Imago: STUDYING THE ONLINE RL RECIPE FOR VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *Imago*, an online reinforcement learning (RL) post-training recipe that matures pretrained text-to-image (T2I) generators into reliable compositional models. While modern T2I generators possess rich visual representations, under compositional constraints they frequently omit entities, bind attributes incorrectly, and violate spatial relations. RL offers a principled framework to resolve this gap, yet prior research focuses primarily on optimization algorithms, leaving the reward signal and the prompt corpus comparatively underexplored. Evaluating candidate rewards on optimization stability and visual fidelity, we find that a zero-shot VQAScore evaluated by a multimodal large language model (MLLM) provides reliable guidance, whereas preference-trained rewards are less reliable, with some triggering reward hacking. Treating the prompt corpus as the exploration environment, we further show that short and long prompts develop complementary spatial and attribute capabilities. While naive corpus mixing degrades performance, a sequential short-to-long curriculum successfully captures the strengths of both distributions. We apply the resulting recipe to FLUX.2-Klein-Base and Z-Image without architectural modifications, producing *Imago*-4B and *Imago*-6B. Both models achieve leading performance in their respective size classes across compositional capability and visual appeal, with *Imago*-6B surpassing substantially larger open-source and proprietary models on medium- and long-prompt composition benchmarks.

1 INTRODUCTION

Modern T2I models possess rich visual priors and strong generative capacity [7; 81; 91; 93]. Yet under compositional constraints, they frequently fail: entities are omitted, attributes bind to incorrect objects, and spatial arrangements collapse [56; 59; 140]. This failure reflects an unreliable mapping from complex language to coherent spatial geometry rather than an absence of visual representation. RL provides a principled framework to resolve this gap [6; 68; 83; 137]. For a capable pretrained generator, online RL does not build visual representations from scratch; it catalyzes the expression of pre-existing latent structures. This transition mirrors biological metamorphosis, where the *imago* denotes the final adult stage in which latent structures reach full expression. We view online RL post-training as guiding T2I models toward this *imago* state.

Visual online RL research remains heavily focused on optimization algorithms, concentrating on convergence acceleration and gradient variance reduction [57; 62]. In contrast, two foundational variables receive far less systematic study. First, the reward signal defining the target behavior: most rewards, such as UnifiedReward [110], are trained as offline evaluators of human preference, which can trigger aesthetic degradation or reward hacking when placed in an online loop [26; 37; 109]. Second, the prompt corpus defining the exploration environment: existing frameworks sample prompts from uncontrolled pools, leaving the impact of prompt distributions on capabilities unexamined.

These variables remain unaddressed partly because standard evaluation tooling has become uninformative. Benchmarks such as GenEval [27] and DPG-Bench [39] have saturated on modern models, failing to separate advanced architectures on complex compositions. Furthermore, common visual appeal metrics correlate strongly with one another, obscuring reward hacking when one metric is optimized and another is used for scoring (Sec. 3). Systematic progress requires a standardized evaluation protocol that spans prompt length regimes while auditing against correlated metrics. We establish this diagnostic protocol and investigate the design space of visual online RL post-training. Our primary contributions are fourfold:

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.

- **Stratified evaluation suite (Sec. 3).** We introduce a diagnostic protocol that stratifies compositional benchmarks across short, medium, and long prompt regimes, paired with an audited visual appeal metric showing minimal correlation to candidate training rewards.
- **Systematic reward analysis (Sec. 4).** We demonstrate that zero-shot prompting of a pretrained MLLM provides reliable guidance for composition and visual appeal, whereas preference-trained rewards are less reliable, with some triggering reward hacking.
- **Curriculum corpus design (Sec. 5).** We find that short and long prompts develop complementary spatial and attribute capabilities. While naive corpus mixing degrades performance, a sequential short-to-long curriculum captures the strengths of both distributions.
- **The *Imago* recipe (Sec. 6).** Combining these insights, we formulate *Imago*, a complete post-training recipe for T2I models. Applied to 4B- and 6B-scale backbones, *Imago* establishes leading accuracy across compositional benchmarks while simultaneously elevating overall visual appeal.

2 Related Work: Online RL for T2I Generation

Online RL for T2I generation involves three core components: the reward signal, the prompt corpus, and the optimization algorithm. We structure this review along these three axes (see Appendix A for an extended survey).

Reward modeling. Reward models for T2I generation predominantly build on human preference datasets [49; 76; 110; 123], sometimes incorporating task-specific or rubric-based criteria [3; 24; 65; 69; 106]. However, optimizing generative models against learned preferences frequently induces reward hacking when scoring functions misalign with true visual fidelity [1; 26; 37; 109]. Rather than introducing another reward model, we benchmark existing evaluators to identify which feedback signal provides the most reliable guidance for RL.

Training corpus. During RL, the prompt corpus defines the distribution over which the model samples outputs and receives feedback. Prior studies typically treat corpus selection as an incidental implementation choice, drawing prompts from uncontrolled, heterogeneous sources [6; 23; 57; 62; 68; 125]. Consequently, the role of prompt distribution in visual RL remains poorly understood. In this work, we treat corpus design as a primary focus and investigate how prompt composition governs downstream compositional capability.

Optimization algorithm. Algorithmic formulations for diffusion RL span score-function gradient estimators [6; 23], direct reward backpropagation [19; 82], offline preference optimization [61; 66; 104; 126; 127], and scalable online flow-matching techniques [57; 62; 63; 68; 87; 98; 125; 137; 142]. While algorithmic research routinely holds the reward and corpus constant to highlight optimizer progress, we take a complementary approach: we hold the algorithm fixed to systematically isolate the contributions of the reward signal and prompt environment.

3 Evaluating T2I Generation

Developing effective online RL recipes for T2I generation requires a rigorous evaluation protocol. We evaluate models along two complementary dimensions: *compositional capability*, measuring how faithfully an image reflects prompt specifications, and *visual appeal*, capturing overall perceptual quality [55; 65]. For compositional capability, we compile a benchmark suite spanning short, medium, and long prompts (Sec. 3.1). For visual appeal, we audit candidate preference evaluators to select a held-out preference proxy that reduces direct reuse of the optimization metric (Sec. 3.2).

3.1 Compositional Benchmarks

Standard compositional benchmarks like GenEval [27] and DPG-Bench [39] rely on brief prompts such as "*a red dog*" and suffer from score saturation on modern models [45; 97]. We therefore select five benchmarks stratified into three prompt length regimes: short (under 100 words: T2I-CompBench [41], TIIF-Bench [113], and SpatialGenEval [111]), medium (100 to 500 words: T2I-CoReBench's composition split [59]), and long (over 500 words: LongBench-T2I [140]). All five benchmarks use Qwen3.5-27B [88] as the default MLLM judge. Evaluating 19 T2I models [7; 8; 20; 22; 81; 86; 99; 100; 105; 114; 119; 132] with scores normalized to $[0, 1]$ confirms that this suite avoids saturation (Appendix D). Average scores range from 0.71 to 0.79 on short prompts and drop to between 0.63 and 0.64 on medium and long prompts (Fig. 1, right), maintaining clear headroom to discriminate model capabilities.

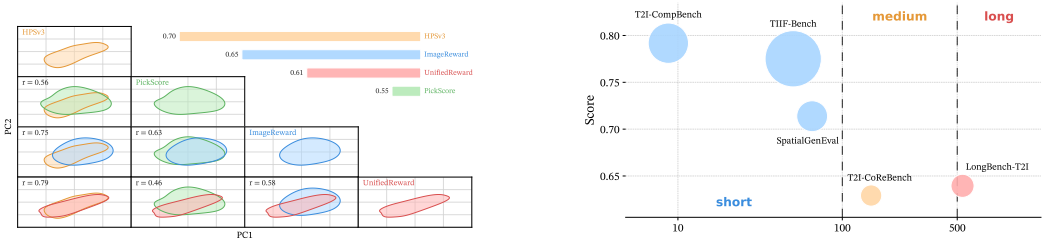


Figure 1: **Left:** PCA projection of four preference evaluators onto two principal components. Pairwise panels report Pearson correlations (r), and the bar chart summarizes each evaluator's mean correlation ($\bar{r}$) with the others. **Right:** Five compositional benchmarks plotted by mean prompt length (x-axis, log scale) and mean score across 19 T2I models (y-axis). Bubble sizes indicate sample counts, and vertical dashed lines delineate the short, medium, and long prompt regimes.

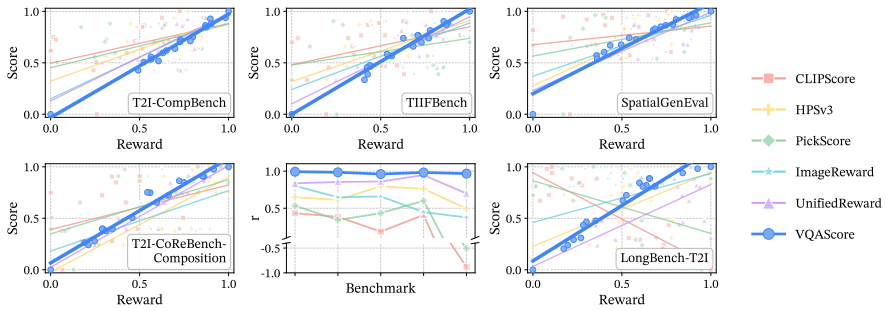


Figure 2: **Cross-model comparison of six candidate rewards across five compositional benchmarks.** The five benchmark panels plot 19 T2I models by their min-max normalized mean reward scores (x-axis) against normalized benchmark scores (y-axis), with linear regression fits. The bottom-center panel reports the corresponding Pearson correlations (r), with benchmarks ordered from left to right by increasing prompt length.

3.2 Visual Appeal Evaluator

Learned human preference models are commonly used to measure visual appeal. However, adopting correlated preference models as both RL rewards and evaluation metrics introduces evaluation bias, favoring models tuned on aligned objectives [1; 26; 68]. To assess this dependency, we evaluate four preference models: PickScore [49], ImageReward [123], HPSv3 [76], and UnifiedReward [110]. Each evaluator scores the prompt-image pairs generated by the 19 T2I models across all five compositional benchmarks, from which we compute pairwise Pearson correlations over the pooled samples. As shown in Fig. 1 (left), PickScore displays the lowest average correlation with other candidates ($\bar{r} = 0.55$), whereas correlations among the remaining models range from 0.61 to 0.70. We adopt PickScore over DrawBench [93] prompts as a held-out proxy for visual preference throughout this work. Its low average correlation reduces direct reuse of the optimization metric.

4 The Choice of Reward as Optimization Signal

Optimizing T2I models via RL requires establishing a target evaluation objective, and then formulating a reliable reward signal to guide optimization toward it. Having established the evaluation suites in Sec. 3, this section focuses on the reward signal. We analyze candidate rewards from two perspectives: first as evaluators, measuring how reliably they track compositional capability (Sec. 4.1); and second as training objectives, assessing how effectively they guide fine-tuned models (Sec. 4.2). Finally, we investigate whether scaling the reward backbone yields further improvements (Sec. 4.3).

4.1 Reward as Evaluator: Does It Track Compositional Capability?

Before deploying a reward model within the RL loop, we evaluate whether its scores correlate with the compositional benchmarks we aim to optimize. Specifically, we assess candidate rewards on the prompt-image pairs generated by 19 T2I models across the five benchmarks from Sec. 3.1. For each benchmark, we compute the Pearson correlation between the mean reward scores and the benchmark scores across all 19 models, measuring how reliably each reward tracks compositional capability.

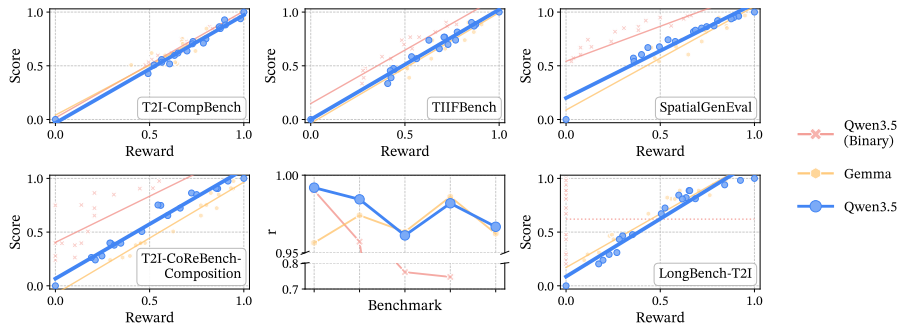


Figure 3: **Controlled comparison of VQAScore scoring formats and MLLM backbones.** Starting from graded Qwen3.5-9B, we ablate one component at a time by introducing binary scoring or substituting Gemma-4-E4B-it. The five benchmark panels plot 19 T2I models by their min-max normalized mean reward scores (x-axis) against normalized benchmark scores (y-axis), with linear regression fits per configuration. The bottom-center panel reports the corresponding Pearson correlations (r), with benchmarks ordered from left to right by increasing prompt length.

Table 1: **Compositional performance across four candidate reward models.** Scores are reported on each benchmark's native scale for the original base models and fine-tuned variants, all trained on the same corpus under a shared RL protocol. **Bold** marks the best score for each benchmark within each base-model block.

Reward	T2I-CompBench	TIIF-Bench	SpatialGenEval	T2I-CoReBench	LongBench-T2I
FLUX.2-Klein-Base-4B					
Base (no RL)	3.983	78.24	74.51	66.51	3.385
CLIPScore	4.024	80.20	74.33	67.18	3.162
HPSv3	3.896	75.85	71.46	65.29	3.035
UnifiedReward	4.012	**80.92**	76.12	71.21	3.506
VQAScore	**4.041**	80.87	**78.00**	**72.41**	**3.808**
Z-Image					
Base (no RL)	4.087	86.44	78.79	85.16	3.848
CLIPScore	4.117	87.06	78.70	86.89	3.834
HPSv3	4.008	82.14	75.80	84.44	3.548
UnifiedReward	4.114	85.95	80.02	87.92	4.008
VQAScore	**4.152**	**87.56**	**80.98**	**88.56**	**4.125**

Rewards. We evaluate six rewards across two paradigms: preference-trained models (HPSv3 [76], PickScore [49], ImageReward [123], and UnifiedReward [110]) and pretrained models evaluated zero-shot (CLIPScore [34] and VQAScore [67]). For VQAScore, we prompt Qwen3.5-9B [88] to score semantic alignment from 1 to 10 using the template in Fig. 12. Because UnifiedReward and VQAScore share the same Qwen3.5-9B backbone and differ only in preference training, this comparison isolates how preference tuning affects a reward's ability to assess compositional quality.

Zero-shot VQAScore tracks compositional capability most reliably. As shown in Fig. 2, VQAScore achieves the highest Pearson correlation across all five benchmarks and maintains high correlation as prompt length increases, whereas other candidate rewards degrade. In the matched-backbone comparison, UnifiedReward consistently underperforms VQAScore despite sharing the same base MLLM. This discrepancy indicates that preference fine-tuning can bias the model toward broad perceptual preferences, weakening its ability to strictly verify compositional constraints.

Graded scoring generalizes across different MLLM backbones. To isolate the key factors in VQAScore, we independently ablate the scoring format and the MLLM backbone. First, we replace the 1-to-10 graded score with a binary criterion that requires complete prompt-image alignment for a positive score. Second, to verify that high correlations do not stem from shared biases with our Qwen3.5-27B benchmark judge, we introduce a cross-family control by substituting Qwen3.5-9B with Gemma-4-E4B-it [30] under identical settings.

As shown in Fig. 3, binary scoring causes a sharp drop in correlation as prompt length increases. Under binary scoring, a single missed detail turns a partial match into a complete failure, obscuring fine-grained differences in semantic alignment. In contrast, changing the MLLM backbone has minimal impact: Gemma-4-E4B-it achieves correlations comparable to Qwen3.5-9B across all benchmarks. This confirms that the strong correlations are driven by the graded scoring design rather than shared model biases with the benchmark judge.

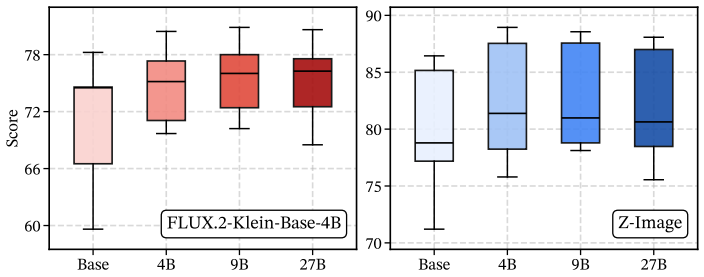


Figure 4: **Compositional performance across VQAScore backbone sizes.** Box plots show the distribution of normalized scores across the five compositional benchmarks for each base model before and after training with graded VQAScore across three Qwen3.5 scales. Center lines indicate medians, and whiskers indicate the score ranges across benchmarks.

Table 2: **PickScore comparison.** Mean PickScore ($\times 100$) on DrawBench. Best per column in **bold**.

Rewards	*FLUX.2-Klein-Base-4B*	*Z-Image*
Baseline	22.35	22.37
+ CLIPScore	22.46	22.41
+ HPSv3	22.80	23.09
+ UnifiedReward	**22.91**	**23.17**
+ VQAScore	**22.91**	22.78

Table 3: **Benchmark correlation across VQAScore backbone sizes.** Values report Pearson correlations across the 19 T2I models between graded VQAScore and compositional benchmark scores. **Bold** marks the highest value per column.

Backbone	Mean	T2I-CompBench	TIIF-Bench	SpatialGenEval	T2I-CoReBench	LongBench-T2I
Qwen3.5-4B	**0.986**	0.989	**0.985**	**0.980**	**0.993**	**0.980**
Qwen3.5-9B	0.977	0.992	0.984	0.961	0.982	0.967
Qwen3.5-27B	0.971	**0.993**	0.982	0.949	0.975	0.956

4.2 Reward as Signal: Does a Better Evaluator Produce a Better Model?

While strong correlation with benchmarks validates a reward as an evaluation metric, it does not guarantee its effectiveness as an optimization signal. To test whether benchmark correlation translates into training performance, we use graded Qwen3.5-9B VQAScore, CLIPScore, HPSv3, and UnifiedReward as RL rewards. To isolate the effect of each reward signal, we hold all other training configurations constant. Specifically, we fine-tune FLUX.2-Klein-Base-4B for 500 steps and Z-Image for 300 steps using DiffusionNFT [137]. During training, we apply neither classifier-free guidance (CFG) [36] nor explicit KL regularization [51]; at evaluation, each model uses the standard guidance configuration of its backbone. The complete experimental setup is detailed in Appendix C.

Benchmark correlation translates to training performance. In Tab. 1, training with VQAScore yields the highest compositional scores across both base models on most benchmarks. Crucially, the comparison between VQAScore and UnifiedReward directly mirrors our findings in Sec. 4.1: the reward model that tracks benchmark scores better also trains a better model. This consistency confirms that benchmark correlation reliably predicts how well a reward guides model training.

VQAScore improves composition without compromising visual appeal. As reported in Tab. 2, VQAScore achieves visual appeal scores comparable to preference-trained rewards across both base models. Qualitative comparisons in Fig. 11 support this finding. Preference-trained rewards such as HPSv3 can exhibit content drift, introducing unprompted subjects that raise aesthetic scores at the expense of prompt fidelity. In contrast, VQAScore preserves prompt fidelity while improving overall visual quality. We attribute this quality gain partly to training without CFG. Because VQAScore evaluates unguided samples during optimization, the generator must produce clear images for the MLLM to recognize and reward semantic alignment, which in turn encourages higher visual quality.

4.3 Scaling Reward Backbone: Does a Larger Model Help?

Having shown that graded Qwen3.5-9B VQAScore provides an effective reward, we next investigate whether scaling the underlying MLLM backbone further improves compositional performance. We evaluate Qwen3.5 across three parameter scales while keeping all other training settings identical to those in Sec. 4.2.

A compact MLLM is sufficient as a reward model. We compare the normalized benchmark scores for both base models before and after training with each reward backbone. As shown in Fig. 4, we find that the 4B backbone alone substantially improves compositional performance on both base models, demonstrating that effective guidance does not require a large reward network.



Figure 5: **Controlled expansion of a seed prompt into four descriptive designs.** Starting from a shared seed prompt (center), four long-prompt designs isolate distinct descriptive axes. Highlighted text denotes scene-graph primitives: entities, attributes, and relations. The balanced design distributes focus across all three primitives, while the remaining three concentrate on individual primitives. Accompanying images are synthesized by Nano-Banana for visualization.

Moreover, performance remains comparable across the 4B, 9B, and 27B configurations, showing no consistent gains from scaling. We attribute this saturation to the fact that the 4B model already tracks compositional metrics reliably. As reported in Tab. 3, all three backbones correlate strongly with benchmark scores across the 19 evaluated models. In fact, the 4B variant achieves the highest correlation on four of the five benchmarks, whereas larger backbones provide no clear advantage. Because larger backbones do not evaluate compositional quality more accurately, scaling the reward backbone does not further improve training outcomes.

> *With a simple graded scoring prompt, a small pretrained MLLM serves as an effective zero-shot reward for T2I RL. It improves compositional performance, matches preference-trained rewards in visual quality, and remains less susceptible to the reward hacking that such models can exhibit.*

5 The Design of Training Corpus

In online RL for T2I synthesis, the prompt corpus constitutes the learning environment itself, directly governing the generator's exploration space and optimization trajectory [11; 48]. Yet, existing methods typically rely on heuristically assembled prompt collections [57; 62; 68; 125], leaving the role of corpus design largely unexplored. To address this gap, we construct and analyze controlled prompt environments. We first decompose prompts into verifiable scene-graph primitives to build corpora varying in prompt design and length (Sec. 5.1), then isolate how each axis impacts fine-tuned models (Sec. 5.2). Finally, observing length-dependent specialization trade-offs, we introduce a short-to-long sequential schedule that combines the strengths of both length regimes (Sec. 5.3).

5.1 Building Controlled Prompt Environments

Decomposition via scene graphs. Grounded in scene-graph representations [2; 43; 44; 50], we decompose text prompts into four elements: *entities* (present objects), *attributes* (verifiable visual properties), *relations* (spatial, physical, or structural interactions), and *holistic modifiers* (style, mood, and global context). We restrict our corpus to the first three objective primitives, omitting holistic modifiers because they lack grounded visual referents. This filtering ensures that downstream generation failures stem directly from compositional errors rather than stylistic ambiguities [84].

Phase 1: Short-prompt curation. We construct the training corpus in two sequential phases, progressively scaling prompt length and structural complexity. In the first phase, we curate a short-prompt base corpus from Fine-T2I [74], which contains roughly 3M prompts averaging 30 words. We embed each prompt with Qwen3-Embedding [135] and apply similarity-aware deduplication to retain 200K unique candidates. We then prompt Qwen3.5-27B to eliminate fragmented or ambiguous descriptions, yielding a refined set of 150K prompts designated ***Imago*-Short**.

Phase 2: Long-prompt expansion. Because prompts in *Imago*-Short average only 30 words, they lack the descriptive detail needed to train models on complex multi-object scenes. The second phase therefore expands each short prompt into four parallel prompts averaging roughly 200 words via Qwen3.5-27B, each adopting a distinct prompt design:

- **Balanced**: Distributes descriptions evenly across entities, attributes, and relations.

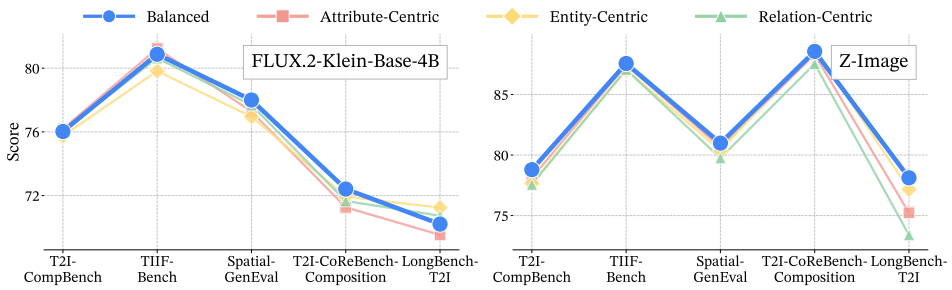


Figure 6: **Balanced prompt design achieves the strongest overall compositional performance.** We report benchmark scores for FLUX.2-Klein-Base-4B (left) and Z-Image (right) fine-tuned under four distinct prompt designs. Although single-axis designs occasionally lead on specific benchmarks, the balanced design (blue) delivers the most consistent performance across both base models.

Table 4: **Complementary specialization across prompt lengths.** Benchmarks are ordered from left to right by increasing prompt length. The best score for each base model is highlighted in **bold** with green shading, where darker hues reflect benchmarks with longer average prompt lengths.

Training Corpus	T2I CompBench	TIIF Bench	Spatial GenEval	T2I CoReBench	LongBench T2I
FLUX.2-Klein-Base-4B					
Base	3.983	78.24	74.51	66.51	3.385
Imago-Short	**4.110**	**82.78**	77.58	**73.68**	3.759
Imago-Long	4.041	80.87	**78.00**	72.41	**3.808**
Z-Image					
Base	4.087	86.44	78.79	85.16	3.848
Imago-Short	4.135	**87.76**	**80.98**	**89.69**	4.086
Imago-Long	**4.152**	87.56	**80.98**	88.56	**4.125**

- **Entity-centric**: Focuses on introducing and specifying distinct objects.
- **Attribute-centric**: Focuses on visual properties such as colors, materials, and textures.
- **Relation-centric**: Focuses on spatial arrangements and interactions between objects.

Fig. 5 illustrates the four designs expanded from a shared seed prompt in *Imago*-Short. Complete details on corpus curation, filtering, and expansion templates are provided in Appendix B.

5.2 Effects of Prompt Design and Prompt Length

Using the corpora constructed in Sec. 5.1, we evaluate how prompt design and prompt length affect compositional performance. Across all experiments, we fine-tune FLUX.2-Klein-Base-4B and Z-Image using a graded Qwen3.5-9B VQAScore reward under the RL setup described in Sec. 4.2.

Balanced prompt design yields the strongest overall performance. We fine-tune each base model independently under the four prompt designs. As shown in Fig. 6, balanced prompt design achieves the highest scores across most benchmarks on both base models. We attribute this advantage to its structural symmetry: distributing descriptive focus evenly across entities, attributes, and relations encourages models to acquire these compositional skills jointly, avoiding bias toward a single descriptive axis. Based on these results, we designate the balanced corpus as ***Imago*-Long** and adopt it for subsequent experiments.

***Imago*-Short and *Imago*-Long exhibit complementary strengths.** Beyond prompt design, the short and expanded prompt distributions induce different performance profiles across benchmarks. As reported in Tab. 4, while both corpora outperform the unoptimized base models, they display distinct advantages across benchmark lengths. *Imago*-Short yields stronger performance on concise benchmarks emphasizing direct attribute binding, such as TIIF-Bench. In contrast, *Imago*-Long excels on detailed prompts in LongBench-T2I. These results show that neither corpus consistently outperforms the other across benchmarks of varying prompt lengths.

5.3 Sequential Training: From Short Binding to Long Composition

The complementary strengths of *Imago*-Short and *Imago*-Long (Sec. 5.2) naturally raise the question of how to integrate both corpora during training so that a single model inherits the advantages of each. To investigate this, we compare two schedules under the same total number of parameter updates. Whereas **joint training** samples uniformly from both corpora throughout optimization,

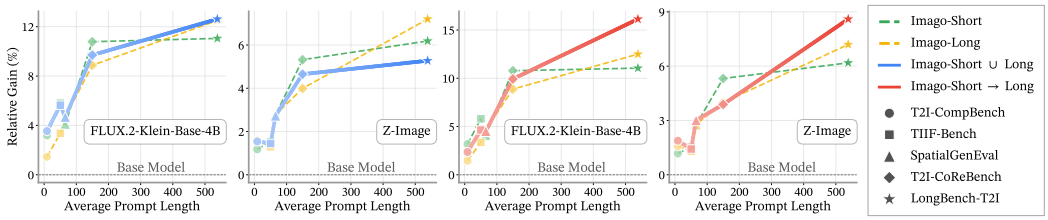


Figure 7: **Sequential training resolves the trade-off across prompt lengths.** Relative performance gains over the base model are plotted against average benchmark prompt length. The left pair illustrates joint training (Imago-Short ∪ Long), where naive mixing fails to reliably match single-corpus bounds. The right pair illustrates sequential training (Imago-Short → Long), which preserves short-prompt accuracy while unlocking long-prompt composition, notably surpassing pure *Imago*-Long training on LongBench-T2I.

sequential training trains on *Imago*-Short for the first 50% of steps before switching to *Imago*-Long, additionally resetting the optimizer.

Joint training degrades performance across prompt lengths. As depicted in Fig. 7, joint training slightly underperforms pure *Imago*-Short on concise prompts and fails to match *Imago*-Long on complex ones across both base models. On LongBench-T2I, jointly trained Z-Image even falls below both single-corpus baselines. We attribute this degradation to conflicting demands across prompt lengths. Short prompts require local attribute binding, whereas long prompts require coordinating many objects and spatial relations. Optimizing for both at once pulls the model in incompatible directions, preventing it from excelling at either concise binding or complex composition.

Sequential training unlocks long-prompt composition. In contrast, sequential training achieves the highest LongBench-T2I scores across both models, notably surpassing pure *Imago*-Long training while retaining competitive performance on short-prompt benchmarks (Fig. 7). This advantage stems from staged capability acquisition. Training on *Imago*-Short first cements rigorous local binding and structural restraint. Building upon this foundation, the subsequent *Imago*-Long phase allows the model to expand scene capacity and coordinate dense, multi-relation scenes far more effectively than training on long prompts from scratch, which forces the model to learn primitive grounding and complex composition simultaneously.

> *The prompt corpus defines the RL environment for T2I generation. Among prompt designs, balancing entities, attributes, and relations yields the strongest overall performance. Across prompt lengths, a short-to-long sequential schedule avoids optimization conflicts from joint mixing, maintaining performance on concise prompts while unlocking long-prompt composition to surpass dedicated long-prompt training.*

6 *Imago*: An Online RL Recipe for Compositional T2I

Integrating the findings from previous sections, we introduce ***Imago***, an online RL post-training recipe designed to improve compositional generation without architectural modifications. We instantiate *Imago* across two parameter scales by fine-tuning FLUX.2-Klein-Base-4B and Z-Image-6B. Training adopts zero-shot VQAScore with a Qwen3.5-9B backbone as the reward signal, applying the graded instruction template described in Sec. 4.1. Prompts follow the sequential curriculum established in Sec. 5.3, transitioning from *Imago*-Short to *Imago*-Long. Both models are optimized using DiffusionNFT [137] without CFG [36] or KL regularization [51]. Hyperparameter and infrastructure details are provided in Appendix C.

Baseline models. We compare *Imago* against 19 T2I models spanning three categories. Proprietary systems include Imagen-4-Fast and Nano-Banana [29]. Unified multimodal models include DeepGen [105], Bagel [20], and GLM-Image [132]. Diffusion and flow-matching models include SANA [119], the SD-3.5 series [22], SDXL [81], Lumina-Image-2 [86], LongCat-Image [100], the FLUX.1 [7] and FLUX.2 [8] families, Qwen-Image [114], and Z-Image [99]. All models are evaluated on the five compositional benchmarks and DrawBench visual appeal suite introduced in Sec. 3.

***Imago* outperforms comparable base models while maintaining visual quality.** As reported in Tab. 5, *Imago* achieves the highest scores across all five compositional benchmarks within both the 4B and 6B size classes. In addition to compositional improvements, *Imago* increases PickScore

Table 5: **Performance comparison with competitive T2I models.** Models are evaluated across five compositional benchmarks alongside visual appeal measured by PickScore on DrawBench prompts. **Bold** marks the best score within each model category; rows corresponding to *Imago* are highlighted.

Model	Size	Short			Medium	Long	PickScore
		T2I-CompBench	TIIF-Bench	SpatialGenEval	T2I-CoReBench	LongBench-T2I	DrawBench
Proprietary Model							
Imagen-4-Fast	–	4.132	79.82	70.05	73.74	3.067	–
Nano-Banana	–	**4.282**	**90.39**	**79.46**	**83.87**	**4.145**	–
Unified Multimodal Model							
DeepGen	5B	**4.147**	**83.46**	**75.85**	61.34	**3.600**	**23.01**
Bagel	7B	3.935	75.13	71.08	50.54	2.937	22.75
GLM-Image	16B	4.017	80.89	74.07	**72.04**	3.540	22.98
Diffusion / Flow Matching							
SANA	1.6B	3.858	69.16	67.10	44.74	2.859	22.66
SD-3.5-Medium	2.5B	3.877	71.75	66.02	42.58	2.508	22.53
SDXL	2.6B	3.468	56.11	47.31	28.62	2.078	22.16
Lumina-Image-2	2.6B	3.839	67.34	69.22	51.62	2.605	22.28
LongCat-Image	6B	3.999	81.86	75.16	72.25	3.288	22.73
SD-3.5-Large	8.1B	3.847	72.07	65.00	43.90	2.446	22.84
FLUX.2-Klein-Base	9B	4.112	85.31	77.99	81.33	3.536	22.62
FLUX.2-Klein	9B	4.155	85.84	79.01	81.10	3.773	23.09
FLUX.1-schnell	12B	3.905	75.74	71.62	51.59	2.916	22.62
FLUX.1-dev	12B	3.781	71.37	69.46	51.72	2.636	22.83
Qwen-Image	20B	3.921	79.57	78.37	77.84	3.676	22.82
FLUX.2-dev	32B	**4.200**	**89.69**	**80.04**	**86.60**	**3.882**	**23.24**
FLUX.2-Klein	4B	4.100	80.92	77.44	70.54	3.677	22.81
FLUX.2-Klein-Base	4B	3.983	78.24	74.51	66.51	3.385	22.35
Imago-4B	4B	**4.102**	**81.78**	**79.10**	**73.50**	**3.953**	**22.92**
Z-Image-Turbo	6B	3.989	81.84	75.60	78.73	3.559	22.86
Z-Image	6B	4.087	86.44	78.79	85.16	3.848	22.37
Imago-6B	6B	**4.191**	**88.77**	**82.71**	**90.26**	**4.185**	**23.08**

over the unmodified base models, verifying that semantic alignment can be strengthened without sacrificing perceptual quality.

***Imago* surpasses larger models on long-prompt composition.** Beyond its size class, *Imago*-6B achieves the highest scores on medium- and long-prompt benchmarks, outperforming substantially larger models such as FLUX.2-dev (32B) and the proprietary Nano-Banana on T2I-CoReBench and LongBench-T2I. Simultaneously, it maintains competitive performance on short-prompt tasks. These results demonstrate that open-source models already possess strong generative capabilities, and that a structured online RL post-training recipe can achieve compositional gains that previously required scaling model parameters.

7 Conclusion

In this work, we systematically investigated the design of online RL post-training recipes for T2I generation, focusing on the reward signal, prompt corpus composition, and curriculum schedule. We demonstrated that a compact pretrained MLLM with graded scoring provides an effective zero-shot reward that is less prone to the reward hacking that preference-trained models can exhibit. Furthermore, treating the prompt corpus as the primary training environment reveals that a sequential short-to-long curriculum mitigates the performance trade-offs observed under naive prompt mixing. Integrating these findings, *Imago* achieves consistent improvements across compositional benchmarks within its size class and outperforms considerably larger proprietary and open-source models on complex, long-prompt tasks. While our study focuses on parameter-efficient fine-tuning via LoRA, scaling to full-parameter optimization offers a promising direction for future exploration. Our results highlight reward and corpus design as consequential components of online RL post-training, complementary to algorithmic improvements.

ETHICS STATEMENT

This work studies reinforcement learning for improving the compositional fidelity of text-to-image (T2I) generative models. All experiments use publicly available pretrained generators and prompt corpora derived from public datasets, and involve no human subjects, private data, or personally identifiable information. As with all T2I systems, the models studied here can be misused to produce misleading or harmful imagery and may reflect biases present in their pretraining data. Our method optimizes prompt fidelity rather than introducing new generative capabilities, and we encourage responsible deployment with appropriate content-safety safeguards.

REPRODUCIBILITY STATEMENT

We provide extensive details to support reproduction of our results. Appendix B documents the complete data-curation and prompt-expansion pipeline, including the language-model templates and auditor protocol used to build the *Imago*-Short and *Imago*-Long corpora. Appendix C specifies all reinforcement learning hyperparameters, the optimization objective, and the hardware configuration used to produce *Imago*-4B and *Imago*-6B. Appendix D describes the generation settings, judge models, and scoring protocols for every quantitative result reported in the main text.

USE OF LARGE LANGUAGE MODELS

Large language models are a core component of the methodology studied in this paper, serving as zero-shot reward models and as data-curation and prompt-expansion tools; their use in that capacity is described in the main text and appendices. Separately, in preparing this manuscript the authors used LLM-based assistants for language polishing and to assist with minor code and LaTeX formatting tasks. All research ideas, experimental design, analyses, and conclusions are the authors' own, and the authors verified the correctness of all AI-assisted text and code. LLMs are not eligible for authorship and did not contribute research contributions.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

[2] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation, 2019. URL https://arxiv.org/abs/1909.05379.

[3] Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences, 2025. URL https://arxiv.org/abs/2506.02095.

[4] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models, 2023. URL https://arxiv.org/abs/2304.05390.

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2022. URL https://arxiv.org/abs/2211.01324.

[6] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. URL https://arxiv.org/abs/2305.13301.

[7] Black Forest Labs. Announcing Black Forest Labs, 2024. URL https://bfl.ai/blog/24-08-01-bfl. Blog post; introduces the FLUX.1 text-to-image model family (schnell, dev, pro).

[8] Black Forest Labs. FLUX.2: Frontier visual intelligence, 2025. URL https://bfl.ai/blog/flux-2. Blog post; introduces the FLUX.2 text-to-image model family.

[9] Muhammad Atif Butt, Alexandra Gomez-Villa, Tao Wu, Javier Vazquez-Corral, Joost Van De Weijer, and Kai Wang. Gencolorbench: A color evaluation benchmark for text-to-image generation models, 2025. URL https://arxiv.org/abs/2510.20586.

[10] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Duojun Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, Yang Li, Zhenxi Li, Zhimin Li, Jiaxin Lin, Lucaz Liu, Shu Liu, Songtao Liu, Yu Liu, Yuhong Liu, Yanxin Long, Fanbin Lu, Qinglin Lu, Yuyang Peng, Yuanbo Peng, Xiangwei Shen, Yixuan Shi, Jiale Tao, Yangyu Tao, Qi Tian, Pengfei Wan, Chunyu Wang, Kai Wang, Lei Wang, Linqing Wang, Lucas Wang, Qixun Wang, Weiyan Wang, Hao Wen, Bing Wu, Jianbing Wu, Yue Wu, Senhao Xie, Fang Yang, Miles Yang, Xiaofeng Yang, Xuan Yang, Zhantao Yang, Jingmiao Yu, Zheng Yuan, Chao Zhang, Jian-Wei Zhang, Peizhen Zhang, Shi-Xue Zhang, Tao Zhang, Weigang Zhang, Yepeng Zhang, Yingfang Zhang, Zihao Zhang, Zijian Zhang, Penghao Zhao, Zhiyuan Zhao, Xuefei Zhe, Jianchen Zhu, and Zhao Zhong. Hunyuanimage 3.0 technical report, 2025. URL https://arxiv.org/abs/2509.23951.

[11] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

[12] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL https://arxiv.org/abs/2405.09818.

[13] Chieh-Yun Chen, Zhonghao Wang, Qi Chen, Zhifan Ye, Min Shi, Yue Zhao, Yinan Zhao, Hui Qu, Wei-An Lin, Yiru Shen, Ajinkya Kale, Irfan Essa, and Humphrey Shi. Mapreduce lora: Advancing the pareto front in multi-preference optimization for generative models, 2026. URL https://arxiv.org/abs/2511.20629.

[14] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models—architecture, training and dataset, 2025. URL https://arxiv.org/abs/2505.09568.

[15] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. URL https://arxiv.org/abs/2310.00426.

[16] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. URL https://arxiv.org/abs/2501.17811.

[17] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation, 2023. URL https://arxiv.org/abs/2305.15328.

[18] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation, 2024. URL https://arxiv.org/abs/2310.18235.

[19] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2023. URL https://arxiv.org/abs/2309.17400.

[20] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

[21] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. URL https://arxiv.org/abs/2105.13290.

[22] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

[23] 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, 2023. URL https://arxiv.org/abs/2305.16381.

[24] Xuelu Feng, Yunsheng Li, Ziyu Wan, Zixuan Gao, Junsong Yuan, Dongdong Chen, and Chunming Qiao. Rubricrl: Simple generalizable rewards for text-to-image generation, 2025. URL https://arxiv.org/abs/2511.20651.

[25] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. URL https://arxiv.org/abs/2203.13131.

[26] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

[27] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

[28] Shelly Golan, Michael Finkelson, Ariel Bereslavsky, Yotam Nitzan, and Or Patashnik. Paretoslider: Diffusion models post-training for continuous reward control, 2026. URL https://arxiv.org/abs/2604.20816.

[29] Google. Nano Banana Pro, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/. Blog post; introduces the Nano Banana Pro image generation model.

[30] Google. Gemma 4: New open models for developers, 2026. URL https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/. Blog post; introduces the Gemma 4 open model family.

[31] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2021. URL https://arxiv.org/abs/2111.14822.

[32] Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization, 2024. URL https://arxiv.org/abs/2406.06382.

[33] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation, 2024. URL https://arxiv.org/abs/2412.18150.

[34] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017. URL https://arxiv.org/abs/1706.08500.

[36] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

[37] Yunqi Hong, Kuei-Chun Kao, Hengguang Zhou, and Cho-Jui Hsieh. Understanding reward hacking in text-to-image reinforcement learning, 2026. URL https://arxiv.org/abs/2601.03468.

[38] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[39] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. URL https://arxiv.org/abs/2403.05135.

[40] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023. URL https://arxiv.org/abs/2303.11897.

[41] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation, 2025. URL https://arxiv.org/abs/2307.06350.

[42] Inclusion AI, Tiwei Bie, Haoxing Chen, Tieyuan Chen, Zhenglin Cheng, Long Cui, Kai Gan, Zhicheng Huang, Zhenzhong Lan, Haoquan Li, Jianguo Li, Tao Lin, Qi Qin, Hongjun Wang, Xiaomei Wang, Haoyuan Wu, Yi Xin, and Junbo Zhao. Llada2.0-uni: Unifying multimodal understanding and generation with diffusion large language model, 2026. URL https://arxiv.org/abs/2604.20796.

[43] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3678, 2015. doi: 10.1109/CVPR.2015.7298990.

[44] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs, 2018. URL https://arxiv.org/abs/1804.01622.

[45] Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation, 2025. URL https://arxiv.org/abs/2512.16853.

[46] Seyed Amir Kasaei, Ali Aghayari, Arash Marioriyad, Niki Sepasian, MohammadAmin Fazli, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Evaluating the evaluators: Metrics for compositional text-to-image generation, 2025. URL https://arxiv.org/abs/2509.21227.

[47] Seungwook Kim and Minsu Cho. Improving text-to-image generation with intrinsic self-confidence rewards, 2026. URL https://arxiv.org/abs/2603.00918.

[48] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

[49] 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, 2023. URL https://arxiv.org/abs/2305.01569.

[50] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73, 2016. URL https://api.semanticscholar.org/CorpusID:4492210.

[51] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. doi: 10.1214/aoms/1177729694.

[52] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

[53] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

[54] Seung Hyun Lee, Yinxiao Li, Junjie Ke, Innfarn Yoo, Han Zhang, Jiahui Yu, Qifei Wang, Fei Deng, Glenn Entis, Junfeng He, Gang Li, Sangpil Kim, Irfan Essa, and Feng Yang. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation, 2024. URL https://arxiv.org/abs/2401.05675.

[55] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models, 2023. URL https://arxiv.org/abs/2311.04287.

[56] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. URL https://arxiv.org/abs/2406.13743.

[57] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2026. URL https://arxiv.org/abs/2507.21802.

[58] Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion, 2026. URL https://arxiv.org/abs/2603.06577.

[59] Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play?, 2026. URL https://arxiv.org/abs/2509.03516.

[60] Rui Li, Ke Hao, Yuanzhi Liang, Haibin Huang, Chi Zhang, Yun Gu, and XueLong Li. Learning to credit the right steps: Objective-aware process optimization for visual generation, 2026. URL https://arxiv.org/abs/2604.19234.

[61] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility, 2024. URL https://arxiv.org/abs/2404.04465.

[62] Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models, 2025. URL https://arxiv.org/abs/2509.06040.

[63] Yuming Li, Qingyu Li, Chengyu Bai, Xiangyang Luo, Zeyue Xue, Wenyu Qin, Meng Wang, Yikai Wang, and Shanghang Zhang. Aegpo: Adaptive entropy-guided policy optimization for diffusion models, 2026. URL https://arxiv.org/abs/2602.06825.

[64] Zijie Li, Yichun Shi, Jingxiang Sun, Ye Wang, Yixuan Huang, Zhiyao Guo, Xiaochen Lian, Peihao Zhu, Yu Tian, Zhonghua Zhai, and Peng Wang. Mmcore: Multimodal connection with representation aligned latent embeddings, 2026. URL https://arxiv.org/abs/2604.19902.

[65] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation, 2024. URL https://arxiv.org/abs/2312.10240.

[66] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization, 2024. URL https://arxiv.org/abs/2406.04314.

[67] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. URL https://arxiv.org/abs/2404.01291.

[68] 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, 2025. URL https://arxiv.org/abs/2505.05470.

[69] Jinlong Liu, Wanggui He, Peng Zhang, Mushui Liu, Hao Jiang, and Pipei Huang. Promptecho: Annotation-free reward from vision-language models for text-to-image reinforcement learning, 2026. URL https://arxiv.org/abs/2604.12652.

[70] Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation, 2026. URL https://arxiv.org/abs/2604.24763.

[71] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[72] Yifu Luo, Xinhao Hu, Keyu Fan, Haoyuan Sun, Zeyu Chen, Bo Xia, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement learning meets masked generative models: Mask-grpo for text-to-image generation, 2025. URL https://arxiv.org/abs/2510.13418.

[73] Xiaoxiao Ma, Jiachen Lei, Tianfei Ren, Jie Huang, Siming Fu, Aiming Hao, Jiahong Wu, Xiangxiang Chu, and Feng Zhao. Mar-grpo: Stabilized grpo for ar-diffusion hybrid image generation, 2026. URL https://arxiv.org/abs/2604.06966.

[74] Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-t2i: An open, large-scale, and diverse dataset for high-quality t2i fine-tuning, 2026. URL https://arxiv.org/abs/2602.09439.

[75] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2411.07975.

[76] Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025. URL https://arxiv.org/abs/2508.03789.

[77] Ziqi Ni, Yuanzhi Liang, Rui Li, Yi Zhou, Haibin Huang, Chi Zhang, and Xuelong Li. Seeing what matters: Visual preference policy optimization for visual generation, 2026. URL https://arxiv.org/abs/2511.18719.

[78] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021. URL https://arxiv.org/abs/2112.10741.

[79] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation, 2025. URL https://arxiv.org/abs/2503.07265.

[80] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

[81] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

[82] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023. URL https://arxiv.org/abs/2310.03739.

[83] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients, 2024. URL https://arxiv.org/abs/2407.08737.

[84] Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. Diffusion beats autoregressive in data-constrained settings, 2025. URL https://arxiv.org/abs/2507.15857.

[85] Rishav Pramanik, Ian E. Nielsen, Jeff Smith, Saurav Pandit, Ravi P. Ramachandran, and Zhaozheng Yin. Saneval: Open-vocabulary compositional benchmarks with failure-mode diagnosis, 2026. URL https://arxiv.org/abs/2602.00249.

[86] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu Qiao, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework, 2025. URL https://arxiv.org/abs/2503.21758.

[87] You Qin, Linqing Wang, Hao Fei, Roger Zimmermann, Liefeng Bo, Qinglin Lu, and Chunyu Wang. Soar: Self-correction for optimal alignment and refinement in diffusion models, 2026. URL https://arxiv.org/abs/2604.12617.

[88] Qwen Team. Qwen3.5, 2025. URL https://qwen.ai/blog?id=qwen3.5. Blog post; introduces the Qwen3.5 language model series.

[89] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

[90] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

[91] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[92] Candace Ross, Melissa Hall, Adriana Romero Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency, 2024. URL https://arxiv.org/abs/2412.13989.

[93] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

[94] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

[95] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback, 2023. URL https://arxiv.org/abs/2307.04749.

[96] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023. URL https://arxiv.org/abs/2307.00716.

[97] Bingda Tang, Boyang Zheng, Xichen Pan, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. *arXiv preprint arXiv:2505.10046*, 2025.

[98] Bingda Tang, Yuhui Zhang, Xiaohan Wang, Jiayuan Mao, Ludwig Schmidt, and Serena Yeung-Levy. V-grpo: Online reinforcement learning for denoising generative models is easier than you think, 2026. URL https://arxiv.org/abs/2604.23380.

[99] Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer, 2025. URL https://arxiv.org/abs/2511.22699.

[100] Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. Longcat-image technical report, 2025. URL https://arxiv.org/abs/2512.07584.

[101] Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing, 2026. URL https://arxiv.org/abs/2603.09877.

[102] Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, Naila Murray, Marjan Ghazvininejad, Mike Lewis, Nicolas Ballas, Amir Bar, Michael Rabbat, Jakob Verbeek, Luke Zettlemoyer, Koustuv Sinha, Yann LeCun, and Saining Xie. Beyond language modeling: An exploration of multimodal pretraining, 2026. URL https://arxiv.org/abs/2603.03276.

[103] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders, 2026. URL https://arxiv.org/abs/2601.16208.

[104] 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, 2023. URL https://arxiv.org/abs/2311.12908.

[105] Dianyi Wang, Ruihang Li, Feng Han, Chaofan Ma, Wei Song, Siyuan Wang, Yibin Wang, Yi Xin, Hongjian Liu, Zhixiong Zhang, Shengyuan Ding, Tianhang Wang, Zhenglin Cheng, Tao Lin, Cheng Jin, Kaicheng Yu, Jingjing Chen, Wenjie Wang, Zhongyu Wei, and Jiaqi Wang. Deepgen 1.0: A lightweight unified multimodal model for advancing image generation and editing, 2026. URL https://arxiv.org/abs/2602.12205.

[106] Haozhe Wang, Cong Wei, Weiming Ren, Jiaming Liu, Fangzhen Lin, and Wenhu Chen. Rationalrewards: Reasoning rewards scale visual generation both training and test time, 2026. URL https://arxiv.org/abs/2604.11626.

[107] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

[108] Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation, 2025. URL https://arxiv.org/abs/2510.18701.

[109] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning, 2025. URL https://arxiv.org/abs/2508.20751.

[110] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation, 2026. URL https://arxiv.org/abs/2503.05236.

[111] Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models, 2026. URL https://arxiv.org/abs/2601.20354.

[112] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, 2022. URL https://arxiv.org/abs/2210.14896.

[113] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions?, 2025. URL https://arxiv.org/abs/2506.02161.

[114] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

[115] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2410.13848.

[116] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yan Zeng, and Weilin Huang. Rewarddance: Reward scaling in visual generation, 2025. URL https://arxiv.org/abs/2509.08826.

[117] 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, 2023. URL https://arxiv.org/abs/2306.09341.

[118] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference, 2023. URL https://arxiv.org/abs/2303.14420.

[119] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. URL https://arxiv.org/abs/2410.10629.

[120] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer, 2025. URL https://arxiv.org/abs/2501.18427.

[121] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2408.12528.

[122] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025. URL https://arxiv.org/abs/2506.15564.

[123] 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, 2023. URL https://arxiv.org/abs/2304.05977.

[124] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths, 2023. URL https://arxiv.org/abs/2305.18295.

[125] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025. URL https://arxiv.org/abs/2505.07818.

[126] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model, 2023. URL https://arxiv.org/abs/2311.13231.

[127] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference, 2024. URL https://arxiv.org/abs/2402.08265.

[128] Zhichao Yang, Tianjiao Gu, Jianjie Wang, Feiyu Lin, Xiangfei Sheng, Pengfei Chen, and Leida Li. Longt2ibench: A benchmark for evaluating long text-to-image generation with graph-structured annotations, 2025. URL https://arxiv.org/abs/2512.09271.

[129] Junchao Yi, Rui Zhao, Jiahao Tang, Weixian Lei, Linjie Li, Qisheng Su, Zhengyuan Yang, Lijuan Wang, Xiaofeng Zhu, and Alex Jinpeng Wang. Flowinone: Unifying multimodal generation as image-in, image-out flow matching, 2026. URL https://arxiv.org/abs/2604.06757.

[130] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. URL https://arxiv.org/abs/2206.10789.

[131] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation, 2024. URL https://arxiv.org/abs/2402.10210.

[132] Z.ai. GLM-Image, 2025. URL https://z.ai/blog/glm-image. Blog post; introduces the GLM-Image text-to-image model.

[133] Huichao Zhang, Liao Qu, Yiheng Liu, Hang Chen, Yangyang Song, Yongsheng Dong, Shikun Sun, Xian Li, Xu Wang, Yi Jiang, Hu Ye, Bo Chen, Yiming Gao, Peng Liu, Akide Liu, Zhipeng Yang, Qili Deng, Linjie Xing, Jiyang Liu, Zhao Wang, Yang Zhou, Mingcong Liu, Yi Zhang, Qian He, Xiwei Hu, Zhongqi Qi, Jie Shao, Zhiye Fu, Shuai Wang, Fangmin Chen, Xuezhi Chai, Zhihua Wu, Yitong Wang, Zehuan Yuan, Daniel K. Du, and Xinglong Wu. Nextflow: Unified sequential modeling activates multimodal understanding and generation, 2026. URL https://arxiv.org/abs/2601.02204.

[134] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation, 2024. URL https://arxiv.org/abs/2405.14705.

[135] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

[136] Ran Zhao, Sheng Jin, Size Wu, Kang Liao, Zerui Gong, Zujin Guo, Yang Xiao, and Wei Li. Knowledge visualization: A benchmark and method for knowledge-intensive text-to-image generation, 2026. URL https://arxiv.org/abs/2604.22302.

[137] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process, 2026. URL https://arxiv.org/abs/2509.16117.

[138] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. URL https://arxiv.org/abs/2408.11039.

[139] Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. Multimodal llms as customized reward models for text-to-image generation, 2025. URL https://arxiv.org/abs/2507.21391.

[140] Yucheng Zhou, Jiahao Yuan, and Qianning Wang. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation, 2025. URL https://arxiv.org/abs/2505.24787.

[141] Jie Zhu, Hanghang Ma, Jia Wang, Yayong Guan, Yanbing Zeng, Lishuai Gao, Junqiang Wu, Jie Hu, and Leye Wang. A simple baseline for unifying understanding, generation, and editing via vanilla next-token prediction, 2026. URL https://arxiv.org/abs/2603.04980.

[142] Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Diffusion reinforcement learning via centered reward distillation, 2026. URL https://arxiv.org/abs/2603.14128.

Appendix

This appendix provides supplementary analyses and implementation details supporting the main paper:

- Sec. A extends the condensed Sec. 2 with a comprehensive survey of T2I generation, compositional and visual-appeal evaluation, and RL post-training.
- Sec. B details the construction pipeline for the *Imago*-Short corpus and the four strategy-specific long-prompt expansions, including the LLM templates and the auditor protocol.
- Sec. C specifies the optimization pipeline and RL fine-tuning hyperparameters that produce *Imago*-4B and *Imago*-6B.
- Sec. D documents the generation, judge, and visual-appeal scoring protocols used for every score reported in the main paper.
- Sec. E provides qualitative comparisons of *Imago* against its base models, covering compositional fidelity across prompt regimes and visual appeal across diverse artistic styles.

A Related Work

A.1 Text-to-Image Generation

Modern text-to-image (T2I) generation has evolved along two primary directions: dedicated image generation models and unified multimodal systems. Early dedicated models synthesized images autoregressively as discrete token sequences [21; 25; 89; 130]. Diffusion and flow-matching frameworks subsequently emerged as the dominant paradigm, driven by text conditioning from language models [78; 90; 93], latent-space denoising [31; 91], and scalable Transformer backbones [15; 80]. Contemporary open-weight models, including the Stable Diffusion [5; 22; 81; 124] and FLUX [7; 8; 53] families alongside other optimized architectures [10; 86; 99; 100; 103; 114; 119; 120], have consistently improved generation quality and efficiency.

In parallel, unified multimodal models integrate visual understanding and image generation within a single architecture. Representative approaches extend next-token prediction to interleaved image-text sequences [12; 107], combine diffusion with autoregressive prediction in a shared Transformer [138], or decouple visual representations for understanding and generation [14; 16; 75; 115; 121; 122]. A wide range of recent models continues to expand this unified design space [20; 42; 58; 64; 70; 101; 102; 105; 129; 133; 141].

Together, both paradigms equip modern generators with strong visual representations. Building on these pretrained foundations, our work focuses on reinforcement learning post-training. Rather than altering model architectures or scaling pretraining, we study how reinforcement learning recipes can systematically enhance compositional capability across different prompt lengths without compromising visual appeal.

A.2 Evaluation and Benchmarking

Evaluating text-to-image generation requires assessing both *compositional capability*, which measures semantic alignment with the prompt, and *visual appeal*, which reflects overall perceptual quality.

Early evaluation metrics such as FID [35] and CLIPScore [34], alongside holistic suites like DrawBench [93; 130] and HEIM [4; 55], capture aggregate performance but obscure fine-grained semantic errors [33; 46; 92]. Subsequent benchmarks introduce targeted prompts to assess compositional capability directly. Initial suites such as GenEval [27] focus on basic object presence and attribute binding [17; 18; 40; 41; 95], whereas recent benchmarks evaluate longer and more complex textual descriptions [9; 56; 59; 67; 79; 85; 108; 111; 113; 128; 136; 140]. However, widely used benchmarks such as GenEval and DPG-Bench [39] have begun to saturate on modern models, limiting their ability to distinguish performance differences [45]. This saturation motivates evaluating models across short, medium, and long prompt regimes that preserve sufficient headroom.

Visual appeal is typically evaluated using learned human preference models, including PickScore [49], HPS [76; 117; 118], ImageReward [123], and UnifiedReward [110]. These models differ in their training supervision, leveraging pairwise human comparisons, expert annotations, multidimensional criteria [134], or unified multimodal objectives. While these models serve as convenient proxies for perceptual preference, using them to evaluate reinforcement learning post-training introduces evaluation bias. When an evaluation metric strongly correlates with the optimization reward, the assessment favors models trained on similar objectives, which obscures reward hacking and inflates measured improvements. Consequently, a reliable evaluation protocol requires selecting a held-out visual preference proxy alongside compositional benchmarks.

A.3 RL for Text-to-Image Generation

An RL recipe for T2I generation requires three core ingredients: the prompt corpus defining the environment, the reward supplying the optimization signal, and the algorithm translating feedback into parameter updates. We organize our review around these three pillars.

Training Corpus. T2I development historically relies on massive image-text and preference datasets [49; 65; 74; 94; 96; 112; 117]. In RL post-training, however, the corpus serves a fundamentally different purpose: it dictates the environment where the policy explores and receives feedback. Existing methods treat corpus selection as incidental, typically sampling from heterogeneous, uncontrolled prompt sources [6; 23; 57; 62; 68; 125]. Consequently, the impact of corpus composition on RL behavior remains underexplored. We elevate the prompt corpus to a primary design variable, investigating how its design choices dictate downstream compositional capabilities.

Reward. Reward modeling for T2I generation is currently dominated by the human-preference paradigm [49; 76; 110; 116–118; 123; 139], occasionally augmented by structured or intrinsic variants [3; 13; 24; 28; 47; 54; 60; 65; 69; 106]. However, optimizing against these learned preferences within an RL loop is notoriously fragile. When the optimization signal imperfectly aligns with true generative quality, the policy rapidly exploits this gap, resulting in reward hacking and aesthetic over-fitting [1; 26; 37; 109]. Rather than proposing a novel evaluator, our work empirically audits this vulnerability. Specifically, we investigate which established reward paradigm provides the most robust and powerful signal for compositional RL.

Algorithm. Algorithmic advancements in T2I RL have progressed in three stages. Early online methods adapted standard policy gradients to the diffusion denoising process [6; 23]. Subsequently, reward-gradient techniques (e.g., AlignProp [82], DRaFT [19]) and preference-optimization variants (e.g., Diffusion-DPO [32; 61; 66; 104; 126; 127; 131]) emerged to reduce variance and eliminate online sampling overhead. The contemporary focus lies in stabilizing and scaling optimization for modern flow-matching and masked generative architectures, exemplified by Flow-GRPO [68] and DiffusionNFT [137], alongside numerous refinements [57; 62; 63; 72; 73; 77; 87; 98; 125; 142]. Across these stages, the underlying reward and training corpus are routinely held fixed to report algorithmic gains. Our work takes a complementary approach: we leverage established RL machinery as a fixed, controlled lens to isolate and study the critical impact of the reward signal and the prompt environment.

B Dataset Construction Details

This section details the construction pipeline for *Imago*-Short and *Imago*-Long, introduced in Sec. 5. Two core principles guide this pipeline. First, every prompt must be visually grounded: prompts that are syntactically well-formed but semantically incoherent introduce optimization noise during reinforcement learning. Second, the training distribution must maintain balance across the three compositional axes: entities, attributes, and relations, preventing the generative model from developing biases toward any single category. We implement this design through a two-stage pipeline executed on a single server equipped with eight B200 GPUs.

B.1 Curation of *Imago*-Short

Raw prompt pool initialization. We initialize the candidate pool from Fine-T2I [74], which aggregates approximately 3M text-to-image prompts with a mean length of roughly 30 words. While this large-scale collection provides broad lexical coverage, it contains considerable noise, including

mid-sentence truncations, encoding artifacts, and non-descriptive boilerplate. We treat this raw pool as a high-recall foundation and enforce precision through subsequent filtering stages.

Embedding-based diversity selection. To eliminate redundant descriptive patterns while preserving compositional breadth, we perform diversity selection in the embedding space. We encode each prompt using Qwen3-Embedding-8B [135] and normalize the resulting representations under the ℓ_2 norm. We construct a cosine-similarity graph $G_\theta = (V, E_\theta)$, where an edge $(u, v) \in E_\theta$ exists if and only if $\langle e_u, e_v \rangle \geq \theta$, and extract a subset using a random-order greedy maximum independent set algorithm. We calibrate the threshold θ by binary search over the pairwise similarity distribution of a 500K-prompt sample to reach a target retention size. This step reduces the candidate pool to approximately 200K prompts with an average pairwise cosine similarity below 0.6, removing near-duplicate prompts while retaining long-tail compositional structures.

Semantic integrity and grounding filter. Because dense text embeddings can represent malformed text without signaling syntactic collapse, we evaluate prompt validity using Qwen3.5-27B [88] as a binary judge. The judge discards prompts that are truncated mid-sentence, contain gibberish or encoding errors, or lack an identifiable visual subject. Prompts default to valid in the rare event of API exceptions, affecting fewer than ten instances across the entire corpus.

Balancing across compositional axes. To ensure comprehensive compositional coverage, the training corpus must expose the generative model equally to entities, attributes, and relations. We employ Qwen3.5-27B to classify each surviving prompt into its primary compositional axis: entity, attribute, relation, or balanced when no single axis dominates. We stream-filter the classified candidates until the representation across all four categories reaches parity. This yields the final $151{,}232$ prompts of *Imago*-Short, which serves as a standalone short-prompt training corpus and as the seed distribution for long-prompt expansion.

B.2 Strategy-Controlled Long-Prompt Expansion

Multi-strategy prompt expansion. Prompts in *Imago*-Short average roughly 30 words, providing limited stress on complex relational configurations. We therefore expand each short seed into four parallel 200-word variants using Qwen3.5-27B [88], following the four compositional strategies defined in Sec. 5.1: Balanced, Entity-centric, Attribute-centric, and Relation-centric. All four templates share an identical preamble, JSON output format, target word count, and instructions excluding non-diegetic stylistic descriptors. The templates differ exclusively in their distribution requirements across the three compositional axes. The Balanced template is shown below; Tab. 6 specifies the operational guidelines for the other three strategies.

Balanced expansion template

You are an expert prompt engineer and visual semanticist. Take a brief base text-to-image prompt and significantly increase its visual complexity by expanding the scene description. Use only three semantic categories: *entities*, the core base nouns including rendered text; *attributes*, verifiable visually observable properties; and *relations*, objective physical, spatial, or structural interactions.

Strict guidelines.

1. *Target length.* A cohesive paragraph of approximately 200 words.

2. *Balanced complexity.* Distribute additions evenly; maintain a roughly equal ratio of distinct entities, attributes, and relations. Do not let any single category dominate the scene.

3. *Harmonious integration.* Forbid mechanical comma-separated lists. Every entity must be objectively qualified and physically anchored to the environment or to other entities via precise spatial prepositions.

4. *Grounded reality.* Avoid subjective adjectives and meta or style instructions; keep the text focused on objective, diegetic content.

Output strictly as JSON: {`"enhanced_prompt": <expanded text>`}.

INPUT PROMPT: {`base prompt`}

Auditor verification and iterative regeneration. Unconstrained language model generation can drift from the specified strategy: entity-focused prompts may accumulate extraneous attributes, and relation-focused prompts may omit structural interactions. To enforce fidelity to each strategy, we employ a Qwen3.5-27B auditor. The auditor evaluates each generated candidate against three explicit criteria: the exclusion of non-diegetic quality descriptors, the absence of comma-separated keyword enumerations, and strict alignment with the targeted distribution of entities, attributes, and

Table 6: **Strategy-specific guidelines for non-Balanced expansion templates.** All other template components, including length constraints, diegetic grounding rules, and JSON output formatting, are identical to the Balanced template.

Strategy	Strategy Guidelines
Entity-centric	*2. Entity focus.* Base nouns and rendered text elements must significantly outnumber attributes and relations to generate a densely populated scene. *3. Structured composition.* Organize entities through clear compositional hierarchies, avoiding unstructured object enumerations.
Attribute-centric	*2. Attribute focus.* Verifiable physical properties must significantly outnumber distinct entities and relations. Maintain a low entity count and qualify each item across multiple visual dimensions. *3. Detailed specification.* Specify the visual, material, and surface identity of each entity to yield a highly detailed yet sparsely populated scene.
Relation-centric	*2. Relation focus.* Spatial prepositions, physical contact, and structural arrangements must significantly outnumber entities and attributes. *3. Structural configuration.* Detail how objects physically interact and occupy space, emphasizing support relationships, multi-layered spatial embeddings, and geometric layouts.

relations. Candidates failing any criterion are regenerated iteratively until they satisfy all checks or exhaust a preset retry limit. This procedure yields four strategy-specific expanded corpora (Balanced, Entity-centric, Attribute-centric, and Relation-centric). As established in Sec. 5.2, the balanced variant achieves the strongest overall performance and is designated *Imago*-Long, which we adopt for the curriculum study (Sec. 5.3) and the final recipe.

C Training Details

This section details the reinforcement learning fine-tuning configurations used throughout our experiments. The primary configuration produces *Imago*-4B and *Imago*-6B, initialized from FLUX.2-Klein-Base [8] and Z-Image [99], respectively. The ablation configuration, applied in Sec. 4.2, Sec. 5.2, and Sec. 5.3, adopts the same formulation with scaled-down capacity as specified below. All models are trained on a single node equipped with eight B200 GPUs.

C.1 Optimization Pipeline and Objectives

Each training step proceeds through three successive stages: rollout sampling, reward assignment, and parameter optimization.

Stage 1: Rollout sampling. Candidate images are generated using an Exponential Moving Average (EMA) copy of the generator parameters, θ^{old}. For each prompt c, the EMA generator produces a group of G candidate images $\{x_0^k\}_{k=1}^G$ over T_{sample} discrete inference steps without classifier-free guidance. Training relies strictly on these generated candidates without external ground-truth images.

Stage 2: Reward assignment and normalization. Each generated image x_0^k receives a scalar reward $r^{\text{raw}}(x_0^k, c)$ evaluated via zero-shot VQAScore from Qwen3.5-9B (Sec. 4), which serves as the sole optimization signal. To compute relative advantages within each group, raw scores are normalized into an optimality probability $r_k \in [0, 1]$:

$$r_k = \frac{1}{2} + \frac{1}{2\,c_{\text{clip}}}\ \text{clip}\left(\frac{r^{\text{raw}}(x_0^k, c) - \mu_R}{\sigma_R + \epsilon_{\text{norm}}}, -c_{\text{clip}}, c_{\text{clip}}\right), \quad c_{\text{clip}} = 5 \tag{1}$$

where $\mu_R = \frac{1}{G}\sum_{j=1}^{G} r^{\text{raw}}(x_0^j, c)$, σ_R is the group standard deviation, and $\epsilon_{\text{norm}} = 10^{-8}$ ensures numerical stability.

Stage 3: Parameter optimization and EMA update. For each generated sample x_0^k, we discretely draw $N_{\text{train}} = 2$ training timesteps from the inference discretization grid, restricted to the high-noise interval $t \in [0.5, 1.0]$. In the rectified flow framework, the forward noising process at continuous

timestep t yields the perturbed state x_t and ground-truth velocity target v:

$$x_t = (1-t)x_0^k + t\epsilon, \quad v = \epsilon - x_0^k, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

During parameter optimization, the active generator velocity $v_\theta(x_t, c, t)$ is evaluated alongside the baseline prediction $v^{\text{old}}(x_t, c, t)$, which is computed by passing the perturbed state x_t through the EMA weights θ^{old} at the sampled timesteps. Using an extrapolation factor β_{NFT}, the objective constructs implicit positive and negative velocity fields:

$$v_\theta^+(x_t, c, t) = (1 - \beta_{\text{NFT}})v^{\text{old}}(x_t, c, t) + \beta_{\text{NFT}} v_\theta(x_t, c, t) \tag{3}$$

$$v_\theta^-(x_t, c, t) = (1 + \beta_{\text{NFT}})v^{\text{old}}(x_t, c, t) - \beta_{\text{NFT}} v_\theta(x_t, c, t) \tag{4}$$

The generator weights θ are updated by minimizing the contrastive flow matching loss:

$$\mathcal{L}_{\text{NFT}}(\theta) = \mathbb{E}_{c, x_0^k, t, \epsilon}\left[\frac{r_k}{w_k^+}\|v_\theta^+(x_t, c, t) - v\|_2^2 + \frac{1 - r_k}{w_k^-}\|v_\theta^-(x_t, c, t) - v\|_2^2\right] \tag{5}$$

$$\text{where} \quad w_k^\pm = \text{sg}\left[\overline{\left|v_\theta^\pm(x_t, c, t) - v\right|}\right] \tag{6}$$

Here $\text{sg}[\cdot]$ denotes the stop-gradient operator and the overline denotes the spatial mean. Normalizing each residual by its stop-gradient mean absolute value produces bounded, approximately L_1 gradient dynamics following DiffusionNFT [137].

Each epoch performs a single optimizer update: gradients are accumulated across the entire rollout ($P_{\text{step}} \times G$ generated images, each supervised at N_{train} timesteps) before executing the optimization step, and collected samples are processed in a single inner epoch without reuse. Throughout training, we do not employ any explicit KL penalty. Following the optimizer update, the sampling generator weights undergo a soft update:

$$\theta^{\text{old}} \leftarrow \eta_i \theta^{\text{old}} + (1 - \eta_i)\theta \tag{7}$$

where $\eta_i = \min(0.001 \cdot i, \eta_{\max})$ with $\eta_{\max} = 0.5$. This soft update stabilizes the rollout distribution while continuously incorporating recent parameter updates.

C.2 Parameter Adaptation and Curriculum

We apply Low-Rank Adaptation (LoRA) [38] exclusively to the DiT backbone, targeting all attention and feed-forward linear layers while freezing the text encoder and VAE modules. Optimization is conducted using AdamW [71] with a constant learning rate η, maintaining fixed optimizer parameters without warmup or decay schedules.

The sequential short to long curriculum transitions the generator from *Imago*-Short to *Imago*-Long (Sec. 5.3). At the transition between phases, the trained LoRA weights from Phase 1 serve as the initialization for Phase 2, while optimizer states, EMA generator weights, and group-level reward statistics are reinitialized.

For ablation experiments, adapter capacity is reduced to rank $r = 64$ and scaling factor $\alpha = 128$. To ensure fair comparison, every ablation run uses the same total budget of 500 update steps for FLUX.2-Klein-Base-4B and 300 for Z-Image: single-corpus and Joint runs consume this budget in a single stage, whereas the Sequential schedule splits it evenly across its two phases ($250{+}250$ and $150{+}150$, respectively). All remaining optimization and loss configurations are held fixed.

C.3 Hyperparameter Specification

Tab. 7 summarizes the complete hyperparameter configurations for *Imago*-4B and *Imago*-6B. While core reinforcement learning hyperparameters are shared across both scales, specific training settings reflect the underlying model architectures: *Imago*-6B utilizes $T_{\text{sample}} = 30$ rollout steps to accommodate the Z-Image sampling schedule, and adopts a higher learning rate of $\eta = 3 \times 10^{-4}$ to ensure effective gradient scaling on the larger backbone. Micro-batch sizes (B_{train} and B_{sample}) are allocated to balance GPU memory utilization during rollout and backward passes while maintaining a constant group size $G = 12$. Because the final models follow the two-phase short-to-long curriculum (Sec. 5.3), this per-phase budget yields 1000 and 600 total updates for *Imago*-4B and *Imago*-6B, respectively, whereas the corpus ablations fix the total budget at $500/300$.

Table 7: **Training configurations for *Imago*-4B and *Imago*-6B.** Parameter symbols correspond to the mathematical formulations in Appendix C.

Component	Symbol	Hyperparameter Description	*Imago*-4B	*Imago*-6B
Model	M_{base}	Base generator architecture	FLUX.2-Klein-Base	Z-Image
	N_{param}	Backbone parameter count	4B	6B
LoRA	r	LoRA rank	256	256
	α	LoRA scaling factor	512	512
Optimization	N_{step}	Parameter updates per phase	500	300
	η	Constant learning rate	1×10^{-4}	3×10^{-4}
	β_1, β_2	AdamW momentum coefficients	0.9, 0.999	0.9, 0.999
	ϵ_{Adam}	AdamW epsilon	10^{-8}	10^{-8}
	λ_{wd}	Weight decay coefficient	10^{-4}	10^{-4}
	c_{grad}	Gradient clipping threshold	1.0	1.0
	B_{train}	Training batch size per GPU	16	6
	B_{sample}	Sampling batch size per GPU	16	36
	P_{step}	Unique prompts per update	96	48
	N_{accum}	Gradient accumulation steps	18	24
DiffusionNFT	G	Group size per prompt	12	12
	β_{NFT}	Target extrapolation factor	10.0	10.0
	c_{clip}	Advantage clipping threshold	5.0	5.0
	$\eta_{\max}$	Maximum EMA decay factor	0.5	0.5
Sampling	T_{sample}	Training rollout steps	8	30
	$H \times W$	Training image resolution	512×512	512×512
	N_{train}	Training timesteps per sample	2	2
	$[t_{lo}, t_{hi}]$	Training timestep interval	[0.5, 1.0]	[0.5, 1.0]

D Evaluation Details

This section details the evaluation protocols for all quantitative results reported in the paper. Model performance is assessed across two primary dimensions: compositional capability across five diagnostic benchmarks, evaluated using multimodal language model judges, and visual appeal on the DrawBench prompt suite, evaluated using PickScore. All models are tested within a unified execution runtime to ensure that comparisons isolate generative capability rather than implementation discrepancies.

D.1 Generation Protocols and Model Configurations

All models are evaluated using their official default generation settings. Tab. 8 summarizes the resolution, denoising step count, and classifier-free guidance (CFG) scale for the 19 evaluated text-to-image models. *Imago*-4B inherits the evaluation configuration of FLUX.2-Klein-Base-4B, while *Imago*-6B inherits that of Z-Image.

Several models incorporate specialized guidance or normalization mechanisms during sampling. A CFG scale of 0.0 corresponds to sampling without guidance, as implemented in FLUX.1-schnell and Z-Image-Turbo. BAGEL applies distinct guidance scales for text (4.0) and image conditioning (1.5), restricted to the normalized timestep interval $[0.4, 1.0]$ with global guidance renormalization. Lumina-Image-2 applies CFG normalization with a truncation ratio of 0.25, LongCat-Image utilizes guidance renormalization, and Qwen-Image applies true classifier-free guidance at a scale of 4.0.

D.2 Compositional Capability Evaluation

The five compositional benchmarks introduced in Sec. 3 are evaluated using Qwen3.5-27B [88] deployed via vLLM [52] with chain-of-thought verification enabled. For SpatialGenEval, the judge evaluates each generated scene across five independent passes and determines the final score via majority voting. For T2I-CoReBench, evaluation is strictly restricted to its composition split, computed as the unweighted mean across four subtasks: multiple instances (C-MI), multiple attributes

Table 8: **Inference configurations for the 19 evaluated text-to-image models.** Denoising steps and classifier-free guidance scales follow official defaults across all benchmarks.

Model	Resolution	Steps	CFG	Model	Resolution	Steps	CFG
SDXL	1024×1024	30	5.0	Z-Image	1024×1024	50	4.0
SD-3.5-Medium	1024×1024	40	4.5	Z-Image-Turbo	1024×1024	9	0.0
SD-3.5-Large	1024×1024	28	3.5	BAGEL	1024×1024	50	Text 4.0, Image 1.5
FLUX.1-schnell	1024×1024	4	0.0	DeepGen	512×512	50	4.0
FLUX.1-dev	1024×1024	50	3.5	GLM-Image	1024×1024	50	1.5
FLUX.2-dev	1024×1024	50	4.0	LongCat-Image	1024×1024	50	4.0
FLUX.2-Klein-4B	1024×1024	4	1.0	Lumina-Image-2	1024×1024	50	4.0
FLUX.2-Klein-9B	1024×1024	4	1.0	Qwen-Image	1024×1024	50	4.0
FLUX.2-Klein-Base-4B	1024×1024	50	4.0	SANA	1024×1024	20	4.5
FLUX.2-Klein-Base-9B	1024×1024	50	4.0				

(C-MA), multiple relations (C-MR), and text rendering (C-TR). The reasoning split is excluded throughout our study to isolate compositional fidelity from open-domain world knowledge.

D.3 Visual Appeal Assessment

Visual appeal is measured by computing the average PickScore [49] over the DrawBench prompt collection [93], generating one image per prompt using single-precision floating-point inference. As established in Sec. 3, PickScore has the lowest average pairwise correlation among the four preference evaluators considered and is therefore used as a held-out proxy for visual preference. This reduces direct reuse of the optimization metric but does not establish independence from the training reward or eliminate evaluation bias; the reported gains should be read as improvements on this proxy rather than a fully independent verification of perceptual quality.

E Qualitative Results

This section provides visual comparisons illustrating the improvements achieved by *Imago* over its base models across compositional fidelity and visual appeal.

Compositional fidelity across prompt regimes. Fig. 8 presents qualitative comparisons across short, medium, and long prompt benchmarks. Across all three regimes, *Imago* systematically corrects semantic errors present in the base models:

- *Short prompts (T2I-CompBench):* The base models frequently omit secondary entities or violate spatial relations, such as missing the square toast or incorrectly positioning the painting relative to the horse. *Imago* consistently recovers the missing objects and respects the specified spatial arrangements.
- *Medium prompts (T2I-CoReBench):* When prompts require simultaneous attribute binding and text rendering, *Imago* accurately renders typographic content (e.g., correctly spelling `Galactic Heroes` on the cinema ticket) and places visual elements in their specified relative locations.
- *Long prompts (LongBench-T2I):* Under dense textual descriptions, the base models struggle to retain all requested scene elements. *Imago* demonstrates higher recall over multi-entity configurations, preserving fine-grained descriptive elements without dropping background entities.

In Fig. 8, red bounding boxes highlight specific semantic failures in the base generations that are resolved by *Imago*.

Visual appeal across diverse artistic styles. Figs. 9 and 10 show that our reinforcement learning recipe, *Imago*, improves visual appeal while preserving the stylistic diversity of the base models across diverse domains:

- Imago-*4B vs. FLUX.2-Klein-Base (Fig. 9):* Across portraits, watercolor landscapes, textured paintings, 3D character renders, and atmospheric lighting scenes, *Imago*-4B enhances surface textures, local contrast, and lighting fidelity without altering the intended artistic genre.



Figure 8: **Qualitative comparison of compositional fidelity across prompt lengths.** Each row pairs a prompt with outputs from the base model and its fine-tuned *Imago* counterpart. Red bounding boxes highlight semantic discrepancies, including missing objects, attribute binding errors, and text layout mistakes in the base model that are corrected by *Imago*.



Figure 9: **Visual appeal comparison between *Imago*-4B and FLUX.2-Klein-Base.** Across diverse genres including portraits, watercolor landscapes, textured paintings, and volumetric lighting scenes, *Imago*-4B improves textural details, lighting coherence, and surface definition without compromising stylistic diversity.

- Imago-*6B vs. Z-Image (Fig. 10):* Across macro photography, 3D clay models, papercraft compositions, landscapes, and graphic posters, *Imago*-6B improves geometric definition, edge clarity, and material rendering.

These comparisons confirm that *Imago* strengthens prompt fidelity across prompt lengths while improving perceptual quality across diverse visual styles.



Figure 10: **Visual appeal comparison between *Imago*-6B and Z-Image.** Across stylized illustrations, papercraft art, nature photography, and graphic compositions, *Imago*-6B improves geometric definition, material realism, and edge clarity relative to the base model.

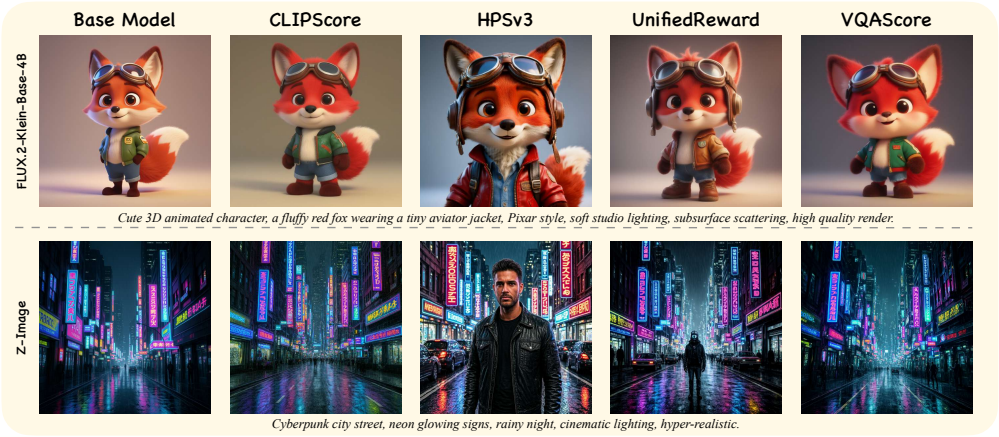


Figure 11: **Qualitative comparison of visual appeal and prompt fidelity.** Preference-trained rewards such as HPSv3 can exhibit content drift and potential reward overoptimization, introducing unprompted subjects (e.g., the person in the bottom row) that raise aesthetic scores at the expense of prompt fidelity, whereas VQAScore preserves prompt fidelity while improving overall visual quality.

You are an expert image evaluator. Please assess how accurately the provided image matches the text prompt. Apply maximum strictness. Score the alignment on a scale from 1 to 10, where 1 means completely mismatched and 10 means perfectly matched. Output your response strictly in the following JSON format:

```
{
"score": <integer_between_1_and_10>
}
```

Figure 12: **VQAScore prompt template.** The graded 1-to-10 semantic-alignment prompt used to instantiate VQAScore in Sec. 4.1.