

Enhancing Text-to-Image Prompts using Reinforcement Learning

Aakash Varma Nadimpalli[§], Varunika Naini^{°,§}

Abhilash Ghanore[†], Lava Bokam[†]

[°]Intern [§]Equal Contribution; listed alphabetically [†]Main mentors

aakashv@dheyo.ai

varunikan@dheyo.ai

abhilash@dheyo.ai lava@dheyo.ai

February 23, 2026

Abstract. Text-to-image diffusion models can generate highly realistic images, but their performance depends heavily on detailed input prompts. While prompt enhancement methods improve visual quality by expanding user inputs, they often fail to preserve explicit constraints such as negation, exact counts, and attribute-specific requirements. Increasing descriptive detail frequently introduces subtle semantic shifts that violate these instructions. We present a reinforcement learning approach for constraint-aware prompt enhancement that incorporates structured visual feedback during training. By evaluating generated images against a structured taxonomy of logical and spatial constraints, and optimizing the prompt generation policy using this feedback, our method strengthens alignment between user intent and visual output. Experiments show improved adherence to compositional and count-based requirements compared to supervised fine-tuning baselines.



1 Introduction

Recent advances in text-to-image (T2I) diffusion models have made it possible to generate photo-realistic images directly from natural language prompts. Models like Stable Diffusion [1], FLUX, and Qwen-Image [2] produce impressive results, but their success depends heavily on detailed and precise input prompts. Users often need to provide elaborate descriptions to achieve high-quality outputs, which creates a significant accessibility barrier. Prompt enhancement methods address this by automatically expanding simple concepts into detailed, stylized prompts that yield better generations [3, 4].

These approaches improve visual appeal, but they struggle to maintain explicit user constraints like negation, count, specifics example: "no scallions," "exactly four dogs," or "all wearing red." During expansion, these instructions often get lost or contradicted. For example, "A bowl of beef noodles, no scallions" might become "garnished with fresh green scallions" due to data-driven associations. This leads to attribute mismatches, incorrect counts, and compositional errors.

To systematically address these failures, we first establish an evaluation framework based on a taxonomy of fine-grained key points, enabling verification of model performance against specific logical and spatial constraints. Building upon this, we develop a training pipeline that starts with

supervised fine-tuning of a lightweight language model to establish baseline capability. To achieve better results, we subsequently enhance original prompt adherence with improved accuracy using RL via Group Relative Policy Optimization (GRPO) [5]. This reinforcement learning stage establishes a closed feedback loop, where the model learns directly from visual evaluation of its generated images. Tying linguistic expansion to image generation outcomes.

By generating images from prompt variants and penalizing mistakes in the visual domain, our approach aligns text generation with pixel-level outcomes, balancing descriptiveness and precision. Our penalty mechanism suppresses prompts below correctness thresholds, preventing skewness toward aesthetics alone. Our contributions are:

- **A comprehensive evaluation framework** using a fine-grained taxonomy of constraint dimensions to verify whether enhanced prompts and their resulting images satisfy user constraints.
- **A reinforcement learning framework** that uses the evaluation metrics to optimize the prompt generation policy, ensuring adherence to logical constraints.
- **A compositional reward function** integrating key point verification, semantic fidelity, and descriptive quality for targeted logical and creative optimization.

Figure 1: Visual constraint satisfaction with GRPO-tuned prompt enhancement. Examples show the Llama-3.2-3B model, fine-tuned with visual feedback, producing prompts that strictly obey complex constraints: negation (top left: “no scallions”), exact numeracy (top right: “four dogs”), shape fidelity (middle left: “guitar-shaped lake”), and material accuracy (bottom left: “ice sculpture”).

2 Evaluation Framework

A fundamental challenge in automated prompt enhancement is ensuring that refined prompts yield images that strictly adhere to user intent without introducing extraneous or contradictory elements. Conventional metrics such as CLIP score are insufficient for detecting subtle logical or constraint

violations. To address this, we introduce an evaluation framework that quantifies both semantic fidelity and constraint compliance using a dedicated set of verification keypoints.

2.1 Pipeline

Each generated image is evaluated through an automated multi-stage pipeline. Let C denote the original user concept, P the enhanced prompt, and I the resulting image. First, a large language model (Gemini-2.5-Pro) functions as a constraint extractor E_{LLM} that decomposes C into a set of relevant keypoints $K \subseteq \mathcal{T}$, where $\mathcal{T}$ is the fixed taxonomy of 24 keypoints:

$$K = E_{\text{LLM}}(C) = \{k_1, k_2, \dots, k_m\} \quad (1)$$

For example, the prompt “A bowl of beef noodles, no scallions” yields keypoints including *Grammar-Negation* and *Attribute-Composition*.

A vision-language model then acts as the visual evaluator V_{VLM} , assigning to each keypoint k_i a binary-verified adherence score $v_i \in [0, 1]$:

$$v_i = V_{\text{VLM}}(I, k_i) \quad (2)$$

The composite Keypoint Adherence Score is defined as the mean over all extracted keypoints:

$$S_{\text{KP}} = \frac{1}{|K|} \sum_{i=1}^{|K|} v_i \quad (3)$$

In parallel, three complementary similarity metrics are computed:

- **Text-to-Image Alignment** (S_{sig}) using SigLIP:

$$S_{\text{sig}} = \sigma(\phi_{\text{sig}}(I) \cdot \phi_{\text{sig}}(P)) \quad (4)$$

- **Image-to-Image Structural Similarity** (S_{clip}) using CLIP:

$$S_{\text{clip}} = \frac{\phi_{\text{clip}}(I) \cdot \phi_{\text{clip}}(I_{\text{baseline}})}{|\phi_{\text{clip}}(I)| |\phi_{\text{clip}}(I_{\text{baseline}})|} \quad (5)$$

- **Text-to-Text Semantic Similarity** (S_{sem}) using Sentence Transformers:

$$S_{\text{sem}} = \frac{SEM(C) \cdot SEM(P)}{|SEM(C)| |SEM(P)|} \quad (6)$$

This pipeline enables fully automated, scalable assessment of prompt enhancement quality across large evaluation sets.

2.2 Keypoint Taxonomy

The evaluation employs a taxonomy of 24 keypoints grouped into six categories (Table 1). These keypoints target specific linguistic, visual, relational, spatial, and reasoning capabilities that are frequently violated by current generation models.

Table 1: Taxonomy of 24 Keypoints for Prompt-Image Constraint Verification

Category	Subcategory	Description	Example Prompt	Evaluation Criteria
Grammar & Linguistic	Grammar-Negation	Correct interpretation of negation	“A bowl of beef noodles, no scallions”	Noodles present; scallions strictly absent
	Linguistic-Co-reference	Resolution of pronoun ambiguity	“The large ball broke the table because it was made of metal”	“it” correctly bound to metal ball; table is wood
Attribute Understanding	Attribute-Consistency	Uniform attribute binding across entities	“Five people all wearing red clothes”	Every person wears red
	Attribute-Quantity	Precise numeracy ($n \geq 3$)	“A picture with four dogs”	Exactly four dogs rendered
	Attribute-Size	Relative size comparison	“A large elephant next to a small mouse”	Clear size distinction
	Attribute-Material	Accurate material rendering	“An ice sculpture of an eagle”	Transparent, crystalline ice appearance
	Attribute-Expression	Facial emotion fidelity	“A strong man with a contemptuous expression”	Clear contempt/disdain visible
Style	Attribute-Composition	Multi-part composition	“A car made of flowers”	Vehicle composed entirely of flowers
	Style-Artistic	Adherence to specified style	“A baroque ballroom scene”	Matches requested style (default: realistic)
Action Understanding	Action-Full Body	Complex full-body dynamics	“A person doing a backflip”	Correct posture and motion
	Action-Hand	Detailed hand/finger structure	“A person pointing with index finger”	Clear, anatomically correct fingers
	Action-Individual	Clear individual action	“A person reading a book”	Distinct reading engagement
	Action-State	Continuous state depiction	“A cat sleeping on a windowsill”	Sustained state (sleep) shown
Relationship	Relationship-Contact	Physical contact	“Two people shaking hands”	Direct hand-to-hand contact
	Relationship-Non-contact	Non-physical interaction	“A person looking at a painting”	Gaze directed at painting without touch
	Relationship-Comparative	Attribute comparison	“A tall person next to a short person”	Visible height difference
	Relationship-Containment	Container-content relation	“A bird in a cage”	Bird inside cage
	Relationship-Similarity	Shape/appearance resemblance	“Two similar-looking dogs”	Shared visual traits
Spatial Understanding	Spatial-Binding	Attribute binding to distinct entities	“A man with a buzz cut wearing a blue shirt”	Correct attribute assignment
	Spatial-Layout	Positional arrangement	“A cat on the left, dog on the right”	Exact spatial placement
Knowledge & Reasoning	World Knowledge	Accurate rendering of known entities	“The Eiffel Tower in Paris”	Recognizable landmark
	Reasoning-Counterfactual	Coherent impossible scenes	“A fish flying through the sky”	Surreal yet visually coherent
Text Rendering	Text-Rendering	Accurate text content	“A sign that says ‘STOP’”	Legible, correct spelling
	Text-Layout	Positional text placement	“The word ‘HELLO’ at the top”	Text in specified location

2.3 Baseline Performance with Generic System Prompt(Zero-Shot)

We first establish a baseline by evaluating Gemini-2.5-Pro conditioned only on a generic system prompt, reflecting standard out-of-the-box behavior without domain-specific fine-tuning or constraint-aware optimization.

Figure 2 presents a heatmap of the three similarity metrics across the prompt evaluation set. Table 6 reports the corresponding keypoint adherence scores. While basic attribute binding and simple action depiction frequently achieve high compliance, performance degrades significantly on negation, strict numeracy, counterfactual reasoning, and precise physical causality.

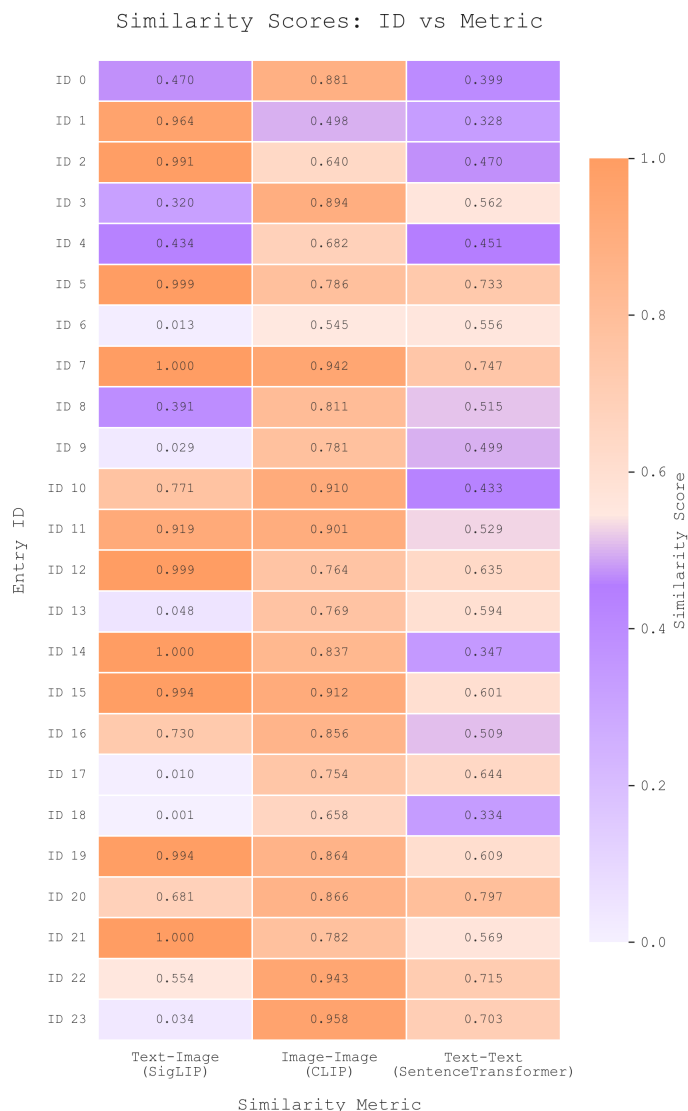


Figure 2: Heatmap of similarity scores across the 24 evaluated prompts. Rows correspond to entry IDs; columns show SigLIP (text-to-image), CLIP (image-to-image), and Sentence Transformers (text-to-text) metrics. Color scale ranges from low (purple, ≈ 0.0) to high (orange, ≈ 1.0) similarity. Darker purple regions indicate substantial deviation from the intended prompt semantics or visual structure.

2.4 Failure analysis

The baseline results reveal systematic limitations in constraint fidelity under minimal guidance. The model consistently prioritizes aesthetic fluency and visual coherence over literal adherence, leading to the insertion of conventional or prototypical elements that violate explicit exclusions (e.g., scallions appear despite negation). Numeracy is relaxed in favor of compositional balance, and causal/physical interactions are frequently softened or omitted. Style requests are overridden by internal priors favoring cinematic or digital-artistic rendering, even when photorealism is specified. Text rendering shows sporadic but non-trivial errors, and counterfactual instructions are approximated rather than precisely realized.

Table 2 summarizes the most severe failures (keypoint scores ≤ 0.5). These cases demonstrate that generic fluency optimization is inadequate when hard logical, physical, or compositional boundaries must be respected.

Table 2: Significant failure cases from the baseline evaluation of Gemini-2.5-Pro under a generic system prompt (keypoint scores ≤ 0.5).

Idx	Metric	Score	Explanation
0	Grammar-Negation	0.0	Explicit exclusion of scallions violated; chopped green onions present.
0	Attr-Composition	0.0	Negation failure propagates to composition.
0	Rel-Containment	0.2	Partial containment success, undermined by prohibited elements.
2	Rel-Contact	0.0	No causal interaction: ball present but table intact.
4	Style-Artistic	0.0	Realistic style overridden by surreal cosmic rendering.
7	Attr-Quantity	0.0	Requested eight horses violated (ten generated).
9	Act-Hand	0.0	Required hand absent; chopsticks appear without operator.
11	Act-Hand	0.0	Finger detail obscured by boxing gloves.
13	Style-Artistic	0.0	Realistic style violated by physically impossible floral vortex.
17	Reason-Counterfactual	0.1	Lake fails to adopt guitar shape.
18	Style-Artistic	0.2	Photorealism overridden by illustrative style.
19	Style-Artistic	0.0	Realistic style violated by cyberpunk aesthetic.
21	Style-Artistic	0.0	Photorealism violated by fantasy rendering.
22	Text-Rendering	0.5	Text rendered as “GAME OTHRONES”.

3 Experimentation on the Fine-Tuned Model

3.1 Model Architecture and Fine-Tuning Data

To overcome the constraint-adherence limitations observed in zero-shot prompting, we adopt a supervised fine-tuning (SFT) strategy. The chosen base model is **Novaeus-Promptist-7B-Instruct**, an instruction-tuned 7-billion-parameter language model optimized for creative prompt enhancement tasks. Fine-tuning was performed on a combined dataset aggregating three high-quality open-source prompt enhancement corpora:

- gokaygokay/prompt-enhancement-75k (73.2k samples)
- gokaygokay/prompt-enhancer-dataset (17.9k samples)
- prithivMLmods/Prompt-Enhancement-Mini (1.16k samples)

This mixed training set provides diverse stylistic and semantic expansions while preserving original intent.

3.2 Quantitative and Qualitative Results

The fine-tuned Novaeus-Promptist-7B model was evaluated using the previously described pipeline, with expanded prompts rendered through the Nano Banana Pro model. Full per-prompt keypoint scores are reported in Appendix 8 (Table 7); two representative cases are shown in Table 3 below.

The model achieves a score of 1.0 on several constructive dimensions, including Attribute-Consistency, Attribute-Quantity, Style-Artistic, Text-Rendering, and Text-Layout. Multi-entity color binding tasks such as “Five people all wearing red clothes,” cross-individual attribute grounding as in “Man with a buzz cut next to a woman with long hair,” and structured poster-style text placement consistently obtain perfect scores, indicating stable preservation of explicitly specified constraints.

Similarly, the prompt “A puppy happily running” achieves 1.0 across Action-Full Body, Action-Individual, Action-State, and Attribute-Expression. This pattern suggests that compositionally aligned affirmative constraints are reliably grounded after fine-tuning.

Table 3: Two representative cases from the Novaeus-Promptist-7B evaluation.

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A bowl of beef noodles, no scallions.		A medium close-up of a bowl of beef noodles with a light brown broth. Two pieces of meat on top of the noodles. The bowl is on a white plate with a small amount of condiments on the side.		Grammar-Negation: 0.0
Five people all wearing red clothes.		A medium shot of five people all wearing red shirts, black pants, and black shoes, all facing right. Each person holds a black cup. A small white sign reading “Brew” sits on the ground in front of them.		Attribute-Consistency: 1.0

3.3 Failure Analysis

Despite clear progress on constructive and descriptive tasks, the model exhibits persistent weaknesses in exclusion, material grounding, and complex motion representation. The negation-sensitive prompt “A bowl of beef noodles, no scallions” receives a score of 0.0 on Grammar-Negation, with the enhanced prompt reintroducing condiments. “An ice sculpture of an eagle” shows weak enforcement of material constraints, and “A girl performing a Thomas flare” fails to preserve full-body action structure. Shape-level similarity in “A lake shaped like a guitar” is also poorly realized.

These systematic failures reveal a core limitation of pure supervised fine-tuning. The objective optimizes token-level likelihood rather than strict constraint satisfaction. High-frequency visual-textual associations dominate generation, while exclusion rules, physical material constraints, and global geometric structure are not explicitly penalized without an external grounding or reward mechanism.

4 Reinforcement Learning

4.1 Motivation

Supervised fine-tuning improves descriptive fluency but does not reliably enforce logical or physical constraints present in the original concept. Prompt variations remain prone to high-frequency co-occurrence biases inherited from training data. To align the model with the visual evaluation signal, we formulate prompt enhancement as a reinforcement learning problem: the prompt generator is the policy π_θ , the frozen diffusion model is the environment, and the composite evaluation score is the reward.

4.2 Group Relative Policy Optimization (GRPO)

We apply Group Relative Policy Optimization (GRPO) [5], which avoids the need for a separate value network by exploiting group-level statistics of sampled outputs.

For each input concept q , the policy π_θ generates G candidate prompts $\{o_1, \dots, o_G\}$. Each o_i is rendered by the diffusion model to produce image I_i and reward r_i . The advantage is computed as:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\}) + \epsilon} \quad (7)$$

The GRPO objective is:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\} \sim \pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^G \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i - \beta \mathbb{D}_{KL}(\pi_\theta || \pi_{ref}) \right) \right] \quad (8)$$

This enables policy updates driven directly by the downstream visual reward while constraining deviation from the reference model.

4.3 Reward Functions

4.3.1 SigLIP-Only Reward

We first consider a reward defined purely through text–image alignment:

$$R_{\text{baseline}} = \sigma(\text{SigLIP}(I, C)) \times 10.0 \quad (9)$$

This objective directly optimizes global semantic similarity without enforcing explicit logical or structural constraints. Comprehensive per-prompt results for both Novaeus and Qwen are reported in Appendix 8, with detailed tables in Table 8 and Table 9. Two representative cases are reproduced in Table 8 for clarity.

Result Analysis. Across both Novaeus and Qwen, optimization with the SigLIP-only reward consistently increases SigLIP_C alignment scores, often approaching saturation after RL. Absolute improvements typically range from +0.10 to +0.20. For instance, “A bowl of beef noodles, no scallions” improves from 0.7374 to 0.9299 (+0.1925), while “A boxer lands a punch on a punching bag” increases from 0.7834 to 0.8913 (+0.1079). Similar gains are observed across identity-binding and attribute-grounding prompts in Appendix 8, confirming that the reward robustly amplifies global semantic correspondence.

However, the effect on structural correctness is not uniform. In action-centric prompts such as the boxing example, A.K.S improves markedly from 0.6763 to 0.94 (+0.2637), indicating stronger pose and interaction grounding. In contrast, negation-sensitive prompts such as “no scallions” exhibit a decline in A.K.S from 0.5866 to 0.375 (0.2116) despite the substantial alignment gain, reflecting reinforcement of dominant visual features rather than strict exclusion.

Overall, the SigLIP-only objective reliably strengthens coarse text–image similarity across both models, but its impact on constraint fidelity remains task-dependent. When semantic alignment and structural correctness are naturally coupled, both metrics improve jointly; when decoupled, optimization favors similarity over logical precision. This trend, consistent across Tables 8 and 9, underscores the limitation of relying solely on global alignment as a reward signal.

4.3.2 Composite Reward

To enforce constraint fidelity, we define a composite reward:

$$R_{\text{final}} = \tanh(S_{\text{base}} - P_{\text{penalty}}) \quad (10)$$

with base score

$$S_{\text{base}} = 0.45 \cdot S_{\text{KP}} + 0.35 \cdot S_{\text{Con}} + 0.20 \cdot S_{\text{Enh}} \quad (11)$$

where

$$S_{\text{KP}} = \frac{1}{N} \sum_{j=1}^N \mathbb{I}_{\text{valid}}(I, k_j), \quad (12)$$

$$S_{\text{Con}} = \sigma(\text{SigLIP}(I, C)), \quad (13)$$

$$S_{\text{Enh}} = \sigma(\text{SigLIP}(I, P)). \quad (14)$$

Table 4: Prompt evaluation results using SigLIP reward and Novaeus model. A.K.S denotes *Average Keypoint Score*.

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A bowl of beef noodles, no scallions.	A steaming bowl of beef noodles sits on a white ceramic plate, the rich brown broth glistening... no scallions are visible...		A detailed image of a bowl of beef noodles with no scallions on a clean white background... focus on the dish itself...		Before RL: A.K.S: 0.5866 SigLIP: 0.7374 After RL: A.K.S: 0.375 SigLIP: 0.9299
A boxer lands a punch on a punching bag.	A powerful boxer delivers a fierce left hook that sends shockwaves through the heavy punching bag...		A vivid image of a boxer delivering a powerful punch in a boxing gym... capturing the raw power...		Before RL: A.K.S: 0.6763 SigLIP: 0.7834 After RL: A.K.S: 0.94 SigLIP: 0.8913

A penalty term is applied as:

$$P_{\text{penalty}} = \begin{cases} 0.25 & \text{if } S_{\text{KP}} < 0.25 \text{ and } N > 0 \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

This design emphasizes structural validity while retaining auxiliary alignment signals. A thresholded penalty enforces minimum keypoint compliance before semantic alignment can positively influence the reward. The final tanh normalization bounds the optimization signal and stabilizes training.

Result Analysis. Under the composite reward, structural fidelity improves consistently across constraint-sensitive prompts, with A.K.S gains observed in the majority of cases (Table 10). Large corrections occur in prompts with strong relational or attribute constraints. For example, “Man with a buzz cut and blue shirt next to a woman with long hair and a yellow shirt” increases from 0.2725 to 0.7613, and “Poster with text ‘Game of Thrones’ at the bottom” improves from 0.3134 to 0.7094. Similarly, motion and pose-sensitive prompts such as “A strong man, low-angle shot, with a contemptuous expression” rise from 0.4254 to 0.6858, and “A gust of wind blows, cherry blossoms dance in the air” increases from 0.6081 to 0.7409.

Counting and compositional constraints also show stable or moderate gains. “A bowl of beef noodles, no scallions” improves from 0.6264 to 0.7253, “Eight galloping horses in Chinese ink wash” from 0.6257 to 0.7321, and “A cup full of soda water” from 0.6434 to 0.7421. Even previously degenerate cases such as “Two large spheres” shift from 0.1919 to 0.0894, indicating partial structural correction.

Importantly, improvements are achieved without degrading already strong cases: prompts such as “Five people all wearing red clothes” remain stable at 0.7605 to 0.7616, and “Einstein looking at Hawking” remains effectively unchanged at 0.7612 to 0.7614. Overall, the composite objective

Table 5: Prompt Evaluation using Composite Reward Function. A.K.S denotes *Average Keypoint Score*.

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A bowl of beef noodles, no scallions.	A steaming bowl of savory beef noodles, served in a rustic ceramic bowl, with tender strands of slow-cooked beef in a rich, dark brown broth, garnished with a sprinkle of toasted sesame seeds and a wedge of lime on the		A steaming bowl of beef noodles sits on a wooden table, with a sprinkle of black pepper on top. The noodles are long and thin, cooked to a perfect texture, and are swimming in a savory beef broth.		Before RL: A.K.S: 0.6264 After RL: A.K.S: 0.7253
Five people all wearing red clothes.	Five people dressed in vibrant red clothing, each wearing a different shade of red, from deep burgundy to bright cherry, stand together in a bustling city square..... drawn in by the vibrant colors and infectious energy.		Five people standing together, each wearing bright red clothing, with bold red scarves and accessories, in a vibrant city street with tall skyscrapers and a bustling atmosphere in the background.		Before RL: A.K.S: 0.7605 After RL: A.K.S: 0.7616

produces broad A.K.S improvements, particularly in logically constrained and relation-heavy prompts, while maintaining stability in cases that were already structurally correct.

5 Inference Pipeline

The final system is the composition of two mappings. Let π_{θ^*} be the GRPO-optimized policy. Given user concept C :

1. Prompt expansion:

$$P^* = f_{\text{enhance}}(C; \theta^*) = \operatorname{argmax}_P \pi_{\theta^*}(P|C) \quad (16)$$

2. Image synthesis:

$$I_{\text{final}} = f_{\text{diff}}(P^*) = \mathcal{D}(z \sim \mathcal{N}(0, \mathbf{I}), P^*) \quad (17)$$

where $\mathcal{D}$ is the frozen Qwen-Image-2512 diffusion model.

The resulting function $I_{\text{final}} = f_{\text{diff}}(f_{\text{enhance}}(C))$ produces images that respect the logical, spatial, and semantic constraints expressed in C .

Figure 3: Pipeline architecture. The policy generates prompt candidates from user concept C , which are rendered by Qwen-Image-2512. Visual evaluation yields rewards for GRPO-based policy optimization.

6 Experimental Setup

Models. The framework comprises four core components. The policy π_θ is based on Llama-3.2-3B-Instruct. The frozen image generator $\mathcal{G}$ is Qwen-Image-2512. Evaluation employs a dual-judge system: (i) a visual evaluator based on Qwen2-VL-7B-Instruct in bfloat16 precision with dynamic resolution to maximize single-GPU throughput, and (ii) a semantic judge using SigLIP-SO400M for embedding-based similarity computation.

Dataset. Training and evaluation draw from a filtered subset of `gokaygokay/prompt-enhancement-75k`. The subset targets constraint-heavy categories from Table 1, including *Grammar-Negation*, *Attribute-Quantity*, and *Spatial-Layout*, to focus on logical and structural alignment.

Implementation. All experiments run on a single NVIDIA H100 GPU. Training uses the AdamW 8-bit optimizer with a cosine learning-rate decay schedule. Group size is fixed at $G = 4$, generating four prompt candidates per concept during exploration. This configuration supports direct advantage computation in GRPO without requiring a learned value baseline.

7 Analysis

Training dynamics were monitored via the average reward trajectory. Under the SigLIP-only reward, semantic similarity increased rapidly, but key point adherence exhibited high variance and oscillations. This behaviour is consistent with reward hacking. The policy discovers linguistically fluent patterns that boost embedding alignment while bypassing logical constraints.

The composite reward, incorporating discrete key point verification and a penalty for low

adherence, yields smoother convergence. The penalty term effectively shapes the policy away from local optima of aesthetic fluency and toward distributions that satisfy both semantic and constraint objectives.

8 Conclusion

We introduced a visually grounded prompt enhancement pipeline that applies Group Relative Policy Optimization (GRPO) [5] to align a 3B-parameter language model with text-to-image generation objectives. By closing the loop between generated text and visual output via an automated multi-model judge, the approach enables robust handling of negation, counting, and spatial constraints that remain challenging for larger models under standard prompting or SFT. The use of GRPO eliminates the need for a critic network, enabling efficient training on a single H100 GPU. These results demonstrate a scalable path toward constraint-aware prompt optimization without reliance on human annotations.

Appendix

A.1 Gemini Baseline Evaluation

Table 6: Prompt Evaluation with KP Scores Using Gemini-2.5-Pro.

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A bowl of beef noodles, no scallions.		A steaming, large ceramic bowl filled with rich, dark beef noodle soup. Tender thick-cut chunks of slow-braised beef brisket rest atop wavy egg noodles. A few vibrant green bok choy leaves on the side. Professional food photography style, warm lighting, shallow depth of field.		Style Artistic: 1.0; Grammar Negation: 0.0; Attr Composition: 0.0; Rel Containment: 0.2
Five people all wearing red clothes.		Five enigmatic figures in a baroque ballroom, each in a unique crimson outfit—gowns of flowing satin, tailored velvet suits. A dramatic spotlight cuts through the darkness, illuminating the vibrant red attire. Cinematic and moody atmosphere.		Style Artistic: 0.9; Attr Quantity: 1.0; Attr Consistency: 1.0
The large ball broke the table because it was made of metal.		A hyper-detailed cinematic shot of a massive, polished metal sphere crashing through an ornate mahogany dining table. Splinters frozen mid-air in a dramatic shaft of light. Sphere dented and scarred, reflecting the dimly lit room. Photorealistic, 8K, dynamic.		Style Artistic: 1.0; Ling Co-reference: 1.0; Attr Size: 1.0; Attr Material: 1.0; Rel Contact: 0.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A picture with four dogs.		Four dogs—a Golden Retriever, French Bulldog, Border Collie, and Samoyed—playing together in a sun-drenched park. The Golden leaps for a frisbee while the others tug a rope toy or bark playfully. Warm golden-hour light, vibrant green grass.		Style Artistic: 1.0; Attr Quantity: 1.0
Two large spheres.		Two colossal spheres hover over a tranquil, shallow body of water. One of polished obsidian absorbing all light, the other a shimmering mother-of-pearl orb refracting a cosmic nebula sky. Still water creates a perfect mirror. Surreal digital painting.		Style Artistic: 0.0; Attr Consistency: 1.0; Attr Size: 1.0; Rel Similarity: 1.0
An ice sculpture of an eagle.		A hyperrealistic ice sculpture of a majestic bald eagle, wings spread wide, crystal-clear and flawlessly transparent. Intricate feathery details etched into the ice. Dramatically backlit by a vibrant multi-colored light display in an elegant winter festival setting.		Style Artistic: 1.0; Attr Material: 1.0; Attr Composition: 1.0
A strong man, low-angle shot, with a contemptuous expression.		An extreme low-angle shot of a man exuding absolute contempt. Chiseled jawline tight with disdain, eyes narrowed in a condescending glare, a cruel smirk on his lips. Dark tailored suit. Chiaroscuro lighting against a rain-slicked neo-noir cityscape.		Style Artistic: 1.0; Attr Expression: 0.6

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
Eight galloping horses in Chinese ink wash.		A masterpiece of Chinese ink wash (Shuimo-hua) capturing eight magnificent horses galloping across a minimalist landscape. Bold calligraphic brushstrokes define powerful muscles and windswept manes. Monochromatic palette on textured rice paper.		Style Artistic: 1.0; Attr Quantity: 0.0; Act Full Body: 1.0; Act State: 1.0
A girl performing a Thomas flare.		A dynamic B-girl in mid-Thomas flare, body a blur of motion with legs swinging in a circular arc, hands planted on a graffiti-covered floor. High-shutter-speed photography, dramatic low-key spotlight lighting, vibrant neon colors.		Style Artistic: 0.6; Act Full Body: 1.0; Act Hand: 1.0; Act Individual: 1.0; World Knowledge: 1.0
A hand using chopsticks to pick up food.		A hyperrealistic close-up of elegant dark ebony chopsticks lifting a piece of glistening, golden-brown gyoza from a steaming minimalist ceramic bowl. Wisps of steam curl around the chopsticks. Warm-toned wooden table background. Soft, dramatic side lighting.		Style Artistic: 1.0; Act Hand: 0.0; Act Individual: 0.5; Act State: 0.5; Rel Contact: 2
A puppy happily running.		A fluffy Golden Retriever puppy captured mid-stride, joyfully bounding through a sun-dappled meadow. Ears floppy, tongue lolling in a happy pant. Vibrant green grass, colorful wildflowers, warm golden-hour light. Low-angle action shot, shallow depth of field.		Style Artistic: 1.0; Attr Expression: 1.0; Act Full Body: 1.0; Act Individual: 1.0; Act State: 1.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A boxer lands a punch on a punching bag.		A dramatic, high-speed photograph of a boxer landing a powerful right hook on a heavy leather punching bag. The impact creates a visible shock-wave, dust and sweat particles frozen mid-air. Chiaroscuro lighting from a single overhead lamp.		Style Artistic: 0.8; Act Full Body: 1.0; Act Hand: 0.0; Act Individual: 1.0; Rel Contact: 1.0
Einstein looking at Hawking.		A poignant digital painting of Albert Einstein and Stephen Hawking in an ethereal library. Einstein in tweed jacket leans forward curiously at Hawking in his wheelchair. Towering bookshelves, blackboard with equations, swirling nebula through a grand window.		Style Artistic: 0.8; Act State: 1.0; Rel Non-contact: 1.0; World Knowledge: 1.0; Reason Counterfactual: 1.0; Spatial Binding: 1.0
A gust of wind blows, cherry blossoms dance in the air.		A digital painting of a powerful gust of wind sweeping through a Japanese garden. A swirling vortex of pink and white cherry blossom petals momentarily obscures an ancient pagoda. Warm golden afternoon light, low-angle composition, ethereal sense of movement.		Style Artistic: 0.0; Attr Quantity: 1.0; Act State: 1.0; Act Individual: 1.0; World Knowledge: 1.0; Spatial Layout: 1.0
Woman in red dress taller than woman in yellow.		A cinematic full-body portrait of two women in a sun-drenched Italian courtyard. The tall woman on the left in a floor-length crimson silk gown gazes slightly downward at the shorter woman in a marigold yellow knee-length dress. Golden-hour light, shallow depth of field.		Style Artistic: 0.9; Attr Size: 1.0; Rel Comparative: 1.0; Spatial Binding: 1.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A cat made of orange slices.		A hyperrealistic digital art piece of a whimsical cat sculpted from vibrant, translucent orange slices, curled up on a rustic dark wooden table. Whiskers of orange zest, eyes of tiny dark seeds. Warm morning light illuminates the cat. Macro shot, magical atmosphere.		Style Artistic: 1.0; Attr Composition: 1.0; Reason Counterfactual: 1.0
A cup full of soda water.		A hyper-detailed macro photograph of a tall Collins glass filled with sparkling soda water. Thousands of tiny bubbles race to the surface, glowing like pearls in crisp backlighting. A vibrant lime wedge rests on the rim. Clean minimalist grey background.		Style Artistic: 1.0; Attr Material: 1.0; Act State: 1.0; Rel Containment: 1.0
A lake shaped like a guitar.		An aerial shot of a mystical lake shaped like a grand acoustic guitar, nestled in an ancient forest of emerald green. Pine trees form the frets on the neck, a rustic wooden bridge mimics the guitar's bridge. Soft ethereal glow of late afternoon sun. Hyperrealistic fantasy blend.		Style Artistic: 0.9; Rel Similarity: 0.6; Reason Counterfactual: 0.1

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
Man with a buzz cut and blue shirt next to a woman with long hair and a yellow shirt.		A photorealistic portrait of a man with a neatly trimmed buzz cut in a casual navy blue t-shirt, beside a woman with long chestnut hair in a vibrant lemon-yellow sundress. Rustic café patio, golden hour sun dappling through leaves. Medium shot, shallow depth of field.		Style Artistic: 0.2; Spatial Binding: 0.9; Spatial Layout: 1.0
A race car on a city track, with a mini-map in the top-left corner.		A hyper-realistic shot of a futuristic crimson race car in mid-drift on a rain-slicked city track at night. Dense cyberpunk cityscape with neon signs. A sleek holographic mini-map overlaid in the top-left corner. Dramatic high-contrast lighting, lens flares.		Style Artistic: 0.0; Act State: 1.0; Rel Contact: 1.0; Rel Containment: 1.0; Spatial Layout: 1.0; World Knowledge: 1.0
A wide shot of The Great Wall of China.		A breathtaking ultra-wide panoramic shot of the Great Wall of China snaking across a mountain range during golden-hour sunset. Warm light accentuates weathered stone textures. Long shadows, gentle mist in the valleys. Fiery orange and purple sky. Hyper-realistic, cinematic.		Style Artistic: 1.0; World Knowledge: 1.0; Spatial Layout: 1.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A girl held onto the stem of a huge dandelion with both hands, suspended above the clouds.		A young girl with a determined grip holds the stem of a colossal, fluffy dandelion suspended high above a sea of soft, billowing clouds. Sky painted in purples, oranges, and pinks. Low-angle shot emphasizing the vast sky. Detailed and vibrant digital painting, dreamlike atmosphere.		Style Artistic: 0.0; Attr Size: 1.0; Act Full Body: 1.0; Act Hand: 1.0; Act Individual: 1.0; Act State: 1.0; Rel Contact: 1.0; Spatial Layout: 1.0; Reason Counterfactual: 1.0
Poster with text “Game of Thrones” at the bottom.		A masterfully designed “Game of Thrones” poster featuring the Iron Throne in a snow-covered northern landscape. Ethereal light beams break through stormy grey clouds. Cold palette of blues, greys, and blacks. “Game of Thrones” in gothic font at the bottom. Wide-angle, epic scale.		Style Artistic: 1.0; Text Rendering: 0.5; Text Layout: 1.0; World Knowledge: 1.0
Poster of a woman on a throne of waves, text “Game of Thrones” at the bottom.		An epic fantasy digital painting of a powerful sea queen on a throne of colossal crashing waves. Her gown blends with the water, hair flows like seafoam. Stormy sky with a single ray of sunlight. Low-angle shot emphasizing regal authority. “Game of Thrones” in cinematic font at bottom.		Style Artistic: 1.0; Text Rendering: 1.0; Text Layout: 1.0; Attr Composition: 1.0; Act State: 1.0; Rel Contact: 1.0; World Knowledge: 1.0; Spatial Layout: 1.0; Reason Counterfactual: 0.9

A.2 Fine-Tuned Model Evaluation

Table 7: Prompt Evaluation with KP Scores Using Novaeus-Promptist-7B.

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A bowl of beef noodles, no scallions.		A medium close-up of a bowl of beef noodles with a light brown broth. Noodles are circular in shape, two pieces of meat on top. The bowl is on a white plate with a small amount of condiments on the side.		Grammar-Negation: 0.0
Five people all wearing red clothes.		A medium shot of five people all wearing red shirts, black pants, and black shoes, all facing towards the right. Each person is holding a black cup. There is a small white sign that says "Brew" on the ground in front of them.		Attribute-Consistency: 1.0
The large ball broke the table because it was made of metal.		A large metal ball rolling on a wooden table, black and white striped pattern, diameter of about 20 cm. The table is rectangular with a small crack in the middle caused by the impact of the ball.		Linguistic-Co-reference: 1.0
A picture with four dogs.		A close-up of four dogs on a concrete floor: a white poodle, a black poodle, a brown poodle, and another black poodle. All sitting with paws on the ground. Gray tile floor with a black metal frame around the edge.		Attribute-Quantity: 1.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
Two large spheres.		A detailed view of two large white spheres with a glossy finish, placed on a concrete floor in front of a black wall. The spheres are on a metal frame, one slightly larger than the other.		Attribute-Size: 1.0; Attribute-Quantity: 0.0
An ice sculpture of an eagle.		A detailed view of an ice sculpture of an eagle facing forward with wings spread out. The eagle has a white head and brown body, wings made of many pieces of ice, sitting on a piece of ice in front.		Attribute-Material: 0.1
A strong man, low-angle shot, with a contemptuous expression.		A medium shot of a strong man in a black shirt and blue jeans, arms crossed, with a contemptuous expression. Bright light illuminates his face, casting a shadow on the concrete wall behind him.		Attribute-Expression: 0.8
Eight galloping horses in Chinese ink wash.		Eight galloping horses in Chinese ink wash, depicted in motion with hooves raised and tails flowing. Ink wash technique creates depth and movement, arranged in a diagonal line on white paper background.		Style-Artistic: 1.0
A girl performing a Thomas flare.		A close-up of a girl performing a Thomas flare on a trampoline, wearing a black tank top with white shorts. Legs in a split position, arms extended. Trampoline surrounded by green grassy area with a wooden fence.		Action-Full Body: 0.1

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A hand using chopsticks to pick up food.		A close-up of a hand holding chopsticks, gripping with thumb and index finger to pick up a small piece of chicken. White surface background with a small plate of rice on the left.		Action-Hand: 1.0
A puppy happily running.		A medium shot of a white and black spotted puppy running through a field of green grass, mouth open with tongue sticking out. Blue sky with white clouds in the background.		Action-Full Body: 1.0; Action-Individual: 1.0; Action-State: 1.0; Attribute-Expression: 1.0
A boxer lands a punch on a punching bag.		A medium shot of a boxer in a gym, wearing red gloves and tank top, landing a punch on a punching bag. Right arm extended, fist making contact with the bag swinging back. Face focused and determined.		Action-Full Body: 1.0; Action-Hand: 1.0; Action-Individual: 1.0; Rel-Contact: 1.0
Einstein looking at Hawking.		A medium shot of Einstein looking at Hawking. Both in black suits at a table with a black tablecloth. A blackboard with equations in the background. A plate with cake and a white cup on the table.		Rel-Non-contact: 0.0
A gust of wind blows, cherry blossoms dance in the air.		A close-up of a tree with pink cherry blossoms, branches covered in pink flowers. The wind is blowing, causing flowers to move back and forth. Blue sky, no clouds, small patch of grass below.		Action-State: 1.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
Woman in red dress taller than woman in yellow.		A medium shot of two women side by side. The woman on the left in a red dress is taller than the woman on the right in a yellow dress. Arms crossed vs. hands in pockets. White wall and black door in background.		Relationship-Comparative: 1.0
A cat made of orange slices.		A medium close-up of a cat made out of orange slices, with a black nose, two green eyes, and a white mouth. Slices arranged to resemble a cat, placed on a white background.		Attribute-Composition: 1.0
A cup full of soda water.		A medium close-up of a white cup with soda water being poured out into a small glass on the right. Cup sitting on a black surface.		Relationship-Containment: 2
A lake shaped like a guitar.		A long, narrow lake shaped like a guitar with a small island forming the guitar body. Calm and clear water reflecting the sky. Trees on the island with a small dock on the shore.		Relationship-Similarity: 0.1
Man with a buzz cut and blue shirt next to a woman with long hair and a yellow shirt.		A man with a buzz cut in a blue shirt stands next to a woman with long hair in a yellow shirt. Both holding white cups with black backpacks. Man looks at woman, she looks back. Tree and fence in background.		Spatial-Binding: 1.0

Original Prompt	Original Image	Enhanced Prompt	Enhanced Image	KP Scores
A race car on a city track, with a mini-map in the top-left corner.		A close-up of a black race car with red and white stripes on a city track. Concrete block track with parked cars on the side. Mini-map in the top-left corner showing the track layout and car's position.		Spatial-Layout: 1.0
A wide shot of The Great Wall of China.		A wide shot of the Great Wall of China visible in the distance, running along the top of a mountain range. Large stone blocks stacked on top of each other, surrounded by green trees and grass. Blue sky with white clouds.		World Knowledge: 1.0
A girl held onto the stem of a huge dandelion with both hands, suspended above the clouds.		A close-up of a girl's hands holding onto the stem of a large, fluffy dandelion. She wears a white shirt and blue shorts, standing on a cloud made of tiny white dots. Blue sky with no other clouds visible.		Reasoning-Counterfactual: 0.5
Poster with text "Game of Thrones" at the bottom.		A poster with the text "Game of Thrones" in bold white letters at the bottom on a black background. Rectangular shape, matte finish. Text is centered horizontally and vertically.		Text-Rendering: 1.0
Poster of a woman on a throne of waves, text "Game of Thrones" at the bottom.		A digital illustration of a woman in a blue dress sitting on a throne made of waves, with waves crashing around it. The text "Game of Thrones" in bold letters at the bottom of the image.		Text-Layout: 1.0

A.3 SigLIP-Only Reward Results

Table 8: Prompt Evaluation using SigLIP Reward Function and Novaeus Model. A.K.S denotes *Average Keypoint Score*. SigLIP_C denotes SigLIP Concept score.

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A bowl of beef noodles, no scallions.	A steaming bowl of beef noodles sits on a white ceramic plate, the rich brown broth glistening with fat droplets... garnished with a sprinkle of red chili flakes, but no scallions are visible...		A detailed, vivid image of a bowl of beef noodles with no scallions on a clean, white background... focus on the dish itself...		Before RL: A.K.S: 0.5866 SigLIP _C : 0.7374 After RL: A.K.S: 0.375 SigLIP _C : 0.9299
A boxer lands a punch on a punching bag.	A powerful boxer delivers a fierce left hook that sends shockwaves through the heavy punching bag... surrounded by rows of similar bags and the faint echoes of past training sessions.		A vivid image of a boxer delivering a powerful punch in a boxing gym... capturing the raw power and skill of the boxer.		Before RL: A.K.S: 0.6763 SigLIP _C : 0.7834 After RL: A.K.S: 0.94 SigLIP _C : 0.8913
A cat made of orange slices.	A playful cat made entirely of vibrant, circular orange slices, each slice seamlessly connecting to form the feline body... tail curled upwards, adding dynamic movement.		A vivid image of a cat made entirely from orange slices... soft, fluffy appearance with overlapping slices for texture and depth.		Before RL: A.K.S: 0.7987 SigLIP _C : 0.9998 After RL: A.K.S: 0.8 SigLIP _C : 0.9986

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A cup full of soda water.	A clear glass cup brimming with sparkling soda water, the surface shimmering with tiny bubbles... a gentle, effervescent foam at the rim.		A detailed image of a full cup containing soda water, effervescent and clear, with bubbles rising to the surface... modern and clean design.		Before RL: A.K.S: 0.5639 SigLIP _C : 0.3930 After RL: A.K.S: 0.5934 SigLIP _C : 0.9531
A girl held onto the stem of a huge dandelion with both hands, suspended above the clouds.	A young girl with a joyful smile clings to the thick, golden stem of a colossal dandelion, floating high above fluffy white clouds... cottony seeds drifting gently around her.		A girl suspended in mid-air above a cloud-filled sky, holding the stem of a massive dandelion with both hands... dreamy, ethereal quality.		Before RL: A.K.S: 0.7820 SigLIP _C : 0.9999 After RL: A.K.S: 0.9625 SigLIP _C : 0.9999
A girl performing a Thomas flare.	A young girl performs an acrobatic Thomas flare, her body arched gracefully in a perfect arc... in a blurred, sunlit gymnasium.		A vivid image of a girl in mid-air, executing the Thomas flare gymnastics move... colorful contemporary gymnastics outfit.		Before RL: A.K.S: 0.7896 SigLIP _C : 0.0765 After RL: A.K.S: 0.275 SigLIP _C : 0.0229
A gust of wind blows, cherry blossoms dance in the air.	A gentle breeze stirs, sending a flurry of delicate cherry blossoms swirling through the air like ethereal snowflakes... drifting gently to the ground.		A gust of wind blowing through a cherry blossom tree, causing flowers to dance and float gracefully in the air... capturing the ephemeral beauty.		Before RL: A.K.S: 0.7963 SigLIP _C : 0.5764 After RL: A.K.S: 0.6939 SigLIP _C : 0.9316

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A hand using chopsticks to pick up food.	A skilled hand delicately maneuvers sleek, black bamboo chopsticks, gently closing around a tender piece of sushi... traditional Japanese dining setting.		A close-up of a hand using chopsticks to pick up food, showing the fingers skillfully manipulating the chopsticks... natural lighting with soft focus background.		Before RL: A.K.S: 0.6360 SigLIP _C : 0.9573 After RL: A.K.S: 0.6676 SigLIP _C : 0.9928
A lake shaped like a guitar.	A serene, glassy lake perfectly mirrors the silhouette of a classic wooden acoustic guitar... surrounded by lush greenery and vibrant wildflowers.		A vivid image of a lake shaped closely like a guitar, with clear, calm water... set against a scenic backdrop of lush greenery and mountains.		Before RL: A.K.S: 0.6774 SigLIP _C : 0.6097 After RL: A.K.S: 0.4 SigLIP _C : 0.9996
A picture with four dogs.	A cozy scene in a sunlit backyard, four dogs of various breeds are playing together... green grass and blooming wildflowers surround them.		A vivid image of four dogs together in the foreground, each with expressive faces and friendly postures... balanced composition with soft, natural lighting.		Before RL: A.K.S: 0.5928 SigLIP _C : 0.5686 After RL: A.K.S: 0.5 SigLIP _C : 0.1032
A puppy happily running.	A joyful golden retriever puppy bounding through a sunny meadow, its tail wagging wildly and ears flopping with each hop... wildflowers dotting the landscape.		A happy puppy running in a green grassy field, wagging its tail and leaving paw prints behind... bright, cheerful color palette.		Before RL: A.K.S: 0.5219 SigLIP _C : 0.9431 After RL: A.K.S: 0.5076 SigLIP _C : 0.9581

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A race car on a city track, with a mini-map in the top-left corner.	A sleek, racing red car zips around a winding city track lined with tall, modern skyscrapers... mini-map in top-left corner marking current position.		A race car speeding on a city track in a futuristic urban environment... mini-map in the top-left corner showing the track layout.		Before RL: A.K.S: 0.8151 SigLIP _C : 0.9775 After RL: A.K.S: 0.7932 SigLIP _C : 0.9975
A strong man, low-angle shot, with a contemptuous expression.	A powerful man shot from a low angle, standing with his arms crossed and a sneer on his face, eyes narrowed in disdain...		A strong, muscular man in a low-angle shot, eyes narrowed in disdain, jaw clenched... dramatic lighting emphasizing powerful presence.		Before RL: A.K.S: 0.8080 SigLIP _C : 0.3048 After RL: A.K.S: 0.8333 SigLIP _C : 0.8836
A wide shot of The Great Wall of China.	A sweeping panoramic view of the Great Wall of China stretching across the landscape, with rugged mountains in the background... a few distant figures walking along the wall.		A wide shot of The Great Wall of China, winding through mountains and valleys... sunny day, high resolution and clarity.		Before RL: A.K.S: 0.8350 SigLIP _C : 0.9563 After RL: A.K.S: 0.8659 SigLIP _C : 0.9412
An ice sculpture of an eagle.	A majestic ice sculpture of an eagle, its wings spread wide in a powerful pose... snowflakes gently falling around the sculpture.		An intricately detailed ice sculpture of an eagle in a winter landscape, clear and pristine ice reflecting the surrounding environment...		Before RL: A.K.S: 0.7545 SigLIP _C : 0.7835 After RL: A.K.S: 0.8604 SigLIP _C : 0.9961

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Eight galloping horses in Chinese ink wash.	Eight powerful horses, their manes and tails flowing freely as they gallop across a vast, misty landscape, rendered in deep, rich tones of Chinese ink wash...		Eight galloping horses in Chinese ink wash style... smooth gradations of black and gray tones conveying depth and texture.		Before RL: A.K.S: 0.8395 SigLIP _C : 0.3183 After RL: A.K.S: 0.775 SigLIP _C : 0.9530
Einstein looking at Hawking.	Einstein, with his signature wild hair and intense gaze, is peering curiously at Stephen Hawking... both men lost in deep intellectual conversation.		Albert Einstein gazing intently at Hawking, standing or sitting next to him... thoughtful look with a soft, warm lighting.		Before RL: A.K.S: 0.7295 SigLIP _C : 0.9997 After RL: A.K.S: 0.68 SigLIP _C : 0.9416
Five people all wearing red clothes.	Five individuals in vibrant red garments gather around a festive table, each garment a deep crimson... against a backdrop of warm, golden sunlight.		Five people all dressed in red clothing, balanced composition, each person in a red outfit... high level of detail and clarity.		Before RL: A.K.S: 0.7086 SigLIP _C : 0.2367 After RL: A.K.S: 0.7 SigLIP _C : 0.9936
Man with a buzz cut and blue shirt next to a woman with long hair and a yellow shirt.	A man with a clean buzz cut wearing a fitted blue shirt stands beside a woman with long, flowing blonde hair in a vibrant yellow shirt... standing in a sunlit park.		A man with a buzz cut wearing a blue shirt standing next to a woman with long hair wearing a yellow shirt... natural lighting, sharp focus.		Before RL: A.K.S: 0.6121 SigLIP _C : 0.8556 After RL: A.K.S: 0.8333 SigLIP _C : 0.9957

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Poster of a woman on a throne of waves, text “Game of Thrones” at the bottom.	A grand poster depicts a majestic woman seated on a throne crafted from swirling, turbulent waves... words “Game of Thrones” written in elegant script at the bottom.		A poster of a woman seated on a throne made entirely of waves... text “Game of Thrones” prominently displayed at the bottom in bold letters.		Before RL: A.K.S: 0.8460 SigLIP _C : 0.3865 After RL: A.K.S: 0.65 SigLIP _C : 0.8161
Poster with text “Game of Thrones” at the bottom.	A large, eye-catching poster featuring the iconic title “Game of Thrones” in bold, metallic gold letters at the bottom... a fierce dragon breaches the sky.		A poster with “Game of Thrones” prominently at the bottom in a bold, dramatic font... dark, moody background with epic medieval elements.		Before RL: A.K.S: 0.7818 SigLIP _C : 0.2591 After RL: A.K.S: 0.96 SigLIP _C : 0.3011
The large ball broke the table because it was made of metal.	A massive metallic ball, gleaming with a reflective sheen, sits in the foreground pushing against a wooden table... splintered and cracked, on the verge of breaking.		A large metal ball that has broken through a table, ball depicted with metallic sheen, table shown in pieces... bright, balanced composition.		Before RL: A.K.S: 0.6528 SigLIP _C : 0.9947 After RL: A.K.S: 0.98 SigLIP _C : 0.9932
Two large spheres.	Two colossal spheres sit side by side on a flat desert landscape, smooth and reflective, catching the light... one deep blue, the other glowing with a warm, golden hue.		Two large spheres, round and smooth with a reflective surface... placed on a clean, flat background with a modern and minimalist aesthetic.		Before RL: A.K.S: 0.6049 SigLIP _C : 0.9642 After RL: A.K.S: 0.5 SigLIP _C : 0.9497

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Woman in red dress taller than woman in yellow.	A striking image of two women on a bustling city street. The woman in the red dress towers over her companion in the yellow dress... both striding with purpose.		Two women standing next to each other, the first in a vibrant red dress standing taller than the second in a bright yellow dress... emphasizing the height difference.		Before RL: A.K.S: 0.6763 SigLIP _C : 0.9998 After RL: A.K.S: 0.98 SigLIP _C : 0.9991

Table 9: Prompt Evaluation using SigLIP Reward Function and Qwen Model. A.K.S denotes *Average Keypoint Score*. SigLIP_C denotes SigLIP Concept score.

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A bowl of beef noodles, no scallions.	A steaming bowl of savory beef noodles in a rustic ceramic bowl, tender slow-cooked beef in a rich dark brown broth, garnished with toasted sesame seeds and a wedge of lime...		A steaming bowl of beef noodles on a wooden table, sprinkle of black pepper on top, long thin noodles in a savory beef broth with tender thinly sliced beef scattered throughout.		Before RL: A.K.S: 0.6264 SigLIP _C : 0.4202 After RL: A.K.S: 0.7253 SigLIP _C : 0.7823
Five people all wearing red clothes.	Five people dressed in vibrant red clothing, each in a different shade from deep burgundy to bright cherry, in a bustling city square... lively and joyful atmosphere.		Five people standing together, each wearing bright red clothing, with bold red scarves and accessories, in a vibrant city street with tall skyscrapers in the background.		Before RL: A.K.S: 0.7605 SigLIP _C : 0.9928 After RL: A.K.S: 0.7616 SigLIP _C : 0.9999

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
The large ball broke the table because it was made of metal.	A gleaming metallic ball atop a sturdy wooden table, surface scratched and dented, with intricate engravings. The table now lies broken and splintered on the floor...		A large metal ball with a dull grey finish, shattered on the floor with fragments scattered. The sturdy wooden table lies broken in two, legs twisted at unnatural angles from the impact.		Before RL: A.K.S: 0.7586 SigLIP _C : 0.9801 After RL: A.K.S: 0.7615 SigLIP _C : 0.9996
Two large spheres.	Two massive, iridescent spheres, each about the size of a small house, hover in a serene, starry night sky, their translucent surfaces etched with swirling patterns...		Two large, glowing spheres, one slightly larger than the other, suspended in mid-air, the smaller sphere rotating slowly around the larger one creating a dance of light and shadow.		Before RL: A.K.S: -0.1919 SigLIP _C : 0.0013 After RL: A.K.S: -0.0894 SigLIP _C : 0.0443
An ice sculpture of an eagle.	A crystal-encrusted ice sculpture of an eagle, wings outstretched with feathers of thin lacy frost, sharp beak, glittering ice eyes, set against a frosty blue backdrop with swirling patterns...		A majestic ice sculpture of a regal eagle, crystalline body glistening, feathers intricately carved, wings spread wide as if frozen in mid-flight. Set against frosted glass with subtle reflections.		Before RL: A.K.S: 0.7394 SigLIP _C : 0.8994 After RL: A.K.S: 0.7612 SigLIP _C : 0.9982

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A picture with four dogs.	Four dogs—a golden retriever, poodle, bulldog, and greyhound—sit on a weathered wooden dock surrounded by bare winter trees. All wearing matching red bandanas...		Four dogs of different breeds and sizes sitting on lush green grass in a sunny meadow, surrounded by vibrant wildflowers and a few tall trees, with a small wooden fence in the background.		Before RL: A.K.S: 0.7459 SigLIP _C : 0.8967 After RL: A.K.S: 0.7507 SigLIP _C : 0.9309
A strong man, low-angle shot, with a contemptuous expression.	A rugged, muscular figure stands tall, gazing down with a scowl, his face illuminated by a single piercing light. Eyes narrow, brow furrowed, lips curling into a sneer...		A powerful, ruggedly handsome man looking directly down at the camera from a low angle, eyes narrowed in disdain, jaw clenched, conveying unyielding contempt. Dark neutral background.		Before RL: A.K.S: 0.4254 SigLIP _C : 0.0018 After RL: A.K.S: 0.6858 SigLIP _C : 0.7402
Eight galloping horses in Chinese ink wash.	Eight majestic horses gallop across a misty, moonlit landscape rendered in delicate Chinese ink wash, with flowing manes and feathery brushstrokes... craggy mountains and twisted bamboo backdrop.		Eight galloping horses in traditional Chinese ink wash, set against a misty mountain range with twisted bamboo and blooming cherry blossoms, on a soft gradient blue background.		Before RL: A.K.S: 0.6257 SigLIP _C : 0.2429 After RL: A.K.S: 0.7321 SigLIP _C : 0.8139

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A girl performing a Thomas flare.	A young woman in a sleek black jumpsuit in a dimly lit industrial setting, grasping a spark plug and pulling an ignition cord, unleashing a brilliant flash of light...		A young woman in a yellow and orange jumpsuit on a desert highway at sunset, igniting a large pyrotechnic flare, flames erupting upward, arms outstretched, hair blowing in the wind.		Before RL: A.K.S: 0.5733 SigLIP _C : 0.0851 After RL: A.K.S: 0.7582 SigLIP _C : 0.9768
A hand using chopsticks to pick up food.	A slender, dexterous hand grasps a pair of delicate curved chopsticks, wooden slats worn smooth, fingers providing gentle pressure, thumb acting as a fulcrum for delicate balance...		A delicate hand grasping chopsticks, fingers curled around the handles, the tips delicately plucking a single piece of food from a steaming plate. Asian-inspired table setting.		Before RL: A.K.S: 0.7063 SigLIP _C : 0.6589 After RL: A.K.S: 0.7607 SigLIP _C : 0.9954
A puppy happily running.	A playful puppy with a fluffy white coat and floppy ears runs across a lush green meadow chasing a fluttering butterfly, tail wagging furiously, paws barely touching the ground...		A small puppy with a fluffy white and brown coat running in a lush green meadow, tongue lolling out, ears flapping wildly. Sunlight casts a warm glow on its fur, highlighting joyful expression.		Before RL: A.K.S: 0.7391 SigLIP _C : 0.8569 After RL: A.K.S: 0.7568 SigLIP _C : 0.9735

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A boxer lands a punch on a punching bag.	A dimly lit boxing gym, the air thick with sweat. The boxer swings with all his might, connecting with a resounding thud as his fist crashes into the punching bag... muscles bulging.		A heavy-duty punching bag in a dimly lit industrial-style gym. A determined boxer throws a powerful hook punch, connecting with a loud thud, bag swinging violently back toward the boxer.		Before RL: A.K.S: 0.7427 SigLIP _C : 0.9335 After RL: A.K.S: 0.7463 SigLIP _C : 0.9591
Einstein looking at Hawking.	Einstein sits in a plush dimly lit armchair, intently focused on Hawking in his wheelchair. Stephen Hawking's face illuminated by a nearby lamp, thin fingers drumming on his armrests...		Albert Einstein in a leather armchair, wearing wire-rimmed glasses, gazing intently at a wheelchair-bound Stephen Hawking, who sits across from him with a faint smile, eyes fixed on Einstein.		Before RL: A.K.S: 0.7612 SigLIP _C : 0.9972 After RL: A.K.S: 0.7614 SigLIP _C : 0.9988
A gust of wind blows, cherry blossoms dance in the air.	A delicate cherry blossom drifts lazily through the air, caught in a whirlwind as if frozen in time. Surrounding blossoms swirl and twirl in ethereal light... a soft, gentle hum in the air.		A delicate dance of pink and white cherry blossoms suspended in mid-air, gently swaying in the breeze, shimmering in soft focus against a pale blue sky with dappled sunlight.		Before RL: A.K.S: 0.6081 SigLIP _C : 0.2661 After RL: A.K.S: 0.7409 SigLIP _C : 0.9196

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Woman in red dress taller than woman in yellow.	A statuesque woman in a bold crimson red evening gown stands tall, towering over a petite woman in a vibrant yellow sundress who looks up with admiration and envy...		A woman in a bold red dress standing tall, dwarfing a woman in a bright yellow dress beside her. The red-dressed woman towers over her counterpart, vibrant attire drawing attention.		Before RL: A.K.S: 0.7213 SigLIP _C : 0.9956 After RL: A.K.S: 0.7615 SigLIP _C : 0.9999
A cat made of orange slices.	A feline figure composed of a mosaic of vibrant orange slices in a swirling organic pattern forming the body, head, and limbs. Eyes are bright shiny marbles embedded in patterned slices...		A whimsical anthropomorphic cat with a body of various orange slices, overlapping and stacked to form the feline shape. Eyes and whiskers made of thin orange peel slices, ears and tail also of orange slices.		Before RL: A.K.S: 0.7615 SigLIP _C : 0.9995 After RL: A.K.S: 0.7615 SigLIP _C : 0.9997
A cup full of soda water.	A clear glass cylinder filled to the brim with pale, bubbly soda water, bubbles clinging to the rim like translucent pearls, ice cubes floating listlessly at the bottom...		A glass filled to the brim with clear, bubbly soda water, with condensation beading on the outside of the glass and a few ice cubes floating gently at the bottom.		Before RL: A.K.S: 0.6434 SigLIP _C : 0.4301 After RL: A.K.S: 0.7421 SigLIP _C : 0.8845

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A lake shaped like a guitar.	A serene turquoise lake in a lush green valley, its surface resembling the curves of a guitar, with gentle ripples reflecting surrounding trees like patterns on a fretboard...		A serene lake shaped like a guitar, body curved in a gentle acoustic manner, neck stretching to the horizon like guitar frets. Calm still surface reflecting the sky, lined with lush greenery.		Before RL: A.K.S: 0.7528 SigLIP _C : 0.9417 After RL: A.K.S: 0.7615 SigLIP _C : 0.9993
Man with a buzz cut and blue shirt next to a woman with long hair and a yellow shirt.	A rugged man with a buzzed dark hairline sits casually next to a stunning woman with cascading honey-blond hair. She wears a sunshine-yellow sundress, his hand resting on her knee...		A man with a buzz cut wearing a blue shirt sitting next to a woman with long, flowing hair wearing a yellow shirt.		Before RL: A.K.S: 0.2725 SigLIP _C : 0.1267 After RL: A.K.S: 0.7613 SigLIP _C : 0.9986
A race car on a city track, with a mini-map in the top-left corner.	A sleek metallic race car zooms down the asphalt reflecting vibrant city lights. A miniature map in the top-left corner glows with a soft blue light showing the car's progress and upcoming turns...		A sleek race car speeding around a bustling city track, mini-map in the top-left corner of the windshield showing a 3D representation of the track. Sun casts a golden glow over towering skyscrapers.		Before RL: A.K.S: 0.7456 SigLIP _C : 0.9301 After RL: A.K.S: 0.7613 SigLIP _C : 0.9997

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A wide shot of The Great Wall of China.	A majestic serpentine structure stretches across sweeping mountains and mist-shrouded valleys, the sun casting a warm golden glow upon its weathered brickwork... snow-capped peaks in the distance.		A sweeping panoramic view of The Great Wall of China against a clear blue sky, its serpentine structure stretching across mountains and rolling hills, with a hint of mist in the valleys below, at sunrise.		Before RL: A.K.S: 0.6490 SigLIP _C : 0.9186 After RL: A.K.S: 0.7533 SigLIP _C : 0.9731
A girl held onto the stem of a huge dandelion with both hands, suspended above the clouds.	A delicate ethereal girl, arms wrapped tightly around the thick sturdy stem of a gigantic dandelion, its bright yellow petals against soft white clouds beneath her... eyes closed in pure joy.		A young girl gazing upwards, holding the stem of a massive dandelion with both hands, arms stretched upward, flower suspended in mid-air surrounded by fluffy white clouds in every direction.		Before RL: A.K.S: 0.7608 SigLIP _C : 0.9993 After RL: A.K.S: 0.7615 SigLIP _C : 0.9999
Poster with text "Game of Thrones" at the bottom.	A dark medieval cityscape at dusk with an imposing castle, knights streaming out preparing for battle. Characters gathered around a feast table with King Robert Baratheon at the head...		A medieval fantasy landscape with dark ominous sky and towering mountains. Characters around a large wooden table with an ornate throne. Text "Game of Thrones" in bold golden letters at the bottom.		Before RL: A.K.S: 0.3134 SigLIP _C : 0.0905 After RL: A.K.S: 0.7094 SigLIP _C : 0.7096

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Poster of a woman on a throne of waves, text “Game of Thrones” at the bottom.	A regal woman atop a throne of undulating blue and green waves, raven hair flowing, golden crown adorned with shells and pearls... flowing white gown of sea foam. “Game of Thrones” at the bottom.		A regal woman on a throne carved from a majestic wave, draped in ocean-blue robes. Long raven-black hair cascades down her back. “Game of Thrones” emblazoned in bold golden letters at the base.		Before RL: A.K.S: 0.6167 SigLIP _C : 0.9882 After RL: A.K.S: 0.6856 SigLIP _C : 0.9865

A.4 Composite Reward Results

Table 10: Prompt Evaluation using Composite Reward Function. A.K.S denotes *Average Keypoint Score*.

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A bowl of beef noodles, no scallions.	A steaming bowl of savory beef noodles in a rustic ceramic bowl, tender slow-cooked beef in a rich dark brown broth, garnished with toasted sesame seeds and a wedge of lime...		A steaming bowl of beef noodles on a wooden table, sprinkle of black pepper on top, long thin noodles in a savory beef broth with tender thinly sliced beef scattered throughout.		Before RL: A.K.S: 0.6264 After RL: A.K.S: 0.7253

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Five people all wearing red clothes.	Five people dressed in vibrant red clothing, each in a different shade from deep burgundy to bright cherry, in a bustling city square... lively and joyful atmosphere.		Five people standing together, each wearing bright red clothing, with bold red scarves and accessories, in a vibrant city street with tall skyscrapers in the background.		Before RL: A.K.S: 0.7605 After RL: A.K.S: 0.7616
The large ball broke the table because it was made of metal.	A gleaming metallic ball atop a sturdy wooden table, surface scratched and dented, with intricate engravings. The table now lies broken and splintered on the floor...		A large metal ball with a dull grey finish, shattered on the floor with fragments scattered. The sturdy wooden table lies broken in two, legs twisted at unnatural angles from the impact.		Before RL: A.K.S: 0.7586 After RL: A.K.S: 0.7615
Two large spheres.	Two massive, iridescent spheres, each about the size of a small house, hover in a serene, starry night sky, their translucent surfaces etched with swirling patterns...		Two large, glowing spheres, one slightly larger than the other, suspended in mid-air, the smaller sphere rotating slowly around the larger one creating a dance of light and shadow.		Before RL: A.K.S: -0.1919 After RL: A.K.S: -0.0894

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
An ice sculpture of an eagle.	A crystal-encrusted ice sculpture of an eagle, wings outstretched with feathers of thin lacy frost, sharp beak, glittering ice eyes, set against a frosty blue backdrop with swirling patterns...		A majestic ice sculpture of a regal eagle, crystalline body glistening, feathers intricately carved, wings spread wide as if frozen in mid-flight. Set against frosted glass with subtle reflections.		Before RL: A.K.S: 0.7394 After RL: A.K.S: 0.7612
A picture with four dogs.	Four dogs—a golden retriever, poodle, bulldog, and greyhound—sit on a weathered wooden dock surrounded by bare winter trees. All wearing matching red bandanas...		Four dogs of different breeds and sizes sitting on lush green grass in a sunny meadow, surrounded by vibrant wildflowers and a few tall trees, with a small wooden fence in the background.		Before RL: A.K.S: 0.7459 After RL: A.K.S: 0.7507
A strong man, low-angle shot, with a contemptuous expression.	A rugged, muscular figure stands tall, gazing down with a scowl, his face illuminated by a single piercing light. Eyes narrow, brow furrowed, lips curling into a sneer...		A powerful, ruggedly handsome man looking directly down at the camera from a low angle, eyes narrowed in disdain, jaw clenched, conveying unyielding contempt. Dark neutral background.		Before RL: A.K.S: 0.4254 After RL: A.K.S: 0.6858

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Eight galloping horses in Chinese ink wash.	Eight majestic horses gallop across a misty, moonlit landscape rendered in delicate Chinese ink wash, with flowing manes and feathery brushstrokes... craggy mountains and twisted bamboo backdrop.		Eight galloping horses in traditional Chinese ink wash, set against a misty mountain range with twisted bamboo and blooming cherry blossoms, on a soft gradient blue background.		Before RL: A.K.S: 0.6257 After RL: A.K.S: 0.7321
A girl performing a Thomas flare.	A young woman in a sleek black jumpsuit in a dimly lit industrial setting, grasping a spark plug and pulling an ignition cord, unleashing a brilliant flash of light...		A young woman in a yellow and orange jumpsuit on a desert highway at sunset, igniting a large pyrotechnic flare, flames erupting upward, arms outstretched, hair blowing in the wind.		Before RL: A.K.S: 0.5733 After RL: A.K.S: 0.7582
A hand using chopsticks to pick up food.	A slender, dexterous hand grasps a pair of delicate curved chopsticks, wooden slats worn smooth, fingers providing gentle pressure, thumb acting as a fulcrum for delicate balance...		A delicate hand grasping chopsticks, fingers curled around the handles, the tips delicately plucking a single piece of food from a steaming plate. Asian-inspired table setting.		Before RL: A.K.S: 0.7063 After RL: A.K.S: 0.7607

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A puppy happily running.	A playful puppy with a fluffy white coat and floppy ears runs across a lush green meadow chasing a fluttering butterfly, tail wagging furiously, paws barely touching the ground...		A small puppy with a fluffy white and brown coat running in a lush green meadow, tongue lolling out, ears flapping wildly. Sunlight casts a warm glow on its fur, highlighting joyful expression.		Before RL: A.K.S: 0.7391 After RL: A.K.S: 0.7568
A boxer lands a punch on a punching bag.	A dimly lit boxing gym, the air thick with sweat. The boxer swings with all his might, connecting with a resounding thud as his fist crashes into the punching bag... muscles bulging.		A heavy-duty punching bag in a dimly lit industrial-style gym. A determined boxer throws a powerful hook punch, connecting with a loud thud, bag swinging violently back toward the boxer.		Before RL: A.K.S: 0.7427 After RL: A.K.S: 0.7463
Einstein looking at Hawking.	Einstein sits in a plush dimly lit armchair, intently focused on Hawking in his wheelchair. Stephen Hawking's face illuminated by a nearby lamp, thin fingers drumming on his armrests...		Albert Einstein in a leather armchair, wearing wire-rimmed glasses, gazing intently at a wheelchair-bound Stephen Hawking, who sits across from him with a faint smile, eyes fixed on Einstein.		Before RL: A.K.S: 0.7612 After RL: A.K.S: 0.7614

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A gust of wind blows, cherry blossoms dance in the air.	A delicate cherry blossom drifts lazily through the air, caught in a whirlwind as if frozen in time. Surrounding blossoms swirl and twirl in ethereal light... a soft, gentle hum in the air.		A delicate dance of pink and white cherry blossoms suspended in mid-air, gently swaying in the breeze, shimmering in soft focus against a pale blue sky with dappled sunlight.		Before RL: A.K.S: 0.6081 After RL: A.K.S: 0.7409
Woman in red dress taller than woman in yellow.	A statuesque woman in a bold crimson red evening gown stands tall, towering over a petite woman in a vibrant yellow sundress who looks up with admiration and envy...		A woman in a bold red dress standing tall, dwarfing a woman in a bright yellow dress beside her. The red-dressed woman towers over her counterpart, vibrant attire drawing attention.		Before RL: A.K.S: 0.7213 After RL: A.K.S: 0.7615
A cat made of orange slices.	A feline figure composed of a mosaic of vibrant orange slices in a swirling organic pattern forming the body, head, and limbs. Eyes are bright shiny marbles embedded in patterned slices...		A whimsical anthropomorphic cat with a body of various orange slices, overlapping and stacked to form the feline shape. Eyes and whiskers made of thin orange peel slices, ears and tail also of orange slices.		Before RL: A.K.S: 0.7615 After RL: A.K.S: 0.7615

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A cup full of soda water.	A clear glass cylinder filled to the brim with pale, bubbly soda water, bubbles clinging to the rim like translucent pearls, ice cubes floating listlessly at the bottom...		A glass filled to the brim with clear, bubbly soda water, with condensation beading on the outside of the glass and a few ice cubes floating gently at the bottom.		Before RL: A.K.S: 0.6434 After RL: A.K.S: 0.7421
A lake shaped like a guitar.	A serene turquoise lake in a lush green valley, its surface resembling the curves of a guitar, with gentle ripples reflecting surrounding trees like patterns on a fretboard...		A serene lake shaped like a guitar, body curved in a gentle acoustic manner, neck stretching to the horizon like guitar frets. Calm still surface reflecting the sky, lined with lush greenery.		Before RL: A.K.S: 0.7528 After RL: A.K.S: 0.7615
Man with a buzz cut and blue shirt next to a woman with long hair and a yellow shirt.	A rugged man with a buzzed dark hairline sits casually next to a stunning woman with cascading honey-blonde hair. She wears a sunshine-yellow sundress, his hand resting on her knee...		A man with a buzz cut wearing a blue shirt sitting next to a woman with long, flowing hair wearing a yellow shirt.		Before RL: A.K.S: 0.2725 After RL: A.K.S: 0.7613

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
A race car on a city track, with a mini-map in the top-left corner.	A sleek metallic race car zooms down the asphalt reflecting vibrant city lights. A miniature map in the top-left corner glows with a soft blue light showing the car's progress and upcoming turns...		A sleek race car speeding around a bustling city track, mini-map in the top-left corner of the windshield showing a 3D representation of the track. Sun casts a golden glow over towering skyscrapers.		Before RL: A.K.S: 0.7456 After RL: A.K.S: 0.7613
A wide shot of The Great Wall of China.	A majestic serpentine structure stretches across sweeping mountains and mist-shrouded valleys, the sun casting a warm golden glow upon its weathered brickwork... snow-capped peaks in the distance.		A sweeping panoramic view of The Great Wall of China against a clear blue sky, its serpentine structure stretching across mountains and rolling hills, with a hint of mist in the valleys below, at sunrise.		Before RL: A.K.S: 0.6490 After RL: A.K.S: 0.7533
A girl held onto the stem of a huge dandelion with both hands, suspended above the clouds.	A delicate ethereal girl, arms wrapped tightly around the thick sturdy stem of a gigantic dandelion, its bright yellow petals against soft white clouds beneath her... eyes closed in pure joy.		A young girl gazing upwards, holding the stem of a massive dandelion with both hands, arms stretched upward, flower suspended in mid-air surrounded by fluffy white clouds in every direction.		Before RL: A.K.S: 0.7608 After RL: A.K.S: 0.7615

Original Prompt	Enhanced Prompt (Before RL)	Image (Before RL)	Enhanced Prompt (After RL)	Image (After RL)	Scores
Poster with text “Game of Thrones” at the bottom.	A dark medieval cityscape at dusk with an imposing castle, knights streaming out preparing for battle. Characters gathered around a feast table with King Robert Baratheon at the head...		A medieval fantasy landscape with dark ominous sky and towering mountains. Characters around a large wooden table with an ornate throne. Text “Game of Thrones” in bold golden letters at the bottom.		Before RL: A.K.S: 0.3134 After RL: A.K.S: 0.7094
Poster of a woman on a throne of waves, text “Game of Thrones” at the bottom.	A regal woman atop a throne of undulating blue and green waves, raven hair flowing, golden crown adorned with shells and pearls... flowing white gown of sea foam. “Game of Thrones” at the bottom.		A regal woman on a throne carved from a majestic wave, draped in ocean-blue robes. Long raven-black hair cascades down her back. “Game of Thrones” emblazoned in bold golden letters at the base.		Before RL: A.K.S: 0.6167 After RL: A.K.S: 0.6856

References

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
- [2] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.
- [3] Yujie Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.
- [4] Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. Beautiful-

prompt: Towards automatic prompt enhancement for text-to-image synthesis. *arXiv preprint arXiv:2311.06752*, 2023.

- [5] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.