting

Anonymous ACL submission

Abstract

Despite the impressive capabilities of text-to-image (T2I) models, an intent–generation gap often persists due to the brevity and ambiguity of user prompts. Existing approaches primarily polish the prompt for fluency and readability. However, the enhancement process still lacks visual grounding. As a result, the rewriter may over-infer missing details, causing an intent-generation gap. To address this limitation, we propose FaithRewriter, a novel prompt-enhancement framework for T2I generation. Specifically, FaithRewriter first leverages a multimodal MLLM to generate an image from the original prompt as an intermediate visual cue. This cue is then combined with the prompt and fed into a large-scale LLM to produce visually grounded augmentations that better reflect how the intended content should appear in images. Finally, these augmentations are distilled into a small-scale LLM for efficient deployment, enhancing its ability to generate effective T2I prompts. Experiments show that FaithRewriter yields prompts that are more faithful to the user intent and more visually plausible than strong baselines, helping narrow the intent–generation gap.

1 Introduction

Recent advances in text-to-image (T2I) synthesis have revolutionized content creation, enabling the generation of high-fidelity images from natural language (Esser et al., 2024; Han et al., 2024; Hao et al., 2023). However, a significant bottleneck persists: the inherent ambiguity and brevity of user prompts often lead to a substantial *intent–generation gap*. To bridge this gap, prompt rewriting has become a standard practice, where Large Language Models (LLMs) are employed to expand concise user inputs (T_0) into detailed descriptions (T_1) (Wang et al., 2025; Cao et al., 2023).

While these methods improve linguistic fluency, we argue that such *blind expansion* does not guarantee better generation. As illustrated in Figure 1, standard LLM rewriting operates purely in the textual domain and cannot verify the visual consequences of the added details. As a result, the expanded prompt T_1 may introduce "textual hallucinations": descriptions that are linguistically plausible but visually or physically incorrect. For example, when rewriting "a glass of water turned upside down," a text-only model may describe the water as remaining suspended, which reads naturally in text but violates basic physical reality in the generated image. In this sense, T_1 can pollute the original intent with visual noise, making the output less faithful than directly using the concise T_0. Therefore, the goal is not simply to lengthen the prompt, but to ensure that any added detail is visually grounded and physically realizable.

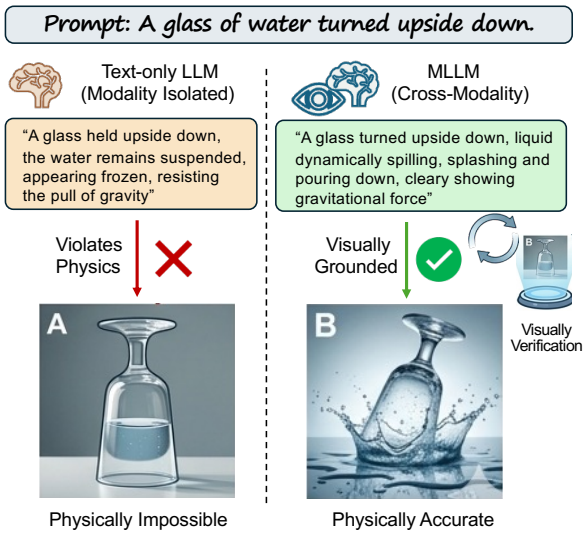

Figure 1: Comparison between standard rewriting and FaithRewriter. Standard rewriting (T_1) often introduces hallucinations due to modal isolation. FaithRewriter utilizes a Visual Anchor to diagnose these errors and produces a verified revision (T_2), which is then distilled into a text-only model via DPO.

From a cognitive perspective, this limitation arises because standard LLMs lack the mechanism

of cross-modal simulation. Dual Coding Theory (Paivio, 1991) suggests that when humans elaborate on a description, they do not merely predict the next word; they construct a mental image, or the "mind's eye," using internal visual cues to verify spatial logic and physical feasibility before speaking. Standard rewriters, lacking this visual feedback loop, operate like an artist trying to paint a complex scene with their eyes closed.

To address this, we propose FaithRewriter, a novel framework that bridges the modal gap via a "Simulate-Then-Distill" paradigm. Our key insight is that while equipping a rewriter with online visual perception is computationally prohibitive, we can leverage a Multimodal LLM (MLLM) as an offline cognitive simulator to teach a lightweight text-only model. Specifically, FaithRewriter operates in two main stages. First, we introduce a simulation-based data construction phase to ensure data quality. For a given user prompt (T_0), we generate an initial rewriting (T_1) and render it into an intermediate image, referred to as a *Visual Anchor*. This anchor serves to externalize the latent hallucinations of T_1. An MLLM then acts as a cognitive agent to diagnose the discrepancies between the user's intent and the visual anchor, generating a grounded revision (T_2). Crucially, unlike T_1 which is a probabilistic guess, T_2 is a corrective revision derived from observing specific visual failures. This ensures that T_2 is strictly superior to T_1 in terms of visual faithfulness.

Second, we perform a two-stage knowledge distillation to transfer the MLLM's visual reasoning capabilities into a compact student LLM. We begin with Supervised Fine-Tuning (SFT) to align the student with the syntax of visual descriptions, followed by Direct Preference Optimization (DPO). Since we have established that T_2 is visually verified while T_1 contains likely hallucinations, we use the pair $(T_2 \succ T_1)$ to suppress *hard negatives*. This sharpens the model's decision boundary, enabling it to reject visually ungrounded descriptions without the need for expensive online reinforcement learning. Extensive experiments demonstrate that FaithRewriter significantly outperforms state-of-the-art baselines in generating faithful and visually plausible prompts, effectively narrowing the intent–generation gap.

The main contributions of this paper are summarized as follows:

- We identify that blind textual expansion often degrades prompt faithfulness due to modal isolation, and propose the Simulate-Then-Distill paradigm as a solution.
- We introduce FaithRewriter, a framework that utilizes Visual Anchors to expose the hallucinations in initial rewrites (T_1) and constructs superior, verified revisions (T_2) through MLLM-driven diagnosis.
- We propose a robust SFT+DPO distillation strategy that enables a lightweight student model to internalize cross-modal priors, effectively learning to reject visually ungrounded descriptions.

2 Related Work

2.1 Prompt Rewriting

Prompt rewriting aims to refine initial instructions for large models to enhance their performance on downstream tasks, thereby reducing reliance on manual crafting (Wang et al., 2025; Cao et al., 2023; Kong et al., 2024; Srivastava et al., 2024; Zhou et al., 2025; Li et al., 2024; Mañas et al., 2024). Existing approaches often treat discrete prompts as parameters to be optimized. For instance, PRewrite (Kong et al., 2024) trains a rewriter LLM using rewards derived from the downstream task performance, allowing the model to explore and identify more effective prompts. Similarly, (Zhou et al., 2025) employs a meta-LLM iteratively to revise the prompt for a given test instance based on the task LLM's output. (Li et al., 2024) proposes to use a powerful LLM to generate rewritten prompts and then distilling this capability into a compact model. Furthermore, works like PromptEnhancer(Wang et al., 2025) focus on generating task-specific or personalized prompts, often for NLP tasks like text classification or personalized email generation. However, they primarily operate within the *textual* modality, lacking an explicit mechanism to ensure the *visual plausibility* of the generated text, which is crucial for text-to-image (T2I) synthesis.

2.2 Text-to-Image Synthesis

The field of text-to-image (T2I) synthesis (Esser et al., 2024; Han et al., 2024; Hao et al., 2023; Podell et al., 2023; Tao et al., 2023) has advanced significantly, with models like HunyuanImage and Qwen-Image achieving high fidelity through advanced architectures. However, research has primarily focused on enhancing the model's inter-

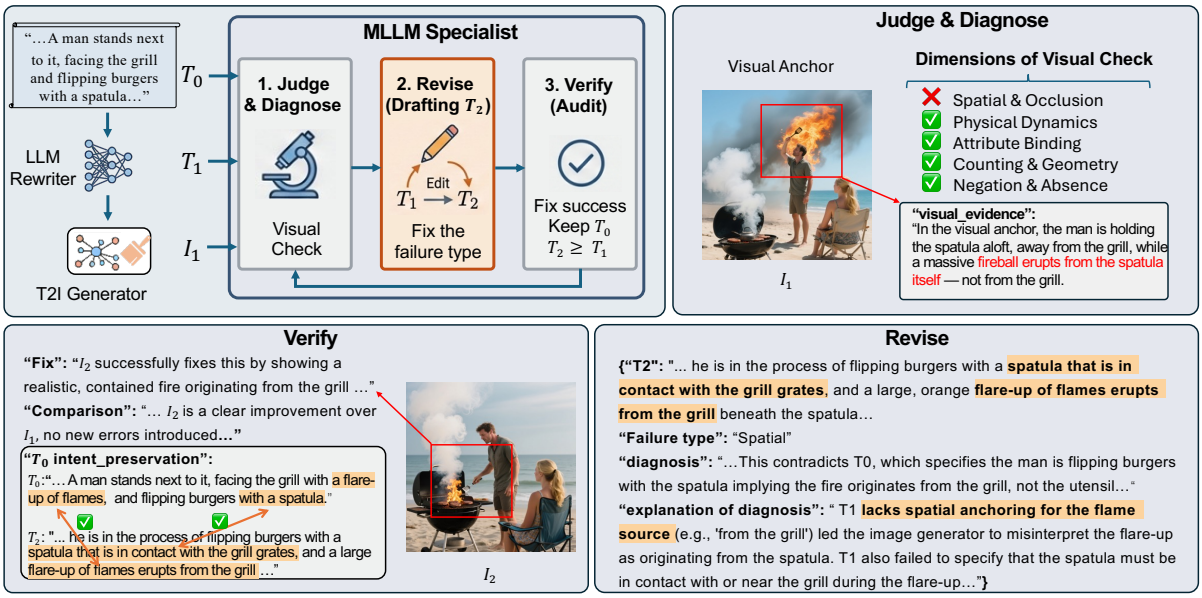

Figure 2: Overview of the FaithRewriter pipeline. Given the original prompt T_0, a text-only rewriter first produces an expanded prompt T_1, which is used to generate an image I_1. An MLLM specialist then diagnoses the visual failure in I_1, revises T_1 into a grounded prompt T_2, and verifies that the new image I_2 fixes the error while preserving the original intent. We omit the DPO training for simplicity.

nal capabilities, often overlooking a critical external factor: the quality of the input prompt itself. The prompt serves as the sole bridge between human intent and the generative process. Ambiguous or under-specified prompts create a bottleneck forcing even powerful models to guess details. This limitation is evident in benchmarks like T2I-CompBench (Huang et al., 2023), which show models struggle with complex object relations and negations—precisely where human descriptions are most prone to ambiguity. Addressing this requires shifting focus from making models understand vague language better to making language clearer for models. Recent work explores this interface: Hunyuan Enhancer trains an LLM to rewrite prompts, improving output alignment, while approaches using rich text or structured formatting provide more explicit signals to guide generation.

3 Method

3.1 Cognitive Insight & Problem Formulation

The Cognitive Mismatch: From "What" to "Where" and "How". Recent text-to-image (T2I) models exhibit strong compositional capabilities in rendering the fundamental *"what"* of a scene (e.g., generating specified objects with basic attributes). However, their limitations become apparent when tasks shift to precisely depicting *"where"* objects are located, *"how"* they are arranged, or *"why"* they interact. From a cognitive perspective, this stems from the fundamental dichotomy between linguistic fluency and visual plausibility. Dual Coding Theory (Paivio, 1991) suggests that accurate description requires *cross-modal simulation*—using an internal visual model to verify physical realizability. General-purpose LLMs, however, are modally isolated: pretrained solely on text, they optimize statistical correlations but lack the perceptual constraints to ground their reasoning.

Formulating the Modal Gap. We formalize this issue as a divergence between two probability distributions over the rewritten prompt y, given an input x. Let $P_{\text{text}}(y|x)$ denote the distribution modeled by text-only LLMs, and $P_{\text{cross}}(y|x)$ represent the ideal distribution of *visually grounded* prompts that yield high-fidelity images. Because textual objectives do not penalize visual impossibility, there exists a subset of hallucinatory prompts where:

$$P_{\text{text}}(y|x) \rightarrow \text{high}, \quad \text{while} \quad P_{\text{cross}}(y|x) \approx 0. \tag{1}$$

For instance, a text model might fluently generate *"A full glass of water held upside down without spilling,"* which is grammatically perfect but violates physical dynamics.

Benchmarking the Gap: Information-Dense Intents & Omni-VQA. To systematically expose this modality gap, we build an automated benchmark engine inspired by recent spatial intelligence evalu-

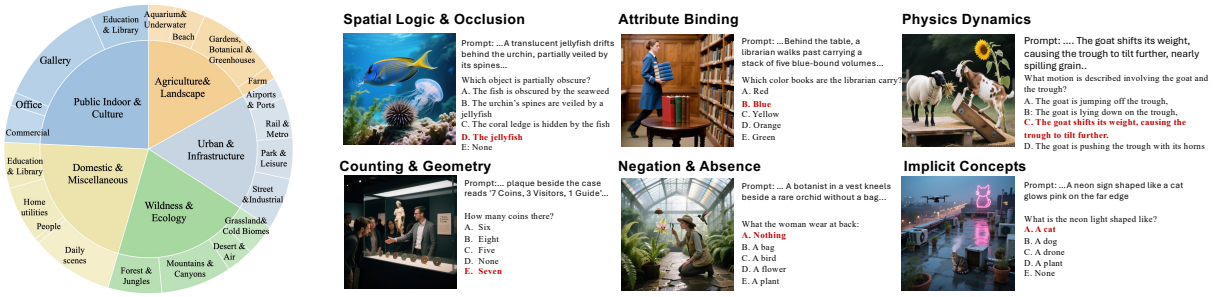

Figure 3: Left: Scene distribution of the proposed FaithT2I benchmark. Right: An illustrative example of the prompt–question–answer pairs used for generated-image evaluation.

ations (Wang et al., 2026). Specifically, we prompt an MLLM to synthesize diverse, information-dense user intents (T_0) across real-world scenes (e.g., living room, forest), with each prompt deliberately designed to probe **six common failure modes** of T2I models.

Hierarchical Taxonomy of Spatial Intelligence. To better diagnose generation failures, we reorganize 10 standard VQA sub-categories into a hierarchical taxonomy tailored to T2I evaluation, yielding six macro-categories as shown in Figure 3

For fine-grained diagnosis, each T_0 is paired with 10 omni-dimensional multiple-choice VQA questions covering both overall image faithfulness and these specific failure modes. To avoid forced guessing when an image is severely misaligned, every question includes an "E: None" option.

The "Simulate-Then-Distill" Paradigm. To bridge this benchmarked gap, the rewriter must align with P_{cross}. Since equipping the rewriter with online visual feedback is computationally prohibitive, we propose the **"Simulate-Then-Distill"** paradigm. We leverage an MLLM as an offline cognitive simulator to construct a visually grounded preference dataset based on the aforementioned failure modes, and subsequently distill this capability into a lightweight text-only model.

3.2 Constructing Visually-Grounded Preference Pairs

The cornerstone of our framework is a fully automated, MLLM-driven data engine designed to synthesize high-quality preference pairs $\mathcal{D} = \{(T_0, T_1, T_2)\}$. To ensure the revision (T_2) is strictly superior to the ungrounded baseline (T_1), we design a rigorous three-phase pipeline: Targeted Intent Generation, Structured Diagnosis, and Cycle-Verification.

3.2.1 Phase 1: Targeted Intent Generation

To maximize the learning signal, the initial user intents (T_0) must expose the inherent weaknesses of text-only models. We prompt an MLLM to systematically generate diverse T_0 samples targeting **six critical failure modes** in Figure 3

3.2.2 Phase 2: Visual Anchor and Structured Diagnosis

For each T_0, we generate a baseline rewrite T_1 using a standard LLM, and synthesize its corresponding image I_1 via a frozen T2I model. The image I_1 acts as a **Visual Anchor**, externalizing the latent hallucinations of T_1. We then instruct an MLLM to act as a visual debugger to revise T_1 into T_2 through a strictly controlled reasoning chain:

Confidence Gating: To prevent unnecessary edits and over-correction, the MLLM first performs a strict visual reality check. If I_1 perfectly reflects T_0, the data point is discarded as it lacks a meaningful error to correct. Revisions are triggered *only* if an objective visual failure is detected.

Categorized Diagnosis: The MLLM pinpoints the exact failure in I_1 mapped to our six predefined dimensions (e.g., identifying "Attribute Leakage: the red color of the apple bled into the pear"). It then critiques T_1 to understand which linguistic ambiguities caused the failure.

Intent Preservation Audit : After drafting the revised prompt T_2 with explicit visual constraints, the MLLM is forced to conduct a retrospective audit against the original T_0. This explicitly prevents the model from dropping background details or secondary subjects while fixating on the primary error, ensuring T_2 is a strict superset of T_0's constraints.

3.2.3 Phase 3: Cycle-Verification via Strict MLLM Judge

Even with careful drafting, T_2 might introduce new artifacts. To provide the ultimate quality guarantee,

we generate a second image I_2 using T_2 and deploy the MLLM as a strict, context-aware judge to evaluate the triplet (T_0, I_1, I_2). The judge evaluates three strict boolean criteria:

The Fix Verification: Did I_2 successfully resolve the exact problem identified in the previous diagnosis phase?

Anti-Regression Check: Did I_2 accurately preserve all original details from T_0 without triggering a "whack-a-mole" effect (i.e., fixing one object but breaking the background)?

Final Verdict: Is I_2 strictly superior and more faithful to T_0 than I_1?

Only when all three boolean gates return `True` do we accept the pair, setting $T_{win} = T_2$ and $T_{lose} = T_1$. This rigorous cycle-verification guarantees that T_{lose} represents a true **hard negative** (linguistically fluent but visually flawed) and T_{win} represents a verified, grounded correction.

3.3 Knowledge Distillation via Hard-Negative Preference Optimization

Using the verified dataset $\mathcal{D}$, we transfer the MLLM's cross-modal reasoning into a compact text-only student LLM. While we first initialize the student policy via Supervised Fine-Tuning (SFT) on the positive pairs (T_0, T_{win}) to teach it the "visual dialect" and descriptive syntax, SFT alone is insufficient. Standard maximum likelihood estimation learns what to say, but fails to explicitly penalize plausible-sounding hallucinations. To instill true cross-modal discrimination, we proceed to reinforcement learning via Direct Preference Optimization (DPO).

The Necessity of Hard Negatives. The core innovation of our alignment strategy lies in the construction of the preference pairs. A naive RL approach might formulate the preference as $(T_{win} \succ T_0)$, penalizing the original concise prompt. However, this inadvertently introduces a trivial *length bias*: the model simply learns a "longer is better" heuristic without acquiring visual grounding.

To systematically address this, we leverage the ungrounded baseline rewrite T_{lose} (i.e., T_1 from Phase 2) as the negative sample. Crucially, T_{lose} acts as a **hard negative**: it shares the exact same verbosity, linguistic complexity, and stylistic fluency as the verified revision T_{win}, yet it harbors the specific visual defects externalized by the visual anchor I_1.

Optimization Objective. By pairing $(T_{win} \succ T_{lose})$, the contrastive signal is stripped of length or style confounders, isolating the *spatial and physical logic* as the only differentiating factor. We optimize the student policy π_θ using DPO:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(T_0, T_{\text{win}}, T_{\text{lose}}) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(T_{\text{win}}|T_0)}{\pi_{\text{ref}}(T_{\text{win}}|T_0)} - \beta \log \frac{\pi_\theta(T_{\text{lose}}|T_0)}{\pi_{\text{ref}}(T_{\text{lose}}|T_0)} \right), \right. \quad (2)$$

where π_{ref} is the SFT-initialized model, and β controls the deviation margin. This optimization forces the student to look beyond surface-level language patterns, explicitly unlearning the tendency to generate textual hallucinations and pushing its distribution toward the cross-modal ideal P_{cross}.

3.4 Benchmark build

4 Experiment

4.1 Experiment Setup

Benchmarks. To comprehensively evaluate FaithRewriter, we utilize both our curated dataset and public benchmarks. FaithT2I (Ours): Constructed via the pipeline in Section 3.2, this dataset focuses on hard cases for T2I generation. We curate splits for training (*FaithT2I-3KFaithT2I-10K*, *FaithT2I-100K*) and a held-out test set (*FaithT2I-Test*, 1,150 samples) covering diverse domains including spatial reasoning and complex interactions.

We evaluate on public benchmarks like Spatial-GenEval (Wang et al., 2026), a benchmark to test if models can handle complex spatial prompts. We also evaluate on DPG-Bench (Hu et al., 2024) and GenEval (Ghosh et al., 2023) to verify generalization capabilities.

Baselines & Models. For the rewriter backbone, we experiment with LLaMA3.1-8B, Hunyuan-7B-Instruct, and Qwen3-8B-Base who have similar size. We compare FaithRewriter against : (1) T0 (Raw): Direct user prompts; (2) Zero-shot: directly generate (3) Text-modal: Fine-tuned on text-only rewritten data(T_1) (4) Cross-modal: Fine-tuned on cross-modality rewritten data (T_2) For image generation, we employ Qwen-Image (Wu et al., 2025) as default and compare with Hunyuan-Image 2.1 (Cao et al., 2025), and FLUX.1-dev.

Training and Inference Details. We use Adam (Diederik, 2014) optimizer, learning rate is 1e-4, and generation random seed is 42. All training and inference experiments were conducted on a server equipped with 8 NVIDIA H20 GPUs

Method	VQA Scoring ↑			Textual Alignment Scoring ↑			Preference Ranking ↓			CLIP ↑	FID ↓
	Gemini	GPT4o	Qwen3VL	Gemini	GPT4o	Qwen3VL	GPT4o	Gemini	Qwen3VL		
T0 (raw)	62.12	60.28	64.39	3.303	3.939	3.994	-	-	-	0.752	23.94
LLama3.1-8B											
Zeroshot	55.00	58.23	60.32	3.314	3.421	3.456	2.20	2.22	2.25	0.762	23.76
Text modal	59.12	62.35	64.52	3.229	3.689	3.977	2.09	1.95	2.07	0.774	23.58
Cross-modal	**63.27**	**65.61**	**70.65**	**4.037**	**4.249**	**4.403**	**1.71**	**1.83**	**1.68**	**0.784**	**23.41**
Hunyuan-7B-Instruct											
PromptEnhancer	68.59	72.69	74.43	3.472	3.491	3.483	2.24	**1.89**	2.28	**0.813**	**22.84**
Zeroshot	64.18	65.82	62.94	3.327	3.364	3.381	3.11	3.39	3.07	0.768	23.67
Text modal	66.37	68.15	67.06	3.176	3.863	3.925	2.81	2.77	2.71	0.796	23.25
Cross-modal	**70.83**	**73.41**	**77.94**	**3.949**	**4.001**	**4.164**	**1.84**	1.95	**1.94**	0.808	22.87
Qwen3-8B-Base											
Zeroshot-LLM	59.73	61.34	70.83	4.122	3.745	3.773	2.29	2.38	2.63	0.762	23.76
Text modal	68.78	69.40	75.65	3.263	3.814	3.963	2.19	2.03	1.82	0.828	22.63
Cross-modal	**74.80**	**73.71**	**78.91**	**4.075**	**4.098**	**4.392**	**1.52**	**1.59**	**1.55**	**0.847**	**22.19**

Table 1: Main results on FaithT2I-test.

Method	Overall	Spatial Foundation		Spatial Perception			Spatial Reasoning			Spatial Interaction	
		Object	Attribute	Position	Orientation	Layout	Comparison	Proximity	Occlusion	Motion	Causal
T0 (raw)	63.09	87.16	80.47	69.23	66.24	73.64	24.42	66.81	38.22	65.25	59.42
Zeroshot	64.58	87.55	80.12	71.10	66.80	74.80	27.20	68.20	39.85	68.10	62.10
Text modal	66.23	87.72	80.47	72.79	67.95	76.34	29.97	71.79	40.92	70.23	64.11
Cross-modal	**70.83**	**88.44**	**82.60**	**74.78**	**73.36**	**80.33**	**35.66**	**76.20**	**53.44**	**73.93**	**69.52**

Table 2: Main results on SpatialGenEval. By utilizing visual anchors, our Cross-modal enhancer demonstrates a comprehensive leap over text-only baselines, particularly in difficult sub-domains like *Occlusion* and *Comparison* .

4.2 Main results

4.2.1 Results on FaithT2I-Test

We first evaluate different prompt enhancement methods on the FaithT2I-10K-Test set, with results summarized in Table 1. We consider three types of baselines: **T0 (raw)**, which directly uses the original user prompt; **Zero-shot**, which applies an off-the-shelf LLM for prompt rewriting without task-specific tuning; and **Text modal**, which is fine-tuned only on textual supervision. We then compare them with our **Cross-modal** FaithRewriter across three backbone LLMs, including Llama3.1-8B, Hunyuan-7B-Instruct, and Qwen3-8B-Base.

Overall, FaithRewriter consistently achieves the best performance across nearly all metrics and backbones. Compared with text-only rewriting, the cross-modal variant yields higher VQA scores, stronger textual alignment, better human-preference-style ranking from multiple MLLM judges, and more favorable CLIP/FID results. These improvements are especially consistent across Gemini, GPT-4o, and Qwen3VL evaluators, showing that the gains are not tied to a single judge. At the same time, all rewriting-based methods outperform the raw prompt baseline, confirming the value of prompt enhancement itself, while the comparison between Text modal and Cross-modal further shows that visual-anchor-based supervision provides additional benefits beyond pure textual fine-tuning.

4.2.2 Results on SpatialGenEval

We further evaluate our method on SpatialGenEval, and the results are shown in Table 2. Compared with the raw, zeroshot, and text-only baselines, the cross-modal model achieves the strongest overall performance and improves almost all sub-dimensions of spatial intelligence. The gains are particularly clear in more challenging categories such as *Comparison*, *Occlusion*, *Motion*, and *Causal*, where text-only enhancement remains limited. These results suggest that FaithRewriter not only improves general prompt faithfulness, but is also especially effective for prompts requiring precise spatial reasoning, physical interaction, and visually grounded logic.

4.2.3 Results on DPG-Bench and GenEval

We further evaluate FaithRewriter on two public benchmarks, DPG-Bench and GenEval, to verify its generalizability beyond FaithT2I. As shown in Table 3, our cross-modal variant consistently outperforms both the raw prompt baseline and the text-only rewriting baseline on both benchmarks, yielding around 2% overall improvement. Although this

Enhancer	DPG-Bench						GenEval						
	Global	Entity	Attribute	Relation	Other	Overall	Single Object	Two Object	Counting	Colors	Position	Attribute	Overall
T0 (raw)	91.32	**91.56**	92.02	94.31	92.73	88.32	**0.99**	0.92	0.89	0.88	0.76	0.77	0.87
Text modal	91.85	91.70	92.25	95.12	92.41	89.21	0.98	0.91	0.90	**0.89**	0.77	0.78	0.88
Cross modal	**92.64**	91.48	**93.15**	**96.53**	**93.08**	**90.34**	**0.99**	**0.94**	**0.93**	0.87	**0.80**	**0.79**	**0.89**

Table 3: Main results on public benchmarks (DPG-Bench and GenEval). FaithRewriter (Cross modal) achieves a robust +2% overall improvement on both benchmarks. **Bold** indicates the best performance.

experiment is not our main focus, it provides additional evidence that the gains of FaithRewriter can transfer to general-purpose public benchmarks.

4.3 Ablation Study

4.3.1 Robustness Across T2I Models

We evaluate the universality of FaithRewriter by applying the same enhanced prompts to three diverse T2I models: Flux, Hunyuan Image 2.1, and Qwen-Image. As shown in Table 4, FaithRewriter consistently boosts visual assessment scores across all generators compared to the raw prompts (T0) and intra-modal baselines. Crucially, the performance gains are observed regardless of the base model's scale or architecture. This demonstrates that FaithRewriter learns model-agnostic visual features effectively bridging the intent gap.

4.3.2 Impact of Training Stages

Table 5 isolates the contributions of each stage in our "Simulate-Then-Distill" paradigm.

SFT Efficiency: The SFT-only model achieves a substantial performance leap over the baseline (T0), confirming that distilling the MLLM's "visual syntax" provides a strong initialization.

DPO Refinement: The addition of the DPO stage yields further consistent gains across all judges. This validates the effectiveness of our hard negative mining strategy. By explicitly penalizing the linguistically fluent but visually flawed y_{lose}, DPO sharpens the model's judgment, reducing hallucinations that SFT alone cannot eliminate.

4.3.3 Data Scaling and Generalization

We investigate the scaling laws of our method by training on a 3k, 10k, 100k dataset. As shown in Table 8, the model trained on 100k samples consistently outperforms the 3k and 10k version. The performance gap between the 100K model and the text-only baseline suggests that our cross-modal distillation process benefits significantly from increased data diversity. The results indicate that the advantage of FaithRewriter is not saturated, showing strong potential for further improvement with larger-scale cross-modal simulation.

Enhancer	T2I Model	VQA Scoring ↑		
		Gemini	GPT-4o	Qwen3VL
T0 (raw)	FLUX-dev1	55.12	54.38	58.21
Cross-modal	FLUX-dev1	**67.25**	**65.92**	**70.43**
T0 (raw)	Hunyuan Image	60.15	58.94	62.15
Cross-modal	Hunyuan Image	**72.84**	**71.56**	**76.54**
T0 (raw)	Qwen-Image	62.12	60.28	64.39
Cross-modal	Qwen-Image	**74.80**	**73.71**	**78.91**

Table 4: Ablation study on different Text-to-Image backbones. **Bold** indicates the best performance.

Configuration	Textual Alignment Scoring ↑		
	Gemini	GPT4o	Qwen3VL
T0 (raw)	3.303	3.939	3.994
Cross-modal (SFT)	3.843	3.825	3.879
Cross-modal (SFT+DPO)	**4.075**	**4.098**	**4.392**

Table 5: Ablation on training stages. Both SFT and DPO contribute to the final performance.

4.4 User Study

Setting. We conduct a large-scale user study with 57 participants and 60 samples from generated data. Given the original prompt and outputs from three configurations (Zeroshot, Text modal, and Cross-modal), participants rank the images according to how faithfully they satisfy the prompt. The results in Table 7 show a clear preference for Cross-modal, which achieves the best overall mean ranking.

VLM-Human Correlation. Taking Qwen-VL as an example, we measure the correlation between VLM preference scores and averaged human rankings across the study samples, obtaining a Pearson correlation of 0.76. This suggests that VLM judgments are reasonably aligned with human.

4.5 Qualitative Case Analysis

To illustrate how FaithRewriter bridges the modal gap and mitigates textual hallucinations, we present qualitative comparisons in Figure 4. We compare images generated from the raw prompt (I_0), the text-only rewritten prompt (I_1), and our

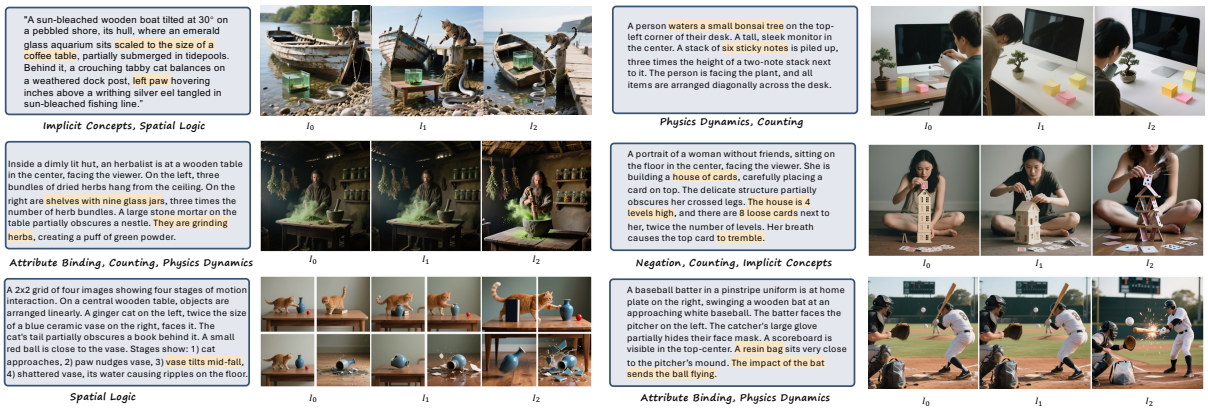

Figure 4: Examplar qualitative results of diverse dimensions, which shows that I_2 maintain the prompt's intent and successfully correct the errors in I_1

Configuration	Textual Alignment Scoring ↑		
	Gemini	GPT-4o	Qwen3VL
T0raw)	3.303	3.939	3.994
Cross-modal (3k)	3.712	3.924	4.105
Cross-modal (10k)	4.075	4.098	**4.392**
Cross-modal (100k)	**4.150**	**4.123**	4.241

Table 6: Ablation study on training data scaling. Performance gains correlate with the size of training data.

Configuration	User Preference Score ↓			Overall ↓	Rank
	LLaMA	Hunyuan	Qwen		
Zeroshot	2.58	2.62	2.55	2.58	3
Text modal	2.15	2.08	2.12	2.12	2
Cross-modal	**1.27**	**1.30**	**1.33**	**1.30**	1

Table 7: User study results. Cross-modal consistently receives the highest human preference.

FaithRewriter prompt (I_2) on several challenging spatial and logical cases.

Resolving Semantic Ambiguity and Implicit Concepts. Text-only rewriters often drift when handling metaphorical or implicit concepts. In the "House of cards" case (Figure 4, Middle-Right), the prompt refers to a structure made of playing cards, but both I_0 and I_1 misinterpret it as a miniature wooden house. In contrast, FaithRewriter (I_2) correctly renders a four-level card structure. A similar issue appears in the "Boat" case (Top-Left), where the aquarium is described as *"scaled to the size of a coffee table"*. The text-only baseline I_1 hallucinates a coffee table, while I_2 preserves the intended scale without introducing unwanted objects.

Enforcing Quantitative and Geometric Constraints. T2I models often fail on precise counting and relative geometry. In the "Sticky notes" case (Top-Right), the prompt requires one stack of six notes to be *"three times the height"* of a two-note stack. Both I_0 and I_1 miss this relation, whereas I_2 satisfies the count and height constraints. In the "Motion Grid" case (Bottom-Left), the prompt requires a strict 2×2 layout showing four sequential stages. Only I_2 follows this structure faithfully, while I_1 produces a disorganized composition.

Restoring Physical Dynamics and Causality. Text-only rewriters may enrich action descriptions linguistically but still fail to induce correct visual dynamics. In the "Baseball" case (Bottom-Right), I_0 and I_1 show a mostly static scene and miss the *"impact of the bat"*. By contrast, I_2 depicts a more dynamic impact with visible physical consequences. This also appears in the "Herb grinding" case (Middle-Left), where only I_2 correctly visualizes the *"puff of green powder"*.

5 Conclusion

In this work, we identified the inherent limitation of text-only prompt rewriting methods, which often produce visually implausible descriptions due to a lack of cross-modal grounding. To address this, we introduced FaithRewriter, a novel framework built on a "Simulate-Then-Distill" (STD) paradigm. FaithRewriter leverages a powerful MLLM as a cognitive simulator to generate visually-grounded prompt enhancements and then efficiently distills this capability into a compact, text-only LLM. And we also propose a corresponding benchmark FaithT2I-test. Extensive experimental results demonstrate that our approach significantly outperforms existing state-of-the-art methods in generating faithful and realistic prompts, effectively bridging the critical intent-generation gap in text-to-image synthesis.

Limitations

Offline Computational Overhead. A primary limitation of our framework lies in the computational cost of the data construction phase. The "Simulate" step requires rendering intermediate images via a diffusion model and invoking a large-scale Multimodal LLM (MLLM) for structured diagnosis and cycle-verification. Consequently, scaling the training dataset (e.g., to millions of pairs) is resource-intensive and time-consuming. However, it is crucial to note that this overhead is strictly confined to the offline training phase. Thanks to our distillation strategy, the deployed student model operates entirely in the text modality, ensuring high inference-time efficiency without requiring visual encoders or online rendering.

Boundaries of the Native Generative Prior. The effectiveness of any prompt rewriting method, including FaithRewriter, is inherently bounded by the intrinsic capabilities of the downstream T2I generator. If the base generative model entirely lacks the prior knowledge of a specific concept (e.g., an extremely rare object) or suffers from hard architectural bottlenecks (e.g., accurately rendering illegible text or generating exactly 15 identical objects), an enhanced prompt cannot magically inject these missing capabilities. While FaithRewriter expertly reorganizes instructions to maximize the generator's potential, it cannot overcome fundamental rendering deficits.

References

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, and 1 others. 2025. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. *arXiv preprint arXiv:2311.06752*.

Kingma Diederik. 2014. Adam: A method for stochastic optimization. *(No Title)*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and 1 others. 2024. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

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

Weize Kong, Spurthi Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Prewrite: Prompt rewriting with reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–601.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, pages 3367–3378.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. 2024. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Saurabh Srivastava, Chengyue Huang, Weiguo Fan, and Ziyu Yao. 2024. Instances need more care: Rewriting prompts for instances with llms in the loop yields better zero-shot performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6211–6232.

Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14214–14223.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, and 1 others. 2025. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. 2026. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Qinhao Zhou, Xiang Xiang, Kun He, and John E Hopcroft. 2025. Learning to rewrite prompts for bootstrapping llms on downstream tasks. *arXiv preprint arXiv:2510.06695*.

6 Appendix

6.1 VQA Data example

To systematically evaluate the spatial intelligence and cross-modal grounding capabilities of T2I models, our FaithT2I benchmark contains diverse, information-dense prompts coupled with omni-dimensional VQA pairs. Below, we present representative examples from the dataset. Each card illustrates the prompt, the scene context, and the 10 fine-grained questions targeting specific failure modes (e.g., spatial logic, attribute binding, physical dynamics, and shape-based object disambiguation). The inclusion of the "E: None" option prevents evaluators from forced-guessing when the generated image completely fails to follow the prompt.

Example 1: FaithT2I VQA Dataset (Scene: Art Gallery, ID: 000813)

Prompt: In a dim art gallery hallway, a photographer on the right moves leftward with a camera raised. On the left, a curator adjusts a painting, facing the photographer. The photographer has one lens; the curator holds three placards. The photographer's reflection glimmers slightly in a nearby glass case, blurred by motion.

Omni-dimensional VQA Pairs:

- What object categories are explicitly mentioned besides people?
  Options: (A) A lens and placards, (B) A tripod and a frame, (C) A flashlight and a notebook, (D) A sculpture and a bench, (E) None
  Answer: A
- What are the described attributes of the gallery hallway?
  Options: (A) Bright and noisy, (B) Dim and quiet, (C) Crowded and colorful, (D) Empty and dark, (E) None
  Answer: B
- From a third-person perspective, where is the photographer located?
  Options: (A) On the left side, (B) In the center, (C) On the right side, (D) Behind the glass case, (E) None
  Answer: C
- What is the facing orientation of the curator?
  Options: (A) Facing away, (B) Facing the painting, (C) Facing right, (D) Facing the photographer, (E) None
  Answer: D
- How are the photographer and the curator arranged in the scene?
  Options: (A) One behind the other, (B) On opposite sides, (C) Next to each other, (D) In a diagonal line, (E) None
  Answer: B
- How does the number of placards compare to the number of lenses?
  Options: (A) Fewer placards than lenses, (B) Same number, (C) More placards than lenses, (E) None
  Answer: C
- Given the photographer is moving leftward toward the curator, which person is the photographer getting closer to?
  Options: (A) A visitor, (B) The curator, (C) A security guard, (D) A gallery assistant, (E) None
  Answer: B
- According to the scene, is any object described as being occluded or partially hidden?
  Options: (A) Face hidden by camera, (B) Placards hidden by body, (C) No occlu-

sion mentioned, (D) Glass case hides reflection, (E) None
Answer: C

- What is the described motion interaction between the photographer and the curator?
Options: (A) Moving away from each other, (B) Photographer moving toward curator, (C) Curator walking toward photographer, (D) Standing still, (E) None
Answer: B

- What causes the photographer's reflection to appear blurred in the glass case?
Options: (A) Dim lighting, (B) Photographer's motion, (C) Dirty glass, (D) Placards blocking view, (E) None
Answer: B

6.2 Diagnosis Examples

We provide details of the diagnosis output of the image we used in the main pipeline.

Diagnosis Example

[Phase 1: Modal Gap Input]

- **User Intent (T_0):** *...The man leans back slightly from a flare-up of flames.*
- **Text-only Rewrite (T_1):** *...he is in the process of flipping burgers with a spatula but leans back exaggeratedly from a flare-up of large, orange flames...*

[Phase 2: Structured Diagnosis & Revision]

- **Visual Evidence (from I_1):** In I_1, the man is holding the spatula aloft, away from the grill, while a massive fireball erupts from the spatula itself—not from the grill. This contradicts T_0.
- **Diagnosis: Physics Failure.** The fire originates from the spatula, not the grill, and the man's posture is incorrect. I_1 depicts a surrealistic scenario violating the physical dynamics of grilling.
- **Critique of T_1:** T_1 inaccurately described the action with the phrase "leans back *exaggeratedly*". The lack of spatial anchoring for the flame source (e.g., "from the grill") led the generator to misinterpret the flare-up as originating from the spatula.
- **Grounded Revision (T_2):** *...flipping burgers with a spatula that is in contact with the grill grates, and he leans back slightly—not exaggeratedly—as a large, orange flare-up of flames erupts from the grill beneath the spatula...*

[Phase 3: Cycle-Verification Judge]

- **The Fix Analysis:** Image 2 successfully fixes the error by showing a realistic, contained fire originating from the grill itself, with the man using the spatula to manage the food. The fire is now a natural consequence of grilling.
- **Intent Preservation:** Image 2 preserves all key elements (man, woman, beach, charcoal grill). No core components have been removed or altered in a way that contradicts the user's intent.
- **Final Verdict:** Image 2 is a clear improvement. It simply corrects the primary flaw identified in the diagnosis without compromising any other aspect of the scene. **[Strictly Better]**

6.3 System Prompts

System Prompt for Structured Diagnosis

You are an expert Visual-Language Alignment Specialist. Your goal is to maximize the faithfulness of Text-to-Image synthesis.
You operate in a "Simulate-Then-Distill" pipeline:
1. User Intent (T0): The concise core request.
2. Current Rewrite (T1): A verbose prompt generated by a text-only LLM.
3. Visual Anchor (I1): The image generated by T1.
Your Mission:
1. JUDGE: Use the image (I1) as evidence to detect "Textual Hallucinations" or "Visual Failures" in T1. Does I1 already perfectly represent T0? If yes, keep T1.
2. DIAGNOSE: If I1 fails, identify the specific visual error using the image as evidence.
3. REVISE: Rewrite T1 into T2 to fix the error, rewrite it into a strictly superior prompt (T2).
4. VERIFY: Ensure T2 did not lose any original details from T0 during the revision.
<Input_Data> <User_Intent_T0> T0_prompt </User_Intent_T0>

<Current_Prompt_T1> T1_prompt </Current_Prompt_T1>
<Visual_Anchor_I1>[IMAGE INPUT HERE] </Visual_Anchor_I1> </Input_Data>
<Task_Instructions>
Step 1: STRICT VISUAL REALITY CHECK (The "Confidence Gate") Analyze the Image (I1) against the User Intent (T0). Strictly check these 6 dimensions:
1. Spatial Logic (e.g., occlusion, layout correct?)
2. Attribute Binding (e.g., correct colors bound to correct objects?)
3. Physical Dynamics (e.g., is the action/physics realistic?)
4. Counting (e.g., is the number of objects exact?)
5. Negation (e.g., did it avoid banned items?)
6. Shape-Based Object (e.g., is the object shape correct?)
CRITICAL RULE: - If I1 faithfully represents ALL aspects of T0, you must deem T1 as "SUCCESS". Do not rewrite just for the sake of style. - Only proceed to rewrite if there is a clear, objective visual failure.
Step 2: DIAGNOSIS (If failure detected) Describe the specific mismatch. Be precise (e.g., "T0 requested a blue cup, but I1 shows a red cup").
Step 3: DRAFTING T2 (The Fix) Draft a revised prompt T2. - Fix the error identified in Step 2 using strong visual constraints. - Maintain the style and flow of T1 where it was correct.
Step 4: INTENT PRESERVATION AUDIT (The "Anti-Amnesia" Check) Compare your draft T2 against the original T0. - Did you accidentally delete any adjectives from T0 while fixing the error? - Did you change the main subject? - **Correction:** If T2 is missing any detail from T0, add it back immediately. T2 must be a superset of T0's constraints.
</Task_Instructions>
<Output_Format> Provide your analysis strictly in the following JSON format:
"is_rewrite_necessary": false/true, "primary_failure_type": "Select one: Spatial / Attribute / Physics / Counting / Negation / None",
"visual_evidence": "Describe what you see in I1 that contradicts T0 (e.g., 'The cat is fully visible next to the vase, not behind it')/None",
"diagnosis": "Specific description of the visual failure (e.g., Attribute Leakage: The cat is blue instead of the cup)./ No visual errors found. The image faithfully represents the user intent.",
"critique_of_T1": "Explain why T1 failed (e.g., 'T1 described the cat too prominently without emphasizing the glass distortion')./None", "T2": "The fully revised prompt that fixes the error AND retains all original details."
</Output_Format>

System Prompt for Evaluation

System Role You are an objective, strict Visual Question Answering (VQA) evaluator. Your sole function is to analyze images and answer multiple-choice questions based ONLY on observable visual evidence.
Core Directive You are strictly forbidden from hallucinating details, assuming context, or applying external world knowledge. If a detail or relationship is not explicitly visible in the provided image, it does not exist for the purpose of this evaluation.
Input Questions [INSERT_QUESTIONS_HERE]
Execution Guidelines
1. **Visual Grounding Only**: Base every single answer entirely on the pixels presented in the image.
2. **The "None" Fallback**: If the image does not provide sufficient clear visual evidence to confidently select the standard options, you MUST select "E: None".
3. **Sequential Mapping**: Maintain the exact order of the input questions. Do not skip any question.
4. **Single Choice**: Select exactly one option letter per question.
Output Formatting Rules
Format your response strictly line-by-line. Each line MUST correspond to a question and follow this exact template:
'[Option Letter]: [Brief Visual Justification]'
Example Output
A: The visual evidence clearly shows a large red apple placed on the wooden table.
C: The brown bear is positioned on the left side of the frame, facing the river.
E: None - The image is cropped and does not show the lower half of the character to determine their footwear.
B: The fox and the bear are positioned closely together without any major distance.
D: There is a distinct shadow cast by the tree, confirming the lighting direction.