# PROMPTECHO: ANNOTATION-FREE REWARD FROM VISION-LANGUAGE MODELS FOR TEXT-TO-IMAGE REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning can improve prompt following in text-to-image (T2I) models, but obtaining high-quality reward signals remains a key challenge. Existing rewards either provide only coarse-grained global alignment signals or rely on costly human preference annotations and additional reward-model training. We introduce PromptEcho, an annotation-free reward that uses a frozen vision-language model (VLM) to compute the image-conditioned, teacher-forced next-token likelihood of the original prompt. This directly converts the VLM's pretraining objective into a continuous token-level reward without training a separate reward model. For evaluation, we develop DenseAlignBench, a benchmark of concept-rich dense captions designed to rigorously test fine-grained prompt following. Relative to the corresponding original baselines, PromptEcho achieves net pairwise advantages of +26.8pp and +16.3pp on DenseAlignBench for Z-Image and QwenImage-2512, respectively, and yields consistent improvements on public benchmarks including GenEval, DPG-Bench, and TIIFBench without any task-specific training. Under the same Z-Image RL configuration, PromptEcho also substantially outperforms CLIP score, ImageReward-v1, and HPSv3, with a particularly pronounced advantage on long, dense prompts. Ablations with VLMs of different sizes further show that PromptEcho can directly benefit from advances in open-source VLMs. Code, checkpoints, and DenseAlignBench are available at `https://github.com/LJLwykqh/prompt_echo`.

## 1 INTRODUCTION

Text-to-image (T2I) generation has progressed rapidly from GANs to diffusion and flow-matching models. Nevertheless, precise prompt following remains difficult: state-of-the-art systems still omit objects, bind attributes incorrectly, confuse spatial relations, or fail at counting when prompts contain many interacting constraints.

Recent RL post-training methods, including DDPO (Black et al., 2024), DPOK (Fan et al., 2023), Flow-GRPO (Liu et al., 2025), and AWM (Xue et al., 2025), have improved prompt following in T2I models by optimizing external feedback. Across these frameworks, performance depends critically on the quality of the external reward.

Yet high-quality reward signals are often costly to obtain. CLIP score (Hessel et al., 2021) is efficient but provides only coarse global image-text similarity. Preference models such as PickScore (Kirstain et al., 2023), ImageReward (Xu et al., 2023), and HPS (Wu et al., 2023), as well as VLM-based reward models such as VisionReward (Xu et al., 2024), RewardDance (Wu et al., 2025b), and UnifiedReward (Wang et al., 2025), obtain stronger evaluation capabilities through dedicated annotations and reward-model training. Zero-shot VLM scoring avoids this training, but typically generates evaluation text autoregressively and expresses the final judgment as a discrete rating. This motivates a reward that directly exposes a pretrained VLM's alignment knowledge without annotation or reward-model training.

**PromptEcho.** We propose PromptEcho, a reward construction paradigm that requires neither human annotation nor reward-model training. Given an image $x$ generated from prompt $c$, PromptEcho feeds $x$ and a guiding query $q$ (e.g., "Describe this image in detail.") into a frozen VLM and uses $c$ as the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

teacher-forced target. The average token-level log-likelihood of $c$ serves as the reward. Intuitively, if the image faithfully depicts the prompt, the VLM should assign high probability to the corresponding prompt tokens—the image *echoes* the prompt through the VLM. In every RL group, $c$, $q$, the frozen VLM, and tokenization remain fixed while only the generated image changes, so group-relative optimization directly uses image-conditioned likelihood differences. PromptEcho thereby exposes pretrained image-text alignment knowledge as a continuous, deterministic token-level reward through one teacher-forcing forward pass, and stronger frozen VLMs can be adopted without collecting new annotations or retraining a reward model.

**Contributions.**

- PromptEcho is the first method to use a frozen VLM's image-conditioned, teacher-forced next-token likelihood of the original prompt as an annotation-free, token-level reward for online RL post-training of T2I models.
- PromptEcho requires neither human reward annotations nor additional reward-model training. It provides a continuous, fine-grained signal through a standard teacher-forcing forward pass and can directly use stronger frozen VLMs without changing the RL framework.
- Across two strong T2I base models, PromptEcho improves DenseAlignBench by +26.8pp and +16.3pp and yields consistent gains across three public benchmarks without any task-specific training. Comparisons with multiple publicly available rewards under the same RL configuration, together with the text-rendering optimization experiment, further demonstrate its effectiveness and broad applicability.

## 2 Related Work

### 2.1 Reinforcement Learning for Text-to-Image Models

Research on applying reinforcement learning to diffusion models has progressed rapidly. DDPO (Black et al., 2024) first formalizes diffusion denoising as a multi-step Markov Decision Process, treating each denoising step as an action and optimizing external rewards with policy gradients. DPOK (Fan et al., 2023) extends DDPO with KL regularization, constraining the optimized policy toward the pretrained distribution to preserve image quality and diversity while improving target attributes.

RL optimization has subsequently been extended to flow-matching models. Flow-GRPO (Liu et al., 2025) adapts Group Relative Policy Optimization to continuous flow matching and updates the policy using group-relative advantages. AWM (Xue et al., 2025) performs advantage-weighted matching directly in the ODE prediction space. PromptEcho is complementary to these optimization algorithms: it provides the reward signal for group-relative RL, and our main experiments use AWM.

### 2.2 Reward Signal Sources

**Traditional reward metrics.** CLIP score (Hessel et al., 2021) directly computes global image-text embedding similarity using a pretrained CLIP model and requires no dedicated reward-model training, but provides a relatively coarse reward signal. PickScore (Kirstain et al., 2023), ImageReward (Xu et al., 2023), and HPS (Wu et al., 2023) train dedicated reward models from human preference comparisons. Diffusion-DPO (Wallace et al., 2024) and Diffusion-RPO (Gu et al., 2024) bypass an explicit reward model but still rely on human-annotated preference pairs.

**VLM-based rewards.** Recent work raises reward-model capacity with VLMs. VisionReward (Xu et al., 2024) fine-tunes a VLM to answer fine-grained binary questions and aggregates their probabilities. RewardDance (Wu et al., 2025b) and UnifiedReward (Wang et al., 2025) train VLMs on annotated preference or instruction data to predict holistic quality scores. These methods leverage the strong multimodal representations of VLMs, but still require purpose-built annotations and additional supervised training of the reward model. Alternatively, a general-purpose VLM can be used directly for zero-shot scoring without additional training. To evaluate this approach, we construct the InferScore baseline, which uses a detailed prompt-following rubric to autoregressively generate a brief rationale and a discrete rating from 1 to 5.

**PromptEcho.** PromptEcho introduces a distinct reward interface: it directly converts the pretrained conditional likelihood of a frozen VLM into a continuous, token-level, and deterministic signal. It requires neither preference annotations nor reward-model training. Every target token in the original prompt contributes directly to the reward, providing a fine-grained token-level signal; meanwhile, a stronger open-source VLM can be adopted simply by replacing the frozen scorer. A recent preprint, SpectraReward (Huang et al., 2026), explores a closely related reward formulation based on image-conditioned teacher-forced prompt likelihood.

### 2.3 Vision-Language Models

**Vision-language models.** The rapid development of open-source VLMs provides a strong foundation for PromptEcho. Representative families include LLaVA (Liu et al., 2023), which established the visual instruction-tuning paradigm; InternVL (Chen et al., 2023), which scales visual encoders and multimodal reasoning; and Qwen-VL (Bai et al., 2023), which has evolved through Qwen2-VL, Qwen2.5-VL, and Qwen3-VL with continued improvements in scale, data, and multimodal understanding. We use a frozen Qwen3-VL-32B-Instruct for the main reward. PromptEcho is model-agnostic: a stronger open-source VLM can be adopted by replacing the frozen scorer, without recollecting preference annotations or retraining a reward model. Our 8B/32B ablation empirically demonstrates this extensibility.

## 3 Method

### 3.1 PromptEcho Reward

Let $G_\theta$ denote a T2I model, c$\sim$D an input prompt, and x$\sim G_\theta$(c) an image generated from c. T2I RL seeks to maximize the expected reward:

$$\theta^\star = \arg\max_\theta \ \mathbb{E}_{c\sim\mathcal{D},\, x\sim G_\theta(c)}[R(x,c)] \tag{1}$$

PromptEcho constructs this reward from a frozen VLM $p_\phi$, whose parameters remain fixed throughout RL.

**Input construction.** To score a generated image x, the VLM receives three components: (1) the image x; (2) a guiding query q, such as "Describe this image in detail."; and (3) the original prompt c as the teacher-forced target. Rather than asking the VLM to generate a new description or an evaluation score, PromptEcho evaluates the conditional likelihood of every target token in c through a single teacher-forcing forward pass.

**PromptEcho reward definition.** The reward is the negative VLM pretraining NTP loss over the original prompt c:

$$R_\phi(x,c,q) = -\mathcal{L}_{\mathrm{NTP}}(c \mid x,q) = \frac{1}{T}\sum_{t=1}^{T} \log p_\phi(c_t \mid x,q,c_{<t}) \tag{2}$$

Here, $T$ is the number of tokens in $c$, and $p_\phi(c_t \mid x,q,c_{<t})$ is the probability assigned by the frozen VLM to the $t$-th prompt token given the image, guiding query, and preceding prompt tokens.

**Group-relative optimization.** For each prompt $c$, the policy samples a group of $K$ images $\{x_j\}_{j=1}^K$ and computes a PromptEcho reward for each image. The group-centered advantage is

$$A_j = R_\phi(x_j,c,q) - \frac{1}{K}\sum_{k=1}^{K} R_\phi(x_k,c,q) \tag{3}$$

All $K$ samples in a group share the same $c$, $q$, frozen VLM, and tokenization; only $x_j$ changes. The centered advantage therefore ranks the generated images by how each visual input changes the likelihood of the same target sequence, providing a direct group-relative learning signal.

**Intuition.** If image x faithfully depicts prompt c (e.g., correctly rendering all objects, attributes, and spatial relationships), the VLM should "recite" each token of c with high probability after observing the image—the image *echoes* the prompt—yielding a high reward. Conversely, omissions or semantic deviations in the image lower the token-level likelihood and thus the reward. This formulation works because the reward is derived directly from the VLM's next-token prediction objective, tapping into the image-text alignment knowledge acquired during pretraining without any additional conversion layers, regression heads, or fine-tuning. The controlled count-and-color experiment in Appendix E.2 directly validates this intuition: for every image, the semantically matched caption receives the highest reward. Further token-level analysis shows that count or color mismatches primarily increase the negative log-likelihood at the corresponding semantic-token positions, thereby reducing the overall reward.

### 3.2 RL Training Framework

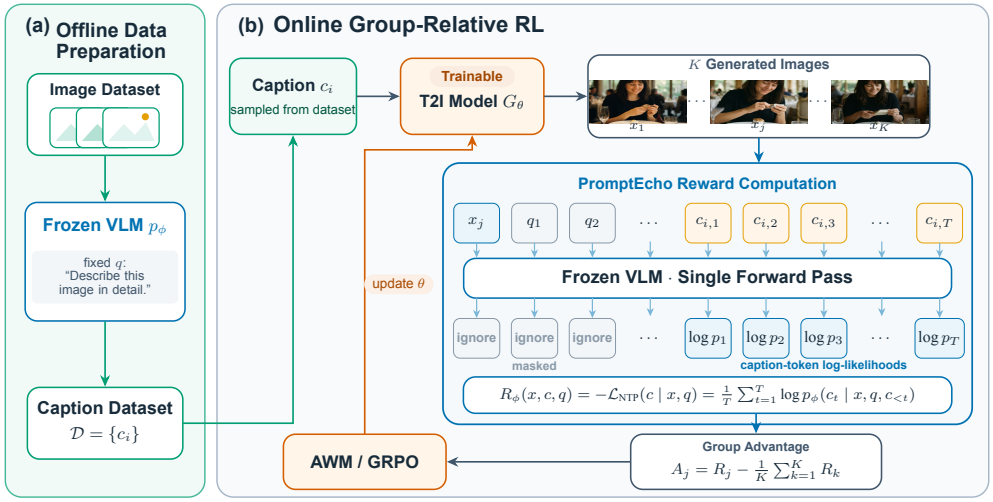



Figure 1: Overview of PromptEcho. **(a)** A frozen VLM converts source images into a caption dataset using a fixed guiding query q. **(b)** For each prompt $c_i$, the T2I model generates K images. The frozen VLM processes each image, the guiding query, and the teacher-forced caption in a single forward pass; the caption-token log-likelihoods are averaged into a continuous reward $R_j$. Group-centered advantages then update the T2I model through AWM/GRPO.

**Training framework.** Training consists of two stages: (1) offline construction of the caption dataset using a fixed guiding query, and (2) online group-relative RL. In each RL iteration, the algorithm samples a caption $c_i$ from the dataset and uses the current T2I model to generate K images. PromptEcho computes a reward for each sampled image: the frozen VLM receives $x_j$ and the same fixed q, evaluates $c_i$ by teacher forcing, and converts its next-token likelihood into a scalar reward. AWM then uses the group-centered advantages to update the T2I model. All main experiments use AWM.

### 3.3 Data Preparation

The training data construction process is simple and efficient. We construct an internal dataset of approximately 100K concept-rich, high-quality images, and use Qwen3-VL-32B to generate detailed captions for these images, using the VLM prompt "Describe this image in detail." Qwen3-VL-32B typically generates detailed captions of 200–400 English words, covering objects, attributes, spatial relationships, colors, textures, and other multi-dimensional information in the images.

After construction, the subsequent RL training process only requires the caption text and no longer needs the original images. This property makes the data preparation cost extremely low, and the caption dataset can be conveniently expanded and updated.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

### 4.1.1 MODELS AND TRAINING CONFIGURATION

**Models and training.** We evaluate PromptEcho on two state-of-the-art open-source T2I models, Z-Image (Cai et al., 2025) and QwenImage-2512 (Wu et al., 2025a). Unless otherwise stated, reward computation uses a frozen Qwen3-VL-32B-Instruct (Bai et al., 2025) with 32 billion parameters, which is not fine-tuned throughout RL training. The main training and generation settings are summarized in Table 1.

Table 1: Training and generation configuration used in the main experiments.

| Setting | Value |
|---|---|
| LoRA | rank 64, $\alpha$=128 |
| Learning rate | $5\times10^{-5}$ |
| Sampling | 32 unique prompts $\times$ 8 images |
| Effective training batch size | 128 |
| Optimization epochs per sampled batch | 4 |
| Training hardware | 32 H20 GPUs |
| Training time | $\approx$100 h (Z-Image) / $\approx$200 h (QwenImage-2512) |
| Train / evaluation resolution | 1024$\times$1024 / 1024$\times$1024 |
| Train sampling | CFG 3.5, 30 ODE steps |
| Evaluation | CFG 4.0, 50 inference steps |

### 4.1.2 EVALUATION BENCHMARKS

**Benchmarks.** DenseAlignBench is a prompt-following stress test containing 2,000 held-out, concept-rich dense captions with an average length of approximately 200 words, covering many objects, attributes, counts, and relations. Further details on data construction and the training/benchmark split are provided in Appendix G. For each prompt, Gemini-3-flash-preview (Google DeepMind, 2025) compares baseline and post-trained images with randomized A/B order using the reasoning-oriented prompt-following rubric in Appendix B.1. To evaluate the prompt-following gains from RL training more broadly, we additionally evaluate on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and TIIFBench (Wei et al., 2025). Their prompt distributions, evaluation targets, and official evaluation procedures differ substantially from DenseAlignBench, providing complementary tests of generalization.

## 4.2 MAIN RESULTS

### 4.2.1 DENSEALIGNBENCH

To directly validate the effectiveness of PromptEcho, we use captions from DenseAlignBench (sharing the same source and distribution as training data) to generate images with both the baseline models (Z-Image and QwenImage-2512) and the PromptEcho-trained models for the same prompts. We use Gemini-3-flash-preview for pairwise evaluation; the complete evaluation protocol and prompt are provided in Appendix B.1. Results for both Z-Image and QwenImage-2512 are shown in Table 2.

Table 2: DenseAlignBench prompt-following comparison using Gemini-3-flash-preview.

| Model | Win Rate | Baseline Win Rate | Tie Rate | Net Advantage↑ |
|---|---|---|---|---|
| Z-Image: + PromptEcho vs Baseline | 61.5% | 34.7% | 3.8% | +26.8pp |
| QwenImage-2512: + PromptEcho vs Baseline | 53.3% | 37.0% | 9.7% | +16.3pp |

On Z-Image, the PromptEcho-trained model achieves an overwhelming advantage in prompt following, with a win rate of 61.5%—nearly double the baseline (34.7%)—and a net advantage of +26.8pp. On QwenImage-2512, it similarly achieves a significant win (53.3% vs 37.0%, net advantage +16.3pp), fully validating the effectiveness of PromptEcho as a reward signal and its cross-model generalization. Additional evaluator-stability, human-validation, and statistical results are reported in Appendix C.

### 4.2.2 PUBLIC BENCHMARK

The test prompts of the following public benchmarks differ significantly in distribution from our training data. We perform no task-specific training; the results entirely reflect generalized improvement in prompt following. Both baselines use officially released models at 1024×1024 with CFG 4.0 and 50 steps.

Table 3: GenEval results on structured prompt-following sub-tasks.

| Model | Single Obj. | Two Obj. | Counting | Colors | Position | Attr. Bind. | Overall |
|---|---|---|---|---|---|---|---|
| Z-Image (Baseline) | 0.99 | 0.91 | 0.75 | 0.86 | 0.41 | 0.59 | 0.75 |
| **+ PromptEcho** | 1.00 | 0.94 | **0.85** | 0.86 | **0.52** | **0.73** | **0.82** |
| QwenImage-2512 (Baseline) | 0.99 | 0.94 | 0.56 | 0.87 | 0.47 | 0.64 | 0.74 |
| **+ PromptEcho** | 0.99 | 0.93 | **0.68** | **0.90** | **0.55** | **0.70** | **0.79** |

**GenEval** As shown in Table 3, on Z-Image, PromptEcho achieves improvements across all GenEval sub-tasks, with Overall increasing from 0.75 to 0.82 (+6.5pp). The most significant improvements include: Attribute Binding +14.0pp (0.59→0.73), indicating a qualitative leap in attribute binding capability; Position +11.0pp (0.41→0.52), substantially enhanced spatial relationship understanding; and Counting +10.0pp (0.75→0.85), significantly improved counting accuracy.

On QwenImage-2512, PromptEcho improves Overall from 0.74 to 0.79 (+5pp), with Counting (+12pp, 0.56→0.68), Position (+8pp, 0.47→0.55), and Attribute Binding (+6pp, 0.64→0.70) showing the most significant gains.

**DPG-Bench** DPG-Bench evaluates T2I models' semantic alignment capability using dense prompts (long, complex descriptions), covering five categories: global, entity, attribute, relation, and other. Results are shown in Table 4.

Table 4: DPG-Bench semantic-alignment results on complex dense prompts.

| Model | Global | Entity | Attribute | Relation | Other | Overall↑ |
|---|---|---|---|---|---|---|
| Z-Image (Baseline) | 91.60 | 91.54 | 90.32 | 92.76 | 91.94 | 86.90 |
| **+ PromptEcho** | **93.05** | **92.76** | **91.87** | **93.89** | 89.99 | **87.92** |
| QwenImage-2512 (Baseline) | 94.40 | 93.27 | 90.01 | 92.82 | 91.34 | 87.32 |
| **+ PromptEcho** | 91.21 | 93.25 | **90.39** | **93.63** | **93.13** | **87.49** |

On Z-Image, PromptEcho improves the DPG-Bench Overall from 86.90 to 87.92 (+1.02), with Attribute (+1.55) and Relation (+1.13) showing the most significant gains.

**TIIFBench** TIIFBench evaluates T2I models' instruction following capability with both short- and long-description prompts. Overall results are summarized in Table 5.

Table 5: TIIFBench Overall results for short- and long-description prompts.

| Model | Overall Short | Overall Long |
|---|---|---|
| Z-Image (Baseline) | 84.91 | 83.16 |
| **+ PromptEcho** | **88.50** | **88.94** |
| QwenImage-2512 (Baseline) | 84.89 | 83.25 |
| **+ PromptEcho** | **85.50** | **86.46** |

PromptEcho improves TIIFBench Overall from 84.91/83.16 to 88.50/88.94 on Z-Image (+3.6pp/+5.8pp) and from 84.89/83.25 to 85.50/86.46 on QwenImage-2512 (+0.6pp/+3.2pp) for short/long prompts. Full fine-grained results are provided in Appendix B.3.

Across all public benchmarks, PromptEcho achieves consistent improvements on two architecturally different T2I models despite large distributional gaps with training data, demonstrating strong cross-distribution generalization stemming from the VLM's broadly pretrained image-text alignment knowledge.

#### 4.2.3 DIVERSE GUIDING-QUERY TRAINING

**Diverse guiding-query training.** The original training captions are constructed using a single guiding query ("Describe this image in detail.") and therefore follow a relatively homogeneous long and detailed descriptive style. To enrich the training-caption distribution, we construct a new caption set using diverse guiding queries that cover different semantic emphases, lengths, and structures, and use it to improve Z-Image's prompt-following capability through RL training. For each caption, the same guiding query used in caption construction is reused as the VLM query when computing its PromptEcho reward during RL. As shown in Table 6, compared with the original single-query setting, the more diverse caption distribution improves the DenseAlignBench net advantage from +26.8pp to +34.9pp and further improves results on public benchmarks including GenEval, DPG-Bench, and TIIFBench.

Table 6: Effect of diverse guiding-query training on prompt-following metrics. DenseAlignBench W/T/L denotes the post-trained model's win rate, tie rate, and baseline win rate (%).

| Setting | Dense W/T/L (net) | GenEval | DPG | TIIF-S/L |
|---|---|---|---|---|
| Z-Image baseline | — | 0.752 | 86.90 | 84.91 / 83.16 |
| Original query | 61.5%/3.8%/34.7% (+26.8pp) | 0.818 | 87.92 | 88.50 / 88.94 |
| **Diverse queries** | **64.7%/5.4%/29.9% (+34.9pp)** | **0.833** | **89.23** | **89.83 / 89.28** |

The gains across DenseAlignBench and all four public overall metrics show that diversifying the guiding queries broadens the coverage of the training captions and further improves both prompt-following performance and cross-benchmark generalization. Exact query templates, mixture proportions, and sampling procedure are provided in Appendix E.1.

### 4.3 ABLATION AND COMPARATIVE EXPERIMENTS

#### 4.3.1 COMPARISON WITH ESTABLISHED REWARDS

**Comparison with established rewards.** Under the same Z-Image RL experimental configuration, we replace PromptEcho with CLIP score, ImageReward-v1, or HPSv3 (Ma et al., 2025), perform RL post-training, and evaluate every checkpoint on the same four benchmarks. Results are summarized in Table 7. PromptEcho provides the strongest and most consistent signal, with a particularly large advantage on long, dense prompts.

We select these baselines to cover representative reward paradigms: CLIP score, like PromptEcho, requires no additional reward annotation or reward-model training and directly uses a pretrained model, but provides global image–text similarity; ImageReward-v1 and HPSv3 are learned from human-preference supervision; and InferScore in Section 4.3.2 represents training-free autoregressive VLM scoring.

Table 7: Reward comparison. DenseAlignBench W/T/L denotes the post-trained model's win rate, tie rate, and baseline win rate (%).

| Reward / model | Dense W/T/L (net) | GenEval | DPG | TIIF-S/L |
|---|---|---|---|---|
| Z-Image baseline | — | 0.752 | 86.90 | 84.9 / 83.2 |
| CLIP score | 40.1%/3.2%/56.6% (−16.6pp) | 0.771 | 86.20 | 84.8 / 83.8 |
| ImageReward-v1 | 30.6%/1.6%/67.8% (−37.1pp) | 0.774 | 87.23 | 84.3 / 84.3 |
| HPSv3 | 37.6%/3.6%/58.8% (−21.2pp) | 0.752 | 86.69 | 81.8 / 84.5 |
| **PromptEcho-32B** | **61.5%/3.8%/34.7% (+26.8pp)** | **0.818** | **87.92** | **88.5 / 88.9** |

These comparisons highlight the value of PromptEcho's full-caption, token-level signal for frontier T2I post-training, especially on dense prompts. CLIP score truncates overlength prompts to its 77-token text context and computes global embedding similarity; ImageReward-v1 truncates prompts to 35 BERT tokens; and HPSv3 targets broad human preference, with long prompts comprising only a small fraction of its training data. These input and training-distribution characteristics limit the ability of all three rewards to provide high-quality signals for dense prompts. Detailed input limits and training-text distributions are provided in Appendix D.1.

#### 4.3.2 PROMPTECHO VS. INFERSCORE

We compare two reward constructions under the same Z-Image base model, frozen Qwen3-VL-32B, training data, and RL configuration. InferScore uses a detailed prompt-following rubric to autoregressively generate a reason and a discrete 1–5 score. PromptEcho instead evaluates the fixed source prompt through teacher forcing and preserves a continuous contribution from every target token. Results are summarized in Table 8; the complete benchmark breakdown and InferScore prompt are provided in Appendix D.2.

Table 8: PromptEcho vs. InferScore. DenseAlignBench W/T/L denotes the post-trained model's win rate, tie rate, and baseline win rate (%).

| Reward | Dense W/T/L (net) | GenEval | DPG | TIIF-S/L |
|---|---|---|---|---|
| Baseline | — | 0.752 | 86.90 | 84.91 / 83.16 |
| **PromptEcho** | **61.5%/3.8%/34.7% (+26.8pp)** | **0.818** | **87.92** | **88.50 / 88.94** |
| InferScore | 28.0%/40.8%/31.3% (−3.3pp) | 0.758 | 86.86 | 82.40 / 84.64 |

PromptEcho achieves +26.8pp on DenseAlignBench and improves all four public overall metrics, whereas InferScore obtains −3.3pp and offers little or inconsistent improvement on the public benchmarks. In group-relative RL, InferScore can output only five discrete values and therefore often assigns the same score to multiple images generated from the same prompt. If all images in a group receive the same score, group centering produces zero advantages and that group provides no learning signal. PromptEcho instead computes a continuous reward from token-level conditional likelihoods, allowing it to distinguish much smaller differences among images in the same group.

#### 4.3.3 VLM CAPACITY

We further train PromptEcho with Qwen3-VL-8B and Qwen3-VL-32B under identical data and hyperparameters. Results are summarized in Table 9.

Table 9: VLM capacity ablation on DenseAlignBench and public benchmarks. DenseAlignBench W/T/L denotes the post-trained model's win rate, tie rate, and baseline win rate (%).

| Reward VLM | Dense W/T/L (net) | GenEval | DPG | TIIF-S/L |
|---|---|---|---|---|
| Z-Image baseline | — | 0.752 | 86.90 | 84.91 / 83.16 |
| **Qwen3-VL-32B** | **61.5%/3.8%/34.7% (+26.8pp)** | **0.818** | **87.92** | 88.50 / **88.94** |
| Qwen3-VL-8B | 57.6%/3.2%/39.3% (+18.3pp) | 0.771 | 87.15 | **89.08** / 87.69 |

Overall, the 32B VLM performs better across DenseAlignBench and the public benchmarks, showing that PromptEcho can directly benefit from stronger VLM visual understanding and grounding.

### 4.4 VALIDATING THE GENERAL-PURPOSE REWARD PARADIGM

To evaluate PromptEcho beyond free-form prompt following, we apply it to text-rendering optimization on an internal e-commerce poster-generation model. We replace the descriptive VLM query and caption target with a structured OCR query and a JSON target extracted from each editing instruction. The core reward computation is unchanged: the frozen VLM evaluates the target through teacher forcing, and its mean next-token log-likelihood serves as the reward. Complete label-construction and reward-query templates are provided in Appendices F.1 and F.2.

We evaluate 5,000 poster samples with Qwen3.5-Plus (Qwen Team, 2025) using strict character-level accuracy. PromptEcho improves all-text accuracy from 68% to 75% (+7pp). Adapting only the VLM query and target lets the same reward optimize a distinct objective on a different T2I model, with no task-specific reward model. The complete evaluation prompt is provided in Appendix F.3.

The same construction also transfers to reference-to-video (R2V) generation: using PromptEcho as the AWM reward, the model increasingly follows fine details of a dense, multi-shot prompt as training proceeds, such as subjects specified by reference images, a small worn accessory, the endpoint of a described motion, or an exact object count. Qualitative cases across training checkpoints are shown in Appendix F.5 (Figure F.3).

### 4.5 Qualitative Analysis

Representative cases in Appendix F.4 compare the QwenImage-2512 baseline with the PromptEcho-trained model across diverse scenarios, including character portraits, animal close-ups, still life, and macro photography. The PromptEcho-trained model shows consistently higher prompt fidelity in attribute details, object features, and pose descriptions. For text rendering, the qualitative examples show clear improvements in stroke completeness, character accuracy, and layout regularity.

## 5 Discussion & Limitations

### 5.1 Pretraining Loss vs. Autoregressive Scoring

InferScore lets the VLM explicitly score images through autoregressive generation, a process subject to hallucinations and sampling randomness. More critically, its discrete scoring mechanism prevents the VLM from distinguishing subtle differences in prompt following across multiple images generated from the same prompt with different seeds—frequently assigning identical scores to all images, effectively nullifying the reward signal. PromptEcho instead directly computes continuous log-likelihood values from the VLM's pretrained next-token distribution, requiring no autoregressive generation. The result is fully deterministic and naturally captures fine-grained distinctions—as confirmed by the large performance gap in our ablation (Section 4.3).

### 5.2 Open-Source VLMs as Free Reward Annotators

PromptEcho transforms the collective progress of the open-source VLM community into freely available reward signals for T2I optimization. Traditional reward models require dedicated annotation and specialized training, whereas PromptEcho directly leverages an existing frozen VLM. As open-source VLMs improve, PromptEcho can adopt them without changing the reward definition, recollecting preference annotations, or retraining a reward model. The 32B versus 8B ablation in Section 4.3.3 provides empirical evidence for this extensibility: the 32B VLM yields a larger DenseAlignBench advantage and leads on most public metrics.

### 5.3 Limitations

Overall, the quality of PromptEcho's reward depends on the visual understanding and grounding capabilities of the frozen VLM. In domains where the VLM lacks specialized visual expertise, such as fine-grained photographic aesthetics or recognition of subtle text errors, PromptEcho may provide less reliable reward signals. As open-source VLMs continue to improve, however, PromptEcho can directly adopt stronger frozen scorers to obtain higher-quality rewards without recollecting annotations or training a reward model.

## 6 Conclusion

We introduced PromptEcho, the first annotation-free reward that directly converts a frozen VLM's image-conditioned, teacher-forced next-token likelihood of the original prompt into a continuous token-level signal for T2I RL. Across Z-Image and QwenImage-2512, PromptEcho delivers substantial gains on DenseAlignBench and consistent improvements on GenEval, DPG-Bench, and TIIFBench without task-specific training. It outperforms established rewards and InferScore under controlled RL configurations, while the text-rendering optimization experiment demonstrates applicability beyond free-form caption following. PromptEcho requires no reward-model training, can expand through richer guiding-query data, and can directly benefit from continuing advances in open-source VLMs.

### AI use statement

Generative-AI tools were used only for language editing and presentation polishing. The research question, method, implementation, experiments, analysis, and conclusions were developed and verified by the authors, who take full responsibility for the manuscript.

### Reproducibility statement

The released materials include the PromptEcho implementation, training and evaluation configurations, trained checkpoints, DenseAlignBench, evaluation prompts, and scripts required to reproduce the reported comparisons. The appendix documents the exact reward computation, generation settings, evaluator prompts, and additional statistical analyses.

## References


Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023.

Google DeepMind. Gemini 3 flash preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview`, 2025.

Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-RPO: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

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

Runhui Huang, Qihui Zhang, Zhe Liu, Yu Gao, Jie Wu, and Hengshuang Zhao. Read it back: Pretrained MLLMs are zero-shot reward models for text-to-image generation. *arXiv preprint arXiv:2607.11886*, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Qwen Team. Qwen3.5-397b-a17b. `https://huggingface.co/Qwen/Qwen3.5-397B-A17B`, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. TIIF-Bench: How does your T2I model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

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*, 2025a.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. RewardDance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, et al. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning RL with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025.

# A Implementation Details and Efficiency

## A.1 PromptEcho scorer

For each image–prompt pair, we form a multimodal chat sequence containing the image and guiding query as the user input and the original prompt as the assistant target. The frozen VLM performs one teacher-forcing forward pass. We average next-token log-probabilities only over the target-caption span; query, chat-template, padding, and assistant-end tokens are excluded. The resulting mean log-probability is the PromptEcho reward.

$$R_\phi(x, c, q) = -\mathcal{L}_{\mathrm{NTP}}(c \mid x, q) = \frac{1}{T}\sum_{t=1}^{T} \log p_\phi(c_t \mid x, q, c_{<t})$$

## A.2 Rollout and minibatch organization

Each iteration produces 256 rollout samples, which are split into two minibatches of 128 samples for optimization. Each sampled batch is reused for four optimization epochs.

### A.3 RUNTIME BREAKDOWN

Z-Image and QwenImage-2512 training take approximately 100 and 200 hours, respectively, on 32 H20 GPUs. Online image generation accounts for approximately 80% of total training time, followed by diffusion-model updates, while PromptEcho VLM reward computation accounts for less than 5%. The frozen reward service can also be shared across training jobs.

## B EVALUATION PROTOCOLS

### B.1 DENSEALIGNBENCH PAIRWISE EVALUATION

For every prompt, the baseline and post-trained images are presented to Gemini-3-flash-preview as Image A and Image B. Their order is swapped with 50% probability. The evaluator is instructed to judge prompt following only, explicitly checking subjects, actions, attributes, composition, text, spatial relations, and counts while ignoring aesthetic preference. The output is parsed from a strict JSON schema with three possible decisions: `image_a`, `image_b`, or `tie`.

**DenseAlignBench complete evaluation prompt**

```
You are a professional image quality assessment expert specializing in
evaluating prompt following accuracy.

Your task is to compare two AI-generated images and determine which
image better follows the given prompt.
Focus only on prompt following accuracy - do not consider aesthetics,
artistic quality, or realism.

**Prompt:**
{prompt}

**Evaluation Process:**
1. Carefully read and understand all requirements in the prompt:
   - Main subjects and objects
   - Actions and poses
   - Visual attributes (color, size, material, texture)
   - Composition and layout
   - Style and atmosphere
   - Any text or written elements
   - Spatial relationships (foreground, background, position)
   - Quantities and counting

2. Examine Image A:
   - Which prompt requirements are accurately depicted?
   - Which prompt requirements are missing or incorrect?
   - Are there elements not mentioned in the prompt?

3. Examine Image B:
   - Which prompt requirements are accurately depicted?
   - Which prompt requirements are missing or incorrect?
   - Are there elements not mentioned in the prompt?

4. Compare:
   - Which image more accurately captures the prompt requirements?
   - If both are equal in accuracy, select "tie"

**Preference Options:**
- "image_a": Image A better follows the prompt than Image B
- "image_b": Image B better follows the prompt than Image A
- "tie": Both images follow the prompt to a similar degree

**Important Notes:**
- Make your decision based solely on prompt following accuracy
- Ignore differences in artistic style, aesthetics, or realism
- Remain objective and unbiased
- Select "tie" when prompt following quality is truly similar

**Output Format (strict JSON, no markdown):**
{
  "reasoning": "<detailed explanation>",
  "preference": "<image_a, image_b, or tie>"
}
```

### B.2 PUBLIC BENCHMARKS

GenEval and DPG-Bench are evaluated with their official prompts, protocols, and evaluation code. Baseline and post-trained checkpoints use identical generation settings: $1024\times1024$ resolution, CFG

4.0, and 50 inference steps; the checkpoint is the only changed component. TIIFBench likewise follows its official short- and long-prompt evaluation protocol.

### B.3 TIIFBench fine-grained results

Tables B.1 and B.2 report the complete TIIFBench breakdown. Each slash-separated cell contains the short/long prompt score.

Table B.1: TIIFBench Overall and basic-metric results. Slash-separated values denote short/long prompts.

| Model | Overall S/L | Basic Attr. S/L | Relations S/L | Reasoning S/L | Basic Avg S/L |
|---|---|---|---|---|---|
| Z-Image (Baseline) | 84.91/83.16 | 89.5/88.9 | 90.3/82.3 | 79.3/86.0 | 86.4/85.7 |
| **+ PromptEcho** | **88.50/88.94** | 94.5/93.0 | 90.9/86.5 | 85.5/89.6 | 90.3/89.7 |
| QwenImage-2512 (Baseline) | 84.89/83.25 | 92.2/86.5 | 84.9/86.8 | 79.4/81.0 | 85.5/84.8 |
| **+ PromptEcho** | **85.50/86.46** | 91.0/87.0 | 84.7/90.0 | 85.5/84.6 | 87.1/87.2 |

Table B.2: TIIFBench advanced, style, text, and real-world results. Slash-separated values denote short/long prompts.

| Model | Attr.+Rela. S/L | Attr.+Reas. S/L | Rela.+Reas. S/L | Advanced Avg S/L | Style S/L | Text S/L | Real World S/L |
|---|---|---|---|---|---|---|---|
| Z-Image (Baseline) | 80.8/81.3 | 75.8/77.4 | 80.9/79.3 | 79.9/79.5 | 82.8/73.3 | 93.2/89.6 | 91.6/90.3 |
| **+ PromptEcho** | 83.5/88.5 | 80.1/88.4 | 84.4/82.8 | 83.3/86.8 | 83.3/83.3 | 99.1/95.0 | 95.2/93.3 |
| QwenImage-2512 (Baseline) | 84.5/84.8 | 73.1/81.7 | 81.6/81.6 | 80.3/82.2 | 80.0/60.0 | 95.5/93.2 | 92.9/93.7 |
| **+ PromptEcho** | 79.1/85.4 | 80.2/88.6 | 81.3/83.1 | 80.6/85.4 | 72.4/69.0 | 99.1/95.4 | 96.2/95.0 |

## C Human Validation and Statistical Reliability

### C.1 Human validation

We randomly sample 200 comparisons from the Z-Image DenseAlignBench experiment, of which 197 receive valid human judgments. Human evaluation yields 56.3% PromptEcho wins, 23.9% ties, and 19.8% baseline wins, corresponding to a +36.5pp net advantage. Gemini produces 66.0% wins, 4.1% ties, and 29.9% baseline wins on the same subset, corresponding to +36.1pp.

Table C.1: Human and Gemini evaluation results on the same 197 valid DenseAlignBench comparisons.

| Evaluator | PromptEcho win | Tie | Baseline win | Net |
|---|---|---|---|---|
| Human | 56.3% | 23.9% | 19.8% | **+36.5pp** |
| Gemini, same subset | 66.0% | 4.1% | 29.9% | **+36.1pp** |

Human and Gemini judgments agree on 77.7% of samples (153/197), with human-versus-Gemini Cohen's $\kappa$=0.601. Among 44 disagreements, 39 involve one evaluator selecting a tie and the other selecting a winner; only five reverse the winning model. Although human evaluation produces more ties, both evaluators reach nearly identical model-level net advantages.

### C.2 Repeated Gemini evaluation

Table C.2: Three independent DenseAlignBench evaluations for each base model.

| Base model | Run | PromptEcho win | Tie | Baseline win | Net |
|---|---|---|---|---|---|
| Z-Image | 1 | 61.5% | 3.8% | 34.7% | +26.8pp |
| Z-Image | 2 | 62.8% | 4.2% | 33.0% | +29.8pp |
| Z-Image | 3 | 63.0% | 4.6% | 32.4% | +30.6pp |
| QwenImage-2512 | 1 | 53.3% | 9.7% | 37.0% | +16.3pp |
| QwenImage-2512 | 2 | 53.2% | 9.5% | 37.2% | +16.0pp |
| QwenImage-2512 | 3 | 53.9% | 9.6% | 36.5% | +17.4pp |

Across the three runs, the net-advantage range is 3.8pp for Z-Image and 1.4pp for QwenImage-2512. Mean pairwise agreement is 79.7% and 73.5%, with Fleiss' $\kappa$ of 0.592 and 0.533, respectively.

### C.3 Bootstrap confidence intervals

A 10,000-sample prompt-level bootstrap gives a 95% confidence interval of [+22.7pp, +31.1pp] for the Z-Image net advantage and [+12.1pp, +20.3pp] for QwenImage-2512. The corresponding PromptEcho-win-rate intervals are [59.4%, 63.7%] and [51.1%, 55.5%].

## D Reward Baselines and InferScore Details

### D.1 Established reward baselines

CLIP score represents an image and its text with global embeddings and accepts at most 77 text tokens, limiting its coverage of the many objects, attributes, counts, and relations in DenseAlignBench captions. ImageReward-v1 is trained as a general human-preference model and truncates its text input to 35 BERT tokens in the official training and inference code. Across the 8,000 public training prompts, the median length is approximately 17 words, the 95th percentile 50 words, and the maximum 137 words; about 35.5% exceed the 35-token input limit.

HPSv3 (Ma et al., 2025) does not use the same 35-token truncation and HPDv3 includes some VLM-generated long captions, but its objective combines aesthetics, visual quality, semantic similarity, and overall coherence rather than specializing in dense prompt following. In the public HPDv3 `train.json`, containing 1,139,924 preference pairs, the pair-weighted prompt length has a median of approximately 24 words and a 95th percentile of 166 words; only about 0.59% reach 200 words. Among 202,274 unique prompts, approximately 0.25% reach 200 words.

### D.2 InferScore

InferScore uses the same frozen Qwen3-VL-32B and the same image–prompt input as PromptEcho, but autoregressively generates a textual reason and a discrete 1–5 prompt-following score. Tables D.1 and D.2 provide the complete benchmark breakdown, followed by the exact prompt used in our experiments.

Table D.1: GenEval breakdown for PromptEcho and InferScore under matched Z-Image RL settings.

| Reward | Single Obj. | Two Obj. | Counting | Colors | Position | Attr. Bind. | Overall |
|---|---|---|---|---|---|---|---|
| Baseline | 0.99 | 0.91 | 0.75 | 0.86 | 0.41 | 0.59 | 0.75 |
| **PromptEcho** | 1.00 | 0.94 | 0.85 | 0.86 | 0.52 | 0.73 | **0.82** |
| InferScore | 1.00 | 0.91 | 0.79 | 0.86 | 0.41 | 0.59 | 0.76 |

Table D.2: DPG-Bench breakdown and TIIFBench Overall results for PromptEcho and InferScore under matched Z-Image RL settings.

| Reward | DPG Global | Entity | Attribute | Relation | Other | DPG Overall | TIIF-S | TIIF-L |
|---|---|---|---|---|---|---|---|---|
| Baseline | 91.60 | 91.54 | 90.32 | 92.76 | 91.94 | 86.90 | 84.91 | 83.16 |
| **PromptEcho** | 93.05 | 92.76 | 91.87 | 93.89 | 89.99 | **87.92** | **88.50** | **88.94** |
| InferScore | 78.74 | 93.17 | 90.21 | 92.56 | 82.49 | 86.86 | 82.40 | 84.64 |

**Complete InferScore prompt**

```
You are a professional image quality evaluator. Please objectively evaluate an AI-generated
     image based on the provided description.

Description/Caption:
{prompt}

Evaluation Dimension (1--5 scale):

Prompt Following

5: Accurately captures and represents all key elements described, including subjects, actions,
      attributes, style, atmosphere, and composition.
4: Captures nearly all key elements, with only minor deviations or a few small missing
     details.
3: Captures the main elements but has noticeable deviations, missing details, or slight
     misinterpretations.
2: Captures only some relevant elements, with significant omissions or misinterpretations.
1: Fails to represent the core content, has major contradictions, or is largely unrelated to
     the description.

Evaluation Guidelines:

Compare the image carefully against the provided description.
Evaluate objectively based only on visible image content.
Do not penalize the image merely because the description is complex.
Briefly explain which requested elements are present, missing, or incorrectly represented.
Ensure the reason is specific and consistent with the assigned score.

Output Format (strict JSON, no markdown):
{"reason": "...", "prompt_following": <1, 2, 3, 4, or 5>}
```

# E Guiding-Query and Token-Level Analyses

## E.1 Diverse guiding-query construction

The diverse-caption experiment uses a fixed Hybrid50 dataset containing 50% captions from the original single-query dataset and 50% newly generated captions. Within the diverse half, guiding-query families are sampled with the weights below. Assignment is deterministic by sample identity, and no benchmark prompt is copied into the training set.

Table E.1: Guiding-query families and mixture proportions in the Hybrid50 dataset.

| Query family | Target length | Share in diverse half | Share in Hybrid50 |
|---|---|---|---|
| Short object/count prompt | 5–15 words | 25% | 12.5% |
| Short compositional prompt | 12–35 words | 25% | 12.5% |
| Long factual prompt | 50–100 words | 20% | 10% |
| Long compositional prompt | 60–110 words | 15% | 7.5% |
| Style/design prompt | 25–80 words | 10% | 5% |
| Visible-text prompt | 15–80 words | 5% | 2.5% |
| Original detailed caption | typically 200–400 words | — | 50% |

**Complete diverse guiding queries**

```
[Short object/count]
Write a minimal English image-generation prompt of 5--15 words. Mention only the main one or
     two visible object categories. If clearly visible, include exact count, color, and one
     simple left/right/above/below relation. Do not add background, mood, style, or unseen
     details. Output the prompt only.

[Short compositional]
Write one concise English image-generation sentence of 12--35 words. Preserve the most
     important visible entities, exact counts, actions, colors, textures, shapes, comparisons,
      and spatial relations. Prefer at least two constraints when supported by the image. Do
     not invent absent or negative facts. Output one sentence only.

[Long factual]
Write a factual English image-generation prompt in 3--5 sentences and 50--100 words. Describe
      the global scene, visible entities, counts, colors, shapes, textures, states, and
     spatial or non-spatial relations. Do not add atmosphere, emotions, sounds, time, or
     unseen details. Output the prompt only.

[Long compositional]
Write a context-rich English image-generation prompt of 60--110 words. Preserve all visible
     subjects, actions, counts, attributes, comparisons, and 2D or 3D relations. Include
     style and composition only when visually evident. Do not invent text, absent objects, or
      hidden facts. Output the prompt only.

[Style/design]
Write a practical English image-generation prompt of 25--80 words. Describe the visible
     subject, artistic medium or visual style, lighting, camera viewpoint, composition, and
     design intent when they are supported by the image. Do not invent branding or text.
     Output the prompt only.

[Visible text]
Write an English image-generation prompt of 15--80 words describing the scene and
     transcribing every clearly readable visible word exactly in quotation marks. Preserve
     capitalization when possible. Do not guess unreadable text or invent additional wording.
      Output the prompt only.
```

### E.2 CONTROLLED COUNT-AND-COLOR ANALYSIS

We construct nine short captions of the form "The image shows {count} {color} apples." using three counts and three colors. Z-Image generates one $1024 \times 1024$ image for each caption with seed 42, 30 inference steps, and guidance scale 4.0; the random generator is reinitialized before every sample so that all nine images start from the same noise. PromptEcho then scores every image against every caption with the fixed query "Describe this image in one sentence briefly.", producing 81 image–caption combinations.




Figure E.1: Controlled images spanning three counts and three colors.

Table E.2: Sequence-level next-token prediction loss grouped by count and color match status.

| **Image–caption relation** | **Pairs** | **Mean loss** | **Increase over exact match** |
|---|---|---|---|
| Count and color match | 9 | 2.0150 | — |
| Count mismatch only | 18 | 2.8154 | +0.8004 |
| Color mismatch only | 18 | 3.0929 | +1.0780 |
| Count and color mismatch | 36 | 3.7585 | +1.7435 |

For every image, the exactly matched caption has the lowest loss. As shown in the figure, mismatches in either color or count lead to a substantial increase in loss.

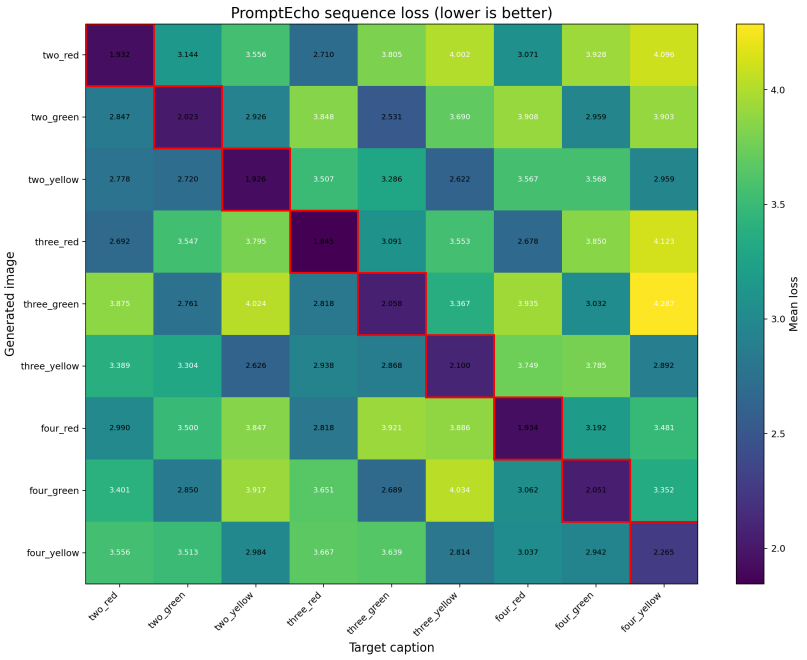



Figure E.2: Sequence-level next-token prediction loss for all 81 combinations (lower is better). Rows are images, columns are candidate captions, and the red diagonal marks exact matches.

Table E.3: Token-level negative log-likelihood for matched and mismatched count and color attributes.

| **Token type** | **Attribute relation** | **Tokens** | **Mean token loss** | **Mismatch − match** |
|---|---|---|---|---|
| Count | Match | 27 | 0.0543 | — |
| Count | Mismatch | 54 | 6.6212 | **+6.5668** |
| Color | Match | 27 | 1.1402 | — |
| Color | Mismatch | 54 | 7.3333 | **+6.1931** |

Count and color mismatches therefore raise the negative log-likelihood primarily at the corresponding semantic tokens, rather than uniformly across the shared sentence template.

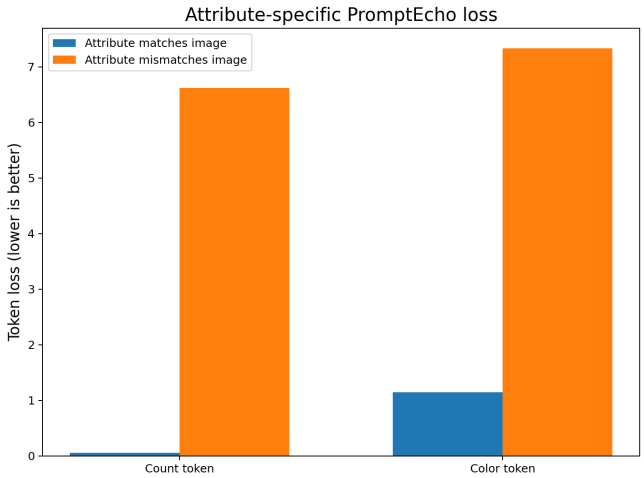



Figure E.3: Negative log-likelihood at count and color tokens under matched and mismatched visual semantics.

# F TEXT RENDERING AND QUALITATIVE RESULTS

## F.1 LABEL-CONSTRUCTION PROMPT

The following prompt extracts the structured target label from each poster-editing instruction.

**Complete label-construction prompt**

```
You are a professional text structure analysis expert. Your task is to
extract all text content required to appear in the image from a given
prompt, and classify them by semantic role in a structured manner.

**Given Prompt:**
{prompt}

**Extraction Rules:**
1. **main_title**: If the title has multiple lines (e.g., top line/
   bottom line), each line is a separate element in the list.
2. **subtitle**: Text inferred to be subtitles, such as product
   titles or secondary headings.
3. **selling_points**: Product selling points, recommendations,
   or introductory phrases.
4. **other_text**: Other text not belonging to the above categories.

**Notes:**
- Each piece of text should be extracted completely, preserving
  the original content
- If a category has no corresponding content, return an empty list []
- Carefully read the context to determine the semantic role of
  each piece of text

**Output Format (must strictly follow JSON format, do not add any
markdown tags):**
{
  "reasoning": "<brief explanation of how you identified and classified these texts>",
  "main_title": ["first line title", "second line title"],
  "subtitle": ["subtitle text"],
  "selling_points": ["point 1", "point 2"],
  "other_text": ["other text 1", "other text 2"]
}
```

## F.2 TRAINING REWARD QUERY

Given a generated poster, this structured OCR query is provided to the frozen VLM, while the structured label from Section F.1 is teacher-forced as the target.

**Complete OCR reward query**

```
You are a professional image text recognition expert. Your task is to
recognize all designed and typeset text content in this e-commerce
marketing image, and classify them by semantic role in a structured
manner.

**Extraction Rules:**
1. **main_title**: The largest and most prominent title text in the
   image. If the title has multiple lines, each line is a separate
   element in the list.
2. **subtitle**: Subtitles or supplementary text, such as product
   titles or secondary headings.
3. **selling_points**: Product selling points, recommendations,
   introductory phrases, efficacy features, etc.
4. **other_text**: Other text not belonging to the above categories.

**Notes:**
- Only extract designed marketing text in the image; do not extract
  text printed on product packaging or physical objects
- Each piece of text should be extracted completely, preserving
  the original content
- Each field is a list of strings, arranged in visual reading order
- If a category has no corresponding content, return an empty list []

**Output Format (must strictly follow JSON format, do not add any
markdown tags or explanations):**
{"main_title": ["first line title", "second line title"],
 "subtitle": ["subtitle text"],
 "selling_points": ["point 1", "point 2"],
 "other_text": ["other text 1", "other text 2"]}
```

### F.3 TEXT-RENDERING EVALUATION PROMPT

Qwen3.5-Plus, selected for its strong Chinese-character recognition capability, performs strict character-level comparison and returns a binary score. An image receives 1 only when all required text is rendered correctly.

**Complete evaluation prompt**

```
You are a professional and strict text rendering quality inspector.
You are extremely sensitive to text details - any character error,
garbled text, missing character, extra character, or positional
deviation is unacceptable.
Your task is to check whether the text rendered in the image exactly
matches the text required by the prompt.
**Prompt:**
{prompt}
**Inspection Steps:**
1. **Extract required text**: Carefully read the prompt and extract
   all text content required to appear in the image. Pay attention
   to distinguishing Chinese/English, uppercase/lowercase, numbers,
   etc. If the prompt does not explicitly require any text to appear
   in the image, expected_text is an empty string, and directly
   give 1 point.
2. **Recognize image text**: Carefully examine the image and
   recognize all text appearing in it character by character.
   Ignore text on product packaging; focus on the text required
   by the prompt.
3. **Character-by-character comparison**: Perform strict
   character-by-character comparison between the prompt-required
   text and the actual text in the image:
   - Every required character must be completely correct;
     garbled text is unacceptable
   - No extra characters, missing characters, wrong characters,
     or garbled text allowed
   - No duplicate text (unless the prompt explicitly requires it)
   - Text order must be correct
4. **Scoring**:
   - 1 point: All required text is rendered completely correctly
     in the image with no character-level errors
   - 0 points: Any text error exists
**Output Format (strict JSON):**
{
  "expected_text": "<all required text extracted from the prompt>",
  "found_text": "<all text actually recognized from the image>",
  "reasoning": "<detailed character-by-character comparison results>",
  "score": <integer 0 or 1>
}
```

## F.4 QUALITATIVE CASES

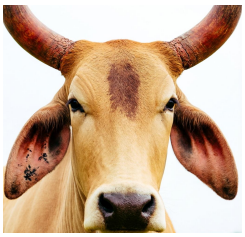
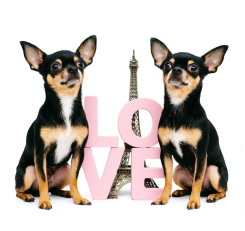
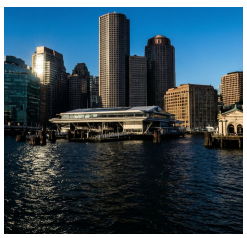
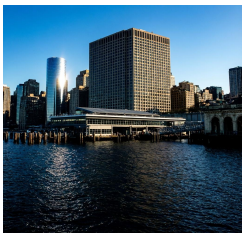
| Prompt | Baseline | Ours |
|---|---|---|
| This is a close-up, eye-level photograph of a light brown or tan-colored cow, likely a breed such as a Zebu or Brahman, given its prominent features. The cow is facing directly toward the camera, creating an intimate and engaging perspective. Head and Face: - The cow has large, curved horns that sweep upward and outward. The horns are thick and appear weathered, with a reddish-brown, almost rust-like coloration, especially near the base where they emerge from the skull. There are visible cracks and rough textures on the horns, suggesting age or exposure to the elements. - Its face is broad and light tan, with a **slightly darker brown patch running down the center of the forehead** and around the eyes. - The eyes are dark, expressive, and appear calm, with a gentle gaze. The area around the eyes is shadowed, giving depth to the face. - The ears are large, floppy, and hang down on either side of the head. The inner parts of the ears are pinkish-brown, and there are some **dark, irregular patches or marks on the left ear (viewer's left)**, possibly scars or natural pigmentation. - The muzzle is broad and light-colored, with a dark, moist nose. |  |  |
| This is a vibrant, close-up photograph of a young girl blowing bubbles outdoors on a sunny day. The image captures a moment of joyful, carefree play. Subject: The girl is the central focus, shown from the chest up. She has long, light brown hair tied into a high ponytail with a light purple or gray hair tie. A few loose strands frame her face. Her eyes are gently closed or looking upward as she blows, and her lips are puckered in concentration. She appears to be around 8–12 years old. Clothing and Accessories: She is wearing a sleeveless top with a black-and-white striped or patterned design, possibly with ruched fabric. Around her neck, she wears a layered necklace: one is a beaded strand with dark blue or black and white beads, accented with **two large, smooth, reddish-brown wooden beads** and a circular metal pendant. Underneath, a thinner, delicate chain is visible. She also wears a small, shiny stud earring in her visible left ear. Action: She is holding a purple bubble wand in her right hand, which is raised near her mouth. The wand has a **small, clear bubble forming at the end**. |  |  |
| This image features two identical-looking Chihuahua dogs sitting side by side against a clean, white background. The dogs are small, with short, smooth coats that are predominantly black, accented with tan markings on their faces, legs, and paws, and a white patch on their chests. Their large, upright ears are pointed and alert, and their dark, expressive eyes gaze upward, giving them an endearing and curious look. Between the two dogs, centered in the composition, is the word "LOVE" spelled out in large, three-dimensional, pastel pink letters. The letters have a smooth, matte finish and are arranged in a slightly staggered, playful manner. **Behind the "LOVE" letters, partially visible through the open spaces** in the letters (especially the "O" and "V"), is a small, detailed model of the Eiffel Tower. The tower is rendered in a metallic bronze or gold color, adding a touch of elegance and a romantic, Parisian theme to the image. The overall composition is symmetrical and balanced, with the two dogs mirroring each other on either side of the central "LOVE" and Eiffel Tower elements. The clean white background emphasizes the subjects,… |  |  |
| This is a striking urban landscape photograph taken from across a body of water, likely a river or harbor, looking toward a dense city skyline under a clear, deep blue sky. The lighting suggests either early morning or late afternoon, with the sun low in the sky, creating strong contrasts and dramatic shadows. **Foreground:** The foreground is dominated by dark, rippling water that reflects the sunlight in shimmering patches. Along the water's edge, there are wooden pilings and docks, indicating a working waterfront. A large, modern-looking structure with a flat roof and glass panels sits on the water's edge. It appears to be a ferry terminal or a public transit hub, with visible openings and what might be ferry slips or loading areas. To its right, there's an older, more ornate structure with arched openings and classical detailing, possibly a historic pier or part of a bridge approach. **Midground:** The midground features a mix of architectural styles. The most prominent building is a massive, rectangular skyscraper with a grid-like facade of windows, standing tall and imposing. To its left, **a sleek, cylindrical skyscraper with a reflective glass surface catches the sunlight**, creating a bright glare. Other buildings of varying heights |  |  |
| This image presents a festive, top-down view of a Christmas gift wrapped in vibrant red and cream holiday-themed paper. The wrapping paper features a repeating pattern of cheerful, cartoonish Santa Claus figures with white beards and red hats, set against a cream background with subtle snowflake and swirl designs. The red sections of the paper are adorned with small decorative elements like stars and hearts. The gift is tied with a glossy red satin ribbon, which is tied into a neat bow on the left side. Attached to the bow is a shiny, golden Christmas bauble, adding a touch of sparkle and holiday cheer. Resting atop the gift are two modern payment-related items: 1. **A white iZettle card reader**: This compact, rectangular device has a black screen displaying a contactless payment symbol **(a stylized wave with a hand above it).** Below the screen is a numeric keypad with black circular buttons labeled 1, 2, 3, 4, 5, and 6 (partially visible). The brand name "iZettle" is printed in blue near the bottom. 2. **A gold-colored credit or debit card**: **Positioned diagonally across the card reader**, this card has a smooth, metallic gold finish. It features a standard magnetic stripe along… |  |  |
| This is a well-composed, professional-style portrait of a man standing outdoors in what appears to be an autumn setting. The image is shot with a shallow depth of field, keeping the man in sharp focus while the background is softly blurred, emphasizing him as the central subject. **Subject:** The man is in his late 20s to early 30s, with dark, neatly styled hair, a light stubble beard, and intense, direct gaze toward the camera. His expression is serious and confident, with **slightly parted lips**, giving him a thoughtful or pensive look. **Attire:** He is dressed in stylish, autumn-appropriate clothing: - A **camel-colored wool coat** with a classic, tailored cut, featuring a single visible button and a notched lapel. - Underneath, he wears a **deep burgundy or maroon sweater**. - Around his neck is a **plaid scarf** with a pattern of navy blue, beige, and white or light gray, adding texture and warmth to his outfit. **Accessories:** He holds a **large, black umbrella** with a curved black handle. The umbrella is open and positioned above his head, suggesting it might be raining or about to rain, or simply used for stylistic effect. The umbrella's canopy is fully extended, creating a strong… |  |  |

Figure F.1: Qualitative comparison of six representative winning cases. Each row presents the complete prompt, the QwenImage-2512 baseline, and the PromptEcho-trained result.

| Input Image | Prompt | Before PromptEcho | After PromptEcho |
|---|---|---|---|
|  | A 9:16 vertical e-commerce poster. Extract the Taier Sauerkraut Fish bowl and the white rice bowl from input image 1 and position them at a slight angle on a light grey sandstone countertop. Intense, warm spotlighting from the left creates high contrast and brings out every detail of the red peppers and garlic bits. **Text Styling:** The main title "太二酸菜鱼" is placed at the top-left in a bold, stylized black font with a silver metallic sheen and hard drop shadow. Directly beneath, "**闲鱼价85元起**" is the focal point, where "85元" is rendered in a giant, bright orange-yellow 3D comic font with sharp yellow speed lines bursting outwards from the center. The character "起" is scaled down and colored dark charcoal. The subtitle "吃喝玩乐先上闲鱼" is stylized as a neon-style graphic at the bottom-right with a soft yellow glow and a thin white stroke. **Decorative Elements:** Droplets of golden soup and a few floating peppercorns are frozen in mid-air around the bowl to suggest fresh preparation. A folded linen napkin is draped partially under the fish bowl. The background is a blurred, light-colored textured plaster wall. Overall style is dynamic, modern, and mouth-watering with clear focus and rich natural shadows. |  |  |
|  | A 9:16 vertical e-commerce poster. Extract the Iced Butter Latte from input image 1 and position it in the bottom third, creating a powerful 3D breakthrough effect where the cup emerges from a circular hole in a yellow-tinted, stylized newspaper background. The newspaper is oriented diagonally, filled with small decorative text blocks describing butter coffee benefits. Lighting is vibrant and high-saturation, coming from multiple directions. **Main Headline Styling**: '淘宝闪购 新人低至0.01元' is rendered in a playful, blocky 3D font at the top; '0.01元' is exceptionally large, colored in bright orange with a thick black comic-style outline and white highlight glints. **Sub-headline Styling**: '最高可得19元外卖红包' is written inside a red slanted badge with a white zigzag border. **Creative Bar**: The text '**淘宝闪购，下单即达不等待**' is styled in a blue neon-tube effect script font centered on a semi-transparent frosted glass bar. **Decorations**: Action lines extend from the product to emphasize movement, accompanied by floating cinnamon sticks and scattered coffee grounds. Overall style is a youthful, energetic Pop-Art inspired commercial poster with high spatial depth. |  |  |
|  | A 9:16 vertical e-commerce poster. Extract the dual-layer and single-layer egg steamers from input image 1. Position the single-layer steamer in the lower center foreground, appearing to physically overlap the edge of a large glowing pink neon circular border, with the double-layer version behind it to create depth. Background & Scene: A dreamy, soft pastel sky blue background decorated with layered paper-cut clouds and light blue pig silhouette patterns. Lighting: Dreamy top-down lighting creating a soft-focus glow around the product domes. Text Styling: The main heading "萌趣早餐·快乐开餐" is at the top in a deep pink brush script with a thick white border. The promotional text "**新人全场**0.01元起" is housed in a translucent frosted glass text box at the bottom; "0.01元" is oversized with a marshmallow-like texture and a soft rainbow gradient. A sub-caption "早起也有好胃口" appears in a white pill-shaped container. Decorative Elements: Voluminous 3D cotton-candy-like steam clouds and a trail of tiny 3D pig footprints leading from the foreground into the background. Overall Style: A soft-focus, whimsical 3D animation style with an extremely high level of cuteness and warmth. |  |  |
|  | A 9:16 vertical e-commerce poster. Extract the Grandpa's Tea (爷爷不泡茶) specialty drinks and the orange gift box from input image 1 and composite them at the center-bottom on a vibrant, matte pink circular podium. The background is a high-end designer scene with a soft orange and pink pastel gradient and floating 3D frosted glass spheres. **Lighting** is a crisp studio setup with white rim light emphasizing the curves of the bottles and cups, creating a professional commercial feel. **Text Styling**: The main title "上淘宝闪购**搜樱**花季" (top line) and "抢樱花季限定盲盒套餐" (bottom line) is rendered in a glossy 3D puffy font. The characters have a milky white center, surrounded by a luminous yellow and orange-red multi-layered outline, creating a cohesive, eye-catching block. The subtitle "上新了 橙意好礼" sits on a small orange ribbon badge below the title. **Decorative Elements**: The product is surrounded by small, artistic floating text saying "**爷爷不泡茶**" in different styles like serif and bold sans-serif. Pink silk ribbons and stylized orange fruit segments float in the mid-ground to add dynamic energy. The overall style is a modern, high-quality 3D digital illustration with a joyful and energetic spring mood. |  |  |

Figure F.2: Text-rendering comparison before and after PromptEcho training. Each row presents the input image, editing prompt, result before PromptEcho, and result after PromptEcho.

### F.5 REFERENCE-TO-VIDEO PROMPT FOLLOWING OVER TRAINING

To probe whether the PromptEcho reward generalizes beyond image generation, we apply it as the reward in a reference-to-video (R2V) setting, where a model renders subjects specified by reference images while following a dense, multi-shot Chinese prompt. We track three cases across seven AWM checkpoints, sampling one frame per checkpoint at a fixed per-case timestamp with the targeted prompt phrase highlighted. As training proceeds, fine details that the base model initially ignores are increasingly realized: a black digital watch is revealed (Sample 1), the subject settles at the viewing-platform edge (Sample 2), and the drawer cabinet converges to the specified five tiers (Sample 3).




Figure F.3: Reference-to-video prompt following across AWM training steps. For three cases, one frame per checkpoint is sampled at a fixed timestamp, and the targeted prompt phrase is highlighted. As training proceeds the phrase is progressively realized: a black digital watch appears (Sample 1), the subject reaches the viewing-platform edge (Sample 2), and the drawer cabinet converges to the specified five tiers, from seven to five (Sample 3). The prompt text shown in the figure is an excerpt (bracketed by ellipses); the complete prompt for each case is given in the text below, and the input reference images are not shown here.

**Sample 1, shot 2 (00:02.0-00:03.5).** [Direct cut][Close-up][Fixed camera] A close view of the grass, where every blade is clearly visible and a few withered leaves are scattered on the ground; the background is the natural gap where the plane-tree roots meet the lawn. The man (@ref1) sits cross-legged on the grass in a waist-level close-up. His red polo shirt (@ref1)(@ref2)(@ref3) forms natural horizontal creases at the belly and hem from the seated pose, and the fabric of his khaki shorts (@ref1) bunches as his legs bend. One hand rests on his knee, revealing the detail of the **black digital watch** (@ref1).

**Sample 2, shot 3 (00:03.5-00:05.0).** [Direct cut][Medium shot][Push-in] A rustic stone arch and an open-air cafe. The arch is built of weathered beige stone, with a few strands of ivy hanging above it. Behind it is an open terrace set with round tables under white cloths and woven rattan chairs. The terrace edge has a black cast-iron railing, beyond which lies a shimmering blue expanse of water dotted with a few small white boats. The woman (@ref1), wearing a straw hat (@ref1) and pearl earrings (@ref1), a white shirt (@ref1)(@ref2)(@ref3) and blue jeans (@ref1)(@ref2)(@ref3), first leans quietly by the stone arch in profile, then walks lightly toward the cafe, and finally **stops at the edge of the viewing platform**, back to the camera, her long hair (@ref1) fluttering in the wind as she tilts her head up to gaze at the distant water.

**Sample 3, shot 3 (00:03.3-00:05.0).** [Direct cut][Wide shot][Slow pull-back] A spacious indoor space with a continuous natural-wood plank floor. Against one wall stands a **five-tier** solid-wood drawer cabinet in dark wood with round brass pulls, a white ceramic vase placed on top. On the other side is a large sliding glass door with a black aluminum frame leading outdoors. The woman (@ref1)(@ref2), in a beige short-sleeve top (@ref1)(@ref2) and a plaid pleated skirt (@ref1)(@ref2), white mid-calf socks (@ref1) and black flats (@ref1)(@ref2), white sunglasses (@ref1) on her head and a black backpack (@ref2), stands before the cabinet at the center of the room holding a bouquet of white flowers. She first looks down at the flowers with a gentle expression, then raises her head to look straight at the camera, relaxed and naturally poised, with a soft smile.

# G Data Construction and Separation

## G.1 Original caption data and DenseAlignBench split

The source images are collected primarily from the Internet, filtered for quality with Gemini-3-flash, and filtered to remove low-resolution samples; retained images are generally at least approximately $1024 \times 1024$. Qwen3-VL-32B generates detailed English captions using "Describe this image in detail.", typically producing descriptions of approximately 200–400 words that cover objects, counts, attributes, colors, materials, and spatial relations.

After removing exactly duplicated images, 2,000 samples are randomly held out to form DenseAlign-Bench and the remaining samples are used for RL training. Neither the held-out source images nor their captions appear in the training split. After caption construction, RL training uses only caption text and does not access the original source images.

## G.2 Diverse-query ablation records

For the diverse-query ablation, the original and newly generated caption subsets are merged with deterministic sample-identity hashing and strict checks for missing, duplicated, unknown, empty, or misaligned records. The evaluation set remains the fixed DenseAlignBench split used by the original experiment.