Improving Text-to-Image Generation with Intrinsic Self-Confidence Rewards

Seungwook Kim[1] Minsu Cho[1,2]
[1]Pohang University of Science and Technology (POSTECH) [2]RLWRLD
https://wookiekim.github.io/SOLACE/

Figure 1. **Qualitative examples of SOLACE on Pick-a-Pic dataset [35].** Best viewed on electronics.

Abstract

*Text-to-image generation powers content creation across design, media, and data augmentation. Post-training of text-to-image generative models is a promising path to improve human preference alignment, factuality, and aesthetics. We introduce SOLACE (**S**elf-**O**riginating **LA**tent **C**onfidence **E**stimation), a post-training framework that replaces external reward supervision with an internal* self-confidence *signal: we re-noise the model's own outputs and measure how accurately it recovers the injected noise, treating low reconstruction error as high self-confidence. SOLACE converts this intrinsic signal into scalar rewards for reinforcement learning, requiring no external reward models, annotators, or preference data. By reinforcing high-confidence generations, SOLACE delivers consistent gains in compositional generation, text rendering, and text-image alignment. Integrating SOLACE with external rewards yields complementary improvements while alleviating reward hacking.*

1. Introduction

Text-to-image (T2I) generation has advanced rapidly with the rise of diffusion and flow-based models, delivering high-fidelity, diverse images from natural language prompts [9, 10, 16, 22, 51, 53, 60, 61]. These models now support a broad range of applications: controllable image editing and inpainting [3, 6, 7, 65, 67, 81, 92]; serving as powerful priors or pre-trained components for text-to-video diffusion models [24, 26, 33, 36, 75, 76, 97]; data creation and augmentation pipelines for downstream perception tasks [69, 78, 85]; and text-to-3D (and 4D) reconstruction via score distillation sampling [1, 31, 32, 34, 55, 66, 70, 79]. Recent studies show that *post-training text-to-image generative models* via reinforcement learning can yield dramatic improvements in visual appeal and aesthetic quality [5, 41, 74], typically by optimizing external rewards derived from human preference models [35, 80, 83] or task-specific validators [14, 21].

However, defining a scalable and reliable reward for

"good" images remains challenging [35, 38, 72, 80, 83]. There are numerous, weakly-aligned criteria a good image has to satisfy, *e.g.,* compositionality, text rendering, aesthetics, and text–image alignment, whose relative importance shifts across domains and prompts [38]. In practice, external-reward post-training is also vulnerable to over-optimization: optimizing a narrow critic can induce reward hacking and regressions on non-target capabilities, degrading coverage or faithfulness even as the targeted score rises [5, 41, 74]. Human-preference based reward models [35, 77, 83] are popular for their efficacy, but require large-scale annotation for training. Operationally, external rewards require running additional evaluators (preference/OCR/safety models) alongside the generator during training, increasing pipeline complexity.

Despite extensive progress in extrinsically supervised post-training, intrinsic signals remain under-explored for text-to-image generation. In this work, we ask: ***can internal feedback from the text-to-image generator itself provide meaningful signals for post-training?*** To this end, we introduce Self-Originating LAtent Confidence Estimation (SOLACE), a post-training framework that uses the model's own *self-confidence* as a reward. Inspired by Score Distillation Sampling [55, 66, 79], which uses a pretrained text-to-image generator as a critic for text-to-3D or -4D generation, we propose to let a text-to-image generator *critique its own generation*. Concretely, given a sampled latent z_0, we re-noise it to selected timesteps $t \in \mathcal{T}$ using the forward noising schedule, and measure how well the model recovers the injected noise. Low reconstruction error indicates high self-confidence. Our hypothesis is that large-scale pretraining endows diffusion models with strong priors over real images and text-image correspondence, so self-confidence should correlate with text alignment and realism.

Empirically, SOLACE yields consistent gains in compositional generation [21], text rendering [14], and text-image alignment [58], while modestly improving human-preference scores [35, 77, 80, 83], all without external rewards. Qualitative comparisons and a user study corroborate these trends, indicating that intrinsic self-confidence aligns with key aspects of image generation quality. Moreover, applying SOLACE on top of an *extrinsically post-trained* model (*i.e*. one already fine-tuned with external rewards) yields further improvements in compositionality, text rendering, and alignment, with only slight drops on the targeted external metric. This shows that intrinsic and extrinsic rewards are complementary, and that SOLACE alleviates the reward hacking commonly observed in external-reward post-training.

The key contributions of our work are as follows:

- We present **SOLACE** (**S**elf-**O**riginating **LA**tent **C**onfidence **E**stimation), a post-training framework using *self-confidence* as reward.
- We define self-confidence as the model's ability to recover noise injected into its own outputs: we re-noise the generated latent, measure reconstruction error, and convert it into a scalar reward for GRPO post-training.
- Across standard benchmarks and a comprehensive user study, SOLACE yields consistent gains in compositionality, text rendering, and text–image alignment, while modestly improving human-preference metrics.
- SOLACE complements *external*-reward pipelines: applying SOLACE on top of externally post-trained models improves non-target capabilities (compositionality, text rendering, alignment) with only mild trade-offs on the targeted external metric, mitigating reward hacking.

2. Related Work

Text-to-image generative models. Text-to-image generation is a rapidly advancing field, which was initially dominated by diffusion models [2, 9, 10, 51, 53, 60, 61]. Recent work increasingly adopts flow matching [3, 16, 68] and sequence models [8, 46, 73, 86] for improved efficiency and generation quality. Advances span architectures [3, 16, 51], image recaptioning [4, 9, 10], and tokenization [30, 73, 87]. In this work, we focus on reinforcement-learning based post-training to improve text-to-image models, using the self-confidence of the generative model as the *intrinsic* reward.

Text-to-image model alignment via post-training. Post-training is emerging as an effective paradigm to align existing text-to-image models toward desired objectives, *e.g.,* human preference. This can take the form of direct fine-tuning given differentiable rewards [13, 56, 57, 83] or Reward Weighted Regression (RWR) [15, 17, 37, 52]. Some schemes build on reinforcement learning to leverage PPO [63]-style policy gradients [5, 18, 25, 48, 94], or perform Direct Preference Optimization (DPO) or its variants [19, 39, 42, 43, 59, 74, 84, 88, 91]. More recently, Flow-GRPO [41] introduces GRPO [64] for flow matching models, by converting the ODE of flow matching sampling to SDEs to inject stochasticity. However, external rewards increase training costs (an additional model must run alongside the generator) and raise the risk of reward hacking [42, 43, 59, 74]. In this work, we define self-confidence as the model's ability to recover noise injected into its own outputs, and use this *intrinsic* signal for post-training, improving compositional generation, text rendering, and text-image alignment without reward hacking.

Intrinsic signals for post-training. Intrinsic signals for post-training have recently gained traction in language modeling as scalable alternatives to human-labeled preference data, leveraging self-derived feedback such as confidence/uncertainty estimates, self-evaluation, and self-consistency to guide reinforcement learning or preference

optimization without annotators [11, 12, 54, 82, 89, 90, 93, 95, 98]. Recently, Intuitor [95] showed that using self-certainty as a confidence-based intrinsic reward enables single-agent reinforcement learning across diverse tasks without relying on explicit feedback, gold labels, or environment-based validation. Bringing the same principle to text-to-image generation is non-trivial: generation proceeds along continuous denoising trajectories and likelihoods are implicit, unlike token-level discrete objectives in LLMs. In this work, we define self-confidence of flow-matching models as their ability to recover noise injected into their own outputs, inspired by score-distillation sampling [55, 66]. This enables dense, on-policy feedback without labeled data or reward models. Empirically, we show that this self-confidence signal aligns with compositionality, text rendering, and text-image alignment.

3. Preliminary: GRPO for Flow Matching

3.1. Flow Matching and Rectified Flow

Flow matching bypasses score learning in conventional diffusion models [27, 71] by directly regressing the target velocity of a transport ODE along a user-chosen path between data and a reference distribution [40, 44]. Recent state-of-the-art generative models [3, 16, 36, 75] adopt the Rectified Flow (RF) framework. Specifically, let $x_0 \sim p_{\text{data}}$ and $x_1 \sim p_1$ (e.g., $\mathcal{N}(0, I)$); RF chooses the straight-line path

$$x_t = (1-t)\, x_0 + t\, x_1, \tag{1}$$

for which the target velocity is constant in t:

$$v^\star = \partial_t x_t = x_1 - x_0. \tag{2}$$

Training reduces to direct regression of this constant velocity at random (x_t, t) pairs:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0 \sim p_{\text{data}},\, x_1 \sim p_1,\, t \sim \mathcal{U}[0,1]} \left\| v^\star - v_\theta(x_t, t) \right\|_2^2. \tag{3}$$

After training, sampling solves the deterministic ODE

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t, t), \qquad t : 1 \to 0, \tag{4}$$

starting from $x_1 \sim p_1$ and transporting to x_0.

3.2. GRPO for Flow Matching

For a policy π_θ, we consider a policy-gradient objective that maximizes expected cumulative reward while regularizing updates toward a reference policy π_{ref} via a KL penalty:

$$\max_\theta \mathbb{E}_{(s_0, a_0, \ldots, s_T, a_T) \sim \pi_\theta} \Big[\sum_{t=0}^{T} R(s_t, a_t) - \beta \sum_{t=0}^{T} D_{\text{KL}}(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_{\text{ref}}(\cdot \mid s_t)) \Big], \tag{5}$$

where $R(s_t, a_t)$ is the per-step reward. Group Relative Policy Optimization (GRPO) [64] proposes to use a group relative formulation to estimate the advantage for each sample to optimize Eq. (5).

Flow-GRPO [41] integrates GRPO into flow matching models for online RL post-training. The iterative denoising process in flow matching can be formulated as a Markov Decision Process [5]: given a text prompt c, the flow model p_θ samples a group of G images $\{x_0^i\}_{i=1}^G$ and the corresponding sampling trajectories $\{(x_T^i, x_{T-1}^i, \cdots, x_0^i)\}_{i=1}^G$. The advantage of the i-th image is calculated by normalizing the group-level rewards:

$$\hat{A}_t^i = \frac{R(x_0^i, c) - \text{mean}(\{R(x_0^i, c)\}_{i=1}^G)}{\text{std}(\{R(x_0^i, c)\}_{i=1}^G)} \tag{6}$$

Finally, GRPO optimizes the policy model by maximizing $\mathcal{J}_{\text{Flow-GRPO}} = \mathbb{E}_{c \sim C, \{x^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|c)} f(r, \hat{A}, \theta, \epsilon, \beta)$, where

$$\begin{aligned} f(r, \widehat{A}, \theta, \epsilon, \beta) &= \underset{i,t}{\text{mean}} \Big[\min\big(r_t^i,\, \text{clip}_\epsilon(r_t^i)\big)\, \widehat{A}_t^i \Big] \;-\; \beta\, \overline{D}_{\text{KL}}, \\ \overline{D}_{\text{KL}} &= \underset{t}{\text{mean}}\, D_{\text{KL}}(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_{\text{ref}}(\cdot \mid s_t)), \\ \text{clip}_\epsilon(r) &\;\triangleq\; \text{clip}(r,\, 1-\epsilon,\, 1+\epsilon). \end{aligned} \tag{7}$$

and $r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i | x_T^i, c)}{p_{\theta_{\text{old}}}(x_{t-1}^i | x_T^i, c)}$. Flow-GRPO then converts the deterministic ODE of Eq. (4) into an equivalent SDE that matches the original model's marginal probability function at all timesteps, in order to meet the GRPO policy update requirements, *e.g.*, stochasticity is necessary for exploration in RL post-training. We adopt Flow-GRPO to post-train flow-matching text-to-image models.

4. Method: SOLACE

Overview. We present **SOLACE** (Self-Originating LAtent Confidence Estimation), a post-training method for text-to-image generators that requires no external reward models. SOLACE uses the model's own *self-confidence* as an intrinsic reward: after generating an output, we re-noise it at selected timesteps and measure how accurately the model recovers the injected noise. Aggregating these per-timestep recovery errors yields a single on-policy scalar reward for reinforcement learning. In the following, we detail the computation of the self-confidence reward (Sec. 4.1) and the stabilization techniques for SOLACE training (Sec. 4.2). An overview of SOLACE is shown in Fig. 2.

4.1. Intrinsic Self-Confidence Reward

Sampling a group of images for GRPO. Given a text prompt c, we sample G independent reverse trajectories in the latent space $\mathcal{Z}$ under the flow policy π_θ:

$$z_T^{(i)} \sim \mathcal{N}(0, I), z_{t-1}^{(i)} \sim \pi_\theta\Big(\cdot \mid z_t^{(i)},\, c\Big), i = 1, \ldots, G. \tag{8}$$

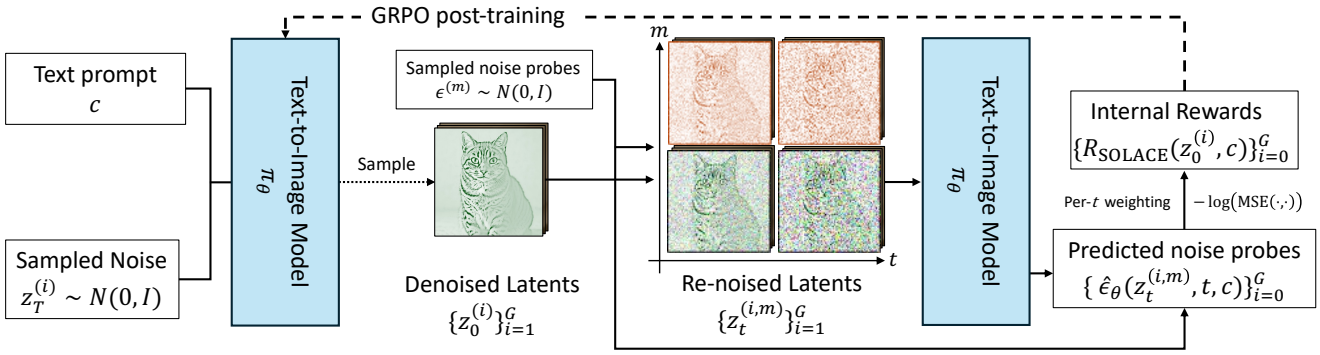

Figure 2. **Overview of SOLACE.** Given a text prompt c, we generate G different latents. Without decoding, we re-noise the latents using K noise probes across $t \in \mathcal{T} \subset [0, 1]$. For each generated latent $z_0^{(i)}$, we formulate the text-to-image generative model's self-confidence of the generated latent as the ability to denoise the re-noised latent. We leverage this self-confidence as an internal reward scalar value, which we use to post-train the text-to-image generative model using GRPO [41, 64]. We omit the KL term in this figure for better readability.

This produces terminal latents $\{z_0^{(i)}\}_{i=1}^G$ and trajectories $\{(z_T^{(i)}, z_{T-1}^{(i)}, \dots, z_0^{(i)})\}_{i=1}^G$. Using multiple independent draws yields the group required for group-relative advantage normalization in GRPO. While we can sample G different images from the same initial noise z_T due to the added stochasticity from [41], we sample different initial noise to improve exploration during GRPO training.

Sampling noise probes for re-noising. We draw a shared set of K noise probes in latent space:

$$\epsilon^{(m)} \sim \mathcal{N}(0, I), \qquad m = 1, \dots, K, \tag{9}$$

so that candidate i and candidate j are perturbed by the *same* probes $\{\epsilon^{(m)}\}_{m=1}^K$. For rectified flow, we re-noise a terminal latent $z_0^{(i)}$ via the linear forward kernel

$$z_t^{(i,m)} \;=\; (1-t)\, z_0^{(i)} \;+\; t\, \epsilon^{(m)}, \qquad t \in \mathcal{T} \subset [0, 1], \tag{10}$$

where $\mathcal{T}$ is the set of re-noising levels used for evaluation. We take K even ($K \geq 2$) and use antithetic pairing to enforce exact mean zero within the probe set, i.e., $\epsilon^{(m+K/2)} = -\epsilon^{(m)}$ for $m = 1, \dots, K/2$.

Calculating self-confidence. For each noised latent $z_t^{(i,m)}$ (Eq. (10)), we query the flow-matching model's velocity field $v_\theta(z_t^{(i,m)}, t, c)$. Under the rectified-flow parameterization, the velocity predicts a linear transform of the injected noise; specifically, we recover a noise estimate via

$$\widehat{\epsilon}_\theta\Big(z_t^{(i,m)}, t, c\Big) \;=\; v_\theta\Big(z_t^{(i,m)}, t, c\Big) \;+\; z_0^{(i)}. \tag{11}$$

We then measure the reconstruction error against $\epsilon^{(m)}$:

$$\text{MSE}_{i,t} \;=\; \frac{1}{K} \sum_{m=1}^{K} \Big\| \widehat{\epsilon}_\theta\Big(z_t^{(i,m)}, t, c\Big) - \epsilon^{(m)} \Big\|_2^2. \tag{12}$$

To turn small errors into large rewards while stabilizing dynamic range, we use the negative log transform,

$$S_{i,t} \;=\; -\log\big(\text{MSE}_{i,t} + \delta\big), \tag{13}$$

where $\delta > 0$ avoids $\log 0$. This choice (i) approximates a Gaussian log-likelihood score under an i.i.d. noise model, (ii) compresses outliers, and (iii) yields additive contributions across timesteps. Aggregating over a set of re-noising levels $\mathcal{T} \subset [0, 1]$ gives the scalar intrinsic reward

$$R_{\text{SOLACE}}\big(z_0^{(i)}, c\big) \;=\; \frac{1}{\sum_{t \in \mathcal{T}} w(t)} \sum_{t \in \mathcal{T}} w(t)\, S_{i,t}. \tag{14}$$

We use $w(t) = 1$ in practice for simplicity. Note that external rewards typically operate in pixel space, $R_{\text{ext}}(x^{(i)}, c)$, where $x^{(i)} = \text{Dec}(z_0^{(i)})$ for a fixed decoder $\text{Dec} : \mathcal{Z} \to \mathcal{X}$. In contrast, R_{SOLACE} is computed *directly in latent space*, avoiding decoding and keeping the signal model-native.

4.2. Stabilization and Efficiency Techniques

Denoising reduction for efficient training. Following Flow-GRPO [41], we shorten the reverse-time horizon by subsampling the denoising steps. This reduces compute without degrading gains: e.g., while SD3.5 uses 40 steps at inference, we use 10 during training. We find that this does not sacrifice image quality at test time, while enabling faster training.

Timestep selection for self-confidence probing. We probe self-confidence at the *exact scheduler timesteps* used by the SD3.5 sampler (same discretization and indices), ensuring alignment with the generation trajectory. This avoids mismatch between sampling and probing, yielding more reliable credit assignment.

Training on selective timesteps. We observe that training on all denoising timesteps easily leads to collapse (e.g., blank or textureless images), a form of reward hacking in which the model steers latents toward regimes where injected noise becomes trivially easy to predict. We mitigate this by training on only a suffix of the schedule, *i.e.* a fixed percentage of the later reverse steps, where the denoising task remains informative but is harder to exploit. Let

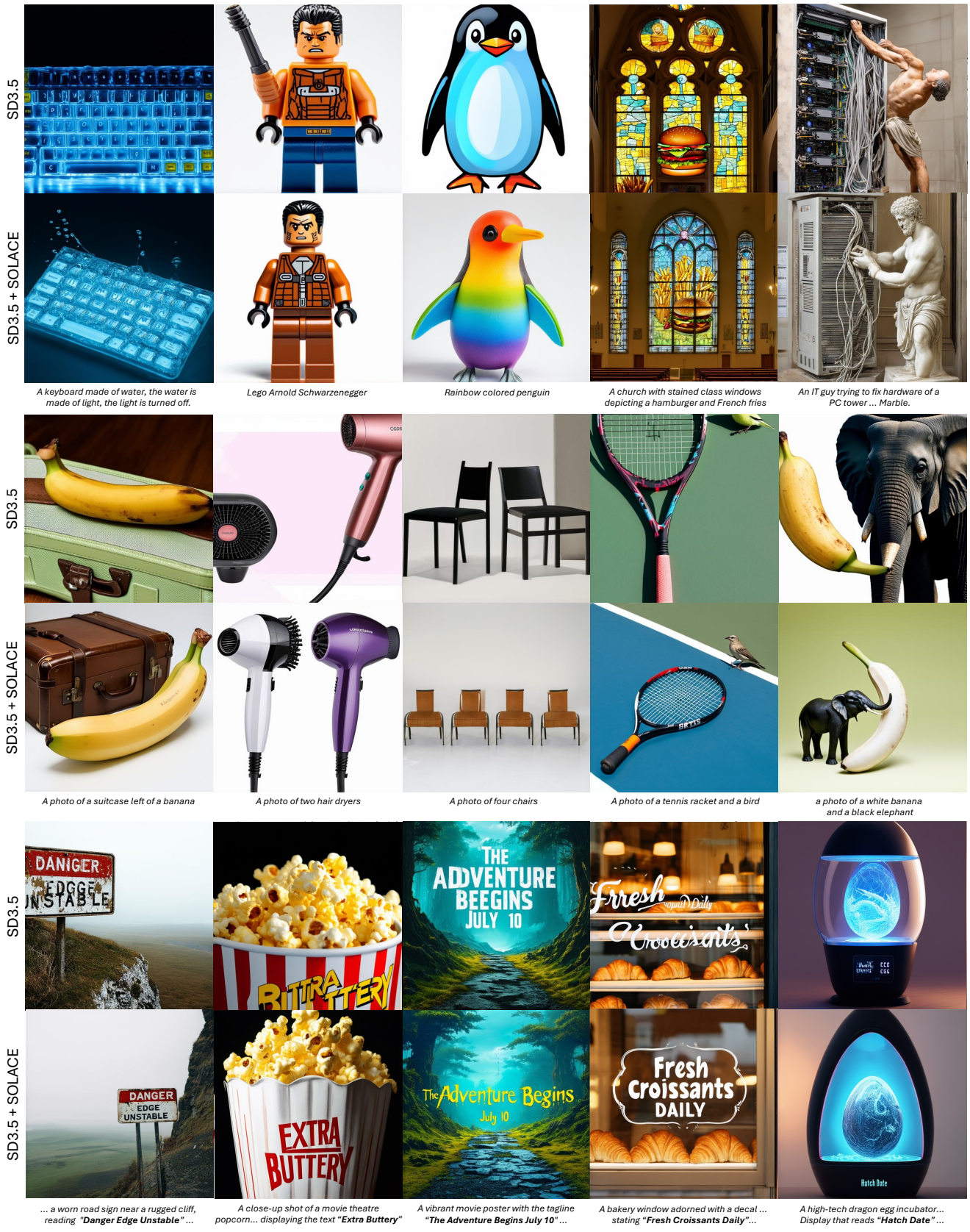

Figure 3. **Qualitative results of SOLACE on SD3.5 [16] across DrawBench [61], GenEval [21] and OCR [14].** SOLACE shows consistent improvements over the baseline SD3.5.

$\mathcal{T}_{\text{train}} \subset \mathcal{T}$ denote this suffix window ($|\mathcal{T}_{\text{train}}| = \lceil \rho |\mathcal{T}| \rceil$); we apply GRPO losses only on $t \in \mathcal{T}_{\text{train}}$, which stabilizes learning without collapse.

CFG-free self-confidence computation. Although G images are sampled with CFG for GRPO training, SOLACE self-confidence is computed *without* CFG. CFG forms a mixture field $v_{\text{cfg}} = v_{\text{uncond}} + s\,(v_{\text{cond}} - v_{\text{uncond}})$; computing self-confidence on this mixture would measure confidence of the guided proxy rather than the base conditional model. Empirically, omitting CFG during self-confidence computation yields stronger and more stable improvements.

Online calculation of self-confidence. We can compute self-confidence either (1) online, using the model being trained (π_θ), or (2) offline, using a fixed base model (π_{ref}). While offline computation does not cause severe over-optimization [20], online computation yields better performance. We conjecture that as the model improves through SOLACE post-training, its self-confidence estimates become more reliable, reinforcing further gains.

5. Experiments

5.1. Implementation details

We use a group size $G = 16$ and number of noise probes $K = 8$ with antithetic pairing in our experiments. While SOLACE requires no external reward models, annotators, or preference data for training, it does require a *prompt corpus* to generate the terminal latents for training; we use the train set of the visual text rendering task [14] from Flow-GRPO [41], which holds longer and more informative prompts compared to Pick-a-Pic [35] or GenEval [21]. We note that SOLACE improves across different prompt sources (see supplementary Sec. 16). We use LoRA [28, 47] with rank $r = 32$ and scaling factor $\alpha = 64$ for parameter-efficient post-training. We use the AdamW [45] optimizer with constant learning rate of $3e^{-4}$, and the KL regularizer weight $\beta = 0.04$. In $|\mathcal{T}_{\text{train}}| = \lceil \rho |\mathcal{T}| \rceil$, we set $\rho = 0.6$, which yields improvements without reward hacking or training collapse. An image resolution of 512×512 is used for both training and testing. We use a CFG guidance scale of 7.0 at inference. All experiments are carried out on 8×NVIDIA RTX PRO 6000 Blackwell GPUs. We include more training details in the supplementary materials.

5.2. Evaluation setting

(1) Compositional image generation. We evaluate on GenEval [21], consisting of complex compositional prompts including object counting, attribute binding, and spatial relations. Evaluation is performed across six tasks: position, counting, attribute binding, colors, two objects, and single object. We follow the official evaluation pipeline, which detects object bounding boxes and colors, then infers spatial relations from the generated image. The scores are calculated in a rule-based manner *e.g.* for object counting, $r = 1 - \frac{|N_{\text{gen}} - N_{\text{ref}}|}{N_{\text{ref}}}$, where N_{gen} is the number of generated objects, while N_{ref} is the specified number of objects in the prompt.

(2) Visual text rendering. We use the 1,000 GPT4o [49]-generated test prompts from [41]. In each prompt, the exact string that should appear in the image (*i.e.* target text) is specified by `"{text}"`. We adhere to [22] to report $r = \max(0, 1 - \frac{N_e}{N_{\text{ref}}})$, where N_e is the minimum edit distance between the rendered text and the target text, and N_{ref} is the non-whitespace length of the target text.

(3) Human preference alignment. We report the model-based reward outputs from Pickscore [35], HPSv2 [80], ImageReward [83] and UnifiedReward [77], trained on large-scale human preference data. We use the test prompts from DrawBench [61] to generate the images for evaluation.

(4) Image quality evaluation. We additionally report the CLIP-Score [58] and Aesthetic Score [62] on DrawBench [61], to evaluate the overall quality of generated images independent of the above task-specific criteria.

5.3. Results

Quantitative results. Results are shown in Tab. 1. Applying SOLACE on SD3.5-M yields consistent gains across task-specific, image quality, and human preference metrics. While improvements in human preference are modest, we observe substantial gains in compositional generation (GenEval [21]), text rendering (OCR [14]), and CLIP-Score [58], nearly matching the performance of SD3.5-L in these metrics despite having less than $\frac{1}{3}$ of the parameters (2.5B vs. 8.1B). This shows that the model's intrinsic self-confidence is strongly correlated with compositionality, text rendering, and text-image alignment.

We also analyze the effect of applying SOLACE *after* post-training SD3.5-M with external rewards via Flow-GRPO [41]. The results show that while performance on the targeted external reward is mildly compromised, we consistently gain improvements across GenEval, OCR, and CLIPScore. This strengthens our hypothesis that intrinsic self-confidence is strongly correlated with compositionality, text rendering, and text-image alignment, and that SOLACE alleviates the reward hacking typically seen in external-reward post-training. In Fig. 5, we show visual examples of SD3.5-M post-trained with FlowGRPO (PickScore), then further post-trained with SOLACE, showing that the two rewards are complementary.

User study. In Fig. 4, we provide the results of a user study on prompts from PartiPrompt [86] and HPSv2 [80], asking users to assess the generated images based on visual appeal/realism and text alignment. We summarize ∼3,600 responses from 40 participants. The results show that SD3.5-M post-trained with SOLACE consistently outperforms the

Model	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
SDXL	0.55	0.14	0.287	5.60	22.42	0.280	0.76	2.93
SD3.5-L	0.71	0.68	0.289	5.50	22.91	0.288	0.96	3.25
SD3.5-M	0.65	0.61	0.282	5.36	22.34	0.279	0.84	3.08
+ SOLACE (Ours)	**0.71**	**0.67**	**0.288**	**5.39**	**22.41**	0.278	**0.87**	**3.11**
SD3.5-M + FlowGRPO	0.95	0.65	0.293	5.32	22.51	0.272	1.06	3.18
	0.67	0.92	0.290	5.32	22.41	0.280	0.95	3.14
	0.54	0.68	0.278	5.90	23.50	0.314	1.26	3.37
SD3.5-M + FlowGRPO +**SOLACE** (Ours)	0.92	**0.71**	**0.294**	**5.35**	22.50	**0.277**	1.06	**3.26**
	0.72	0.89	**0.291**	**5.39**	**22.45**	**0.284**	**0.97**	**3.19**
	0.77	**0.70**	**0.287**	5.63	22.73	0.286	1.07	3.26

Table 1. **Quantitative results of SOLACE.** We evaluate SOLACE on SD3.5 [16] across GenEval [21], Text Rendering, human preference models [35, 77, 80, 83], and image quality metrics. SOLACE yields consistent gains across all quantitative metrics. In the bottom section, each row of SD3.5-M + FlowGRPO corresponds to a different external reward used for FlowGRPO training; the blue cell indicates which metric was used as the external reward.

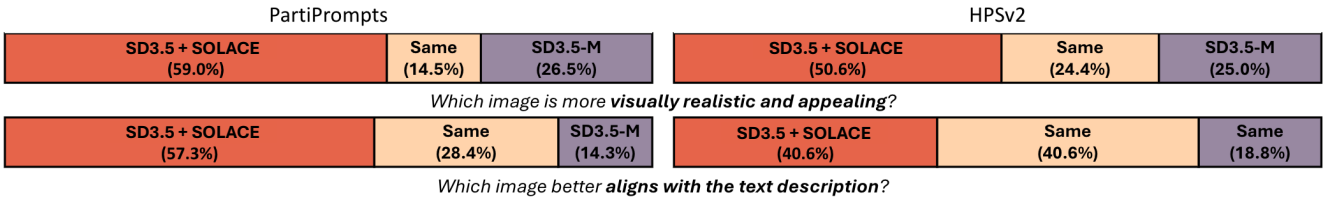

Figure 4. **User study against baseline SD3.5-M [16] on PartiPrompts [61] and HPSv2 [80].** The user study shows that SOLACE post-training yields favorable visual realism/appeal, and text-image alignment.

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
Number of noise probes K								
$K = 4$	0.71	0.66	0.287	5.37	22.34	0.273	0.81	3.08
$K = 8$ (Ours)	0.71	0.67	0.288	5.39	22.41	0.278	0.87	3.11
$K = 16$	0.70	0.67	0.288	5.42	22.34	0.278	0.86	3.09
Classifier-Free Guidance for self-confidence calculation								
O	0.68	0.59	0.287	5.38	22.39	0.278	0.85	3.10
X (Ours)	0.71	0.67	0.288	5.39	22.41	0.278	0.87	3.11
Offline vs Online Self-Confidence								
Offline	0.69	0.61	0.285	5.36	22.36	0.274	0.82	3.07
Online (Ours)	0.71	0.67	0.288	5.39	22.41	0.278	0.87	3.11

Table 2. **Ablation study results of SOLACE.** We validate the design choices of SOLACE over number of noise probes K, the usage of CFG for self-confidence calculation, and online/offline self-confidence calculation. Our current configurations yield superior results.

baseline in both visual realism/appeal and text alignment.

Qualitative comparison. We provide additional qualitative comparisons in Fig. 1 and Fig. 3, showing that SOLACE yields visually appealing results with improved compositionality and text rendering, even without any external reward. We note that SOLACE learns to generate images more tailored to the given prompt; when prompted with detailed descriptions, SOLACE produces realistic outputs.

5.4. Ablation study and analyses

In Tab. 2, we provide ablation study results to validate the design and hyperparameter choices of SOLACE.

Analyses on number of noise probes K. We vary K across $4, 8, 16$. The results show that $K = 8$ yields slightly better results overall. While $K = 16$ slightly outperforms $K = 8$ in aesthetic score, the improvement is negligible relative to the additional compute cost.

Figure 5. **Effect of SOLACE post-training SD3.5-M after post-training on PickScore [35] using FlowGRPO [41].** SOLACE complements external rewards, showing the best compositional generation and visual appeal on GenEval [21]. Post-training on external rewards yields high visual appeal, but sacrifices compositionality as shown above (Column 3: Generates yellow motorcycle instead / generates unwanted human).

CFG for self-confidence. Using CFG during self-confidence computation results in a slight performance drop. We conjecture this is because CFG is an inference-time technique, and using it inside the reward would optimize the *guided proxy* rather than the base conditional policy $\pi_\theta(\cdot \mid z_t, c)$. This may incentivize reward hacking via guidance strength rather than learning a better π_θ.

Online self-confidence vs Offline self-confidence. We compare post-training performance when self-confidence is computed online (*i.e.* using the model being trained, π_θ) versus offline (*i.e.* using the fixed base model, π_{ref}). Using offline self-confidence as a static reward results in lower performance across metrics, suggesting that online computation, which improves alongside the model, provides a stronger training signal.

Observed causes of training collapse. Training collapses when (1) we train on too many timesteps, *i.e.* $\rho > 0.6$ in $|\mathcal{T}_{\text{train}}| = \lceil \rho |\mathcal{T}| \rceil$, or (2) we do not use CFG for sampling the G candidates. In both cases, over-optimization against the self-confidence reward occurs, producing textureless images due to reward hacking. See the supplementary for detailed analysis.

Rationale of self-confidence as reward. We test whether self-confidence correlates with image quality by comparing three inference regimes: (i) 10 steps without CFG, (ii) 10 steps with CFG, and (iii) 20 steps with CFG. As shown in Fig. 6, the self-confidence distribution shifts rightward from (i) to (iii), matching the rise in visual quality. Since the *same* model computes the signal regardless of guidance or step count, better samples are easier to self-denoise, motivating self-confidence as a reward.

Figure 6. **Rationale of SOLACE.** Distributions of self-confidence under three inference settings. The distribution shifts rightward (higher self-confidence) as visual quality improves, showing that noise recovery accuracy is predictive of sample quality.

5.5. Limitations of SOLACE

One limitation is that intrinsic self-confidence does not align strongly with human preference; observed gains on preference metrics are modest. Also, while SOLACE improves compositional generation, text rendering, and text faithfulness, it cannot target a specific alignment objective on its own. However, we showed that SOLACE can be integrated with external rewards to target specific alignments while alleviating reward hacking and improving compositionality or text rendering capabilities (Tab. 1). We note that SOLACE's self-confidence is computed *under the same text conditioning* c in $r(x, c)$, which reduces (but does not eliminate) the risk of reinforcing prompt-agnostic high-density modes; we provide empirical analysis on rare compositions and diversity preservation in the supplementary (Sec. 14).

6. Conclusion

We introduced SOLACE, a post-training framework that replaces external rewards with intrinsic self-confidence, defined as the model's ability to recover noise injected into its own outputs. Across benchmarks and a user study, reinforcing higher self-confidence yields consistent improvements in compositionality, text rendering, and text-image alignment. SOLACE also complements external rewards: applying it on externally post-trained models improves non-target capabilities while alleviating reward hacking. We demonstrate SOLACE's generality across architectures, model scales, resolutions, and modalities in the supplementary. Future directions include (i) multi-view extensions to carry SOLACE to 3D and 4D generation, and (ii) calibrating intrinsic signals for task-targeted reward shaping.

Acknowledgement. This work was supported by the IITP grants (RS-2022-II220290: Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense (40%), RS-2022-II220113: Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework (50%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%), RS-2025-02653113: High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale (5%)) funded by the Korea government (MSIT).

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 1

[2] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023. 2

[3] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 1, 2, 3, 9

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[5] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 1, 2, 3

[6] Frederic Boesel and Robin Rombach. Improving image editing models with generative data refinement. In *The Second Tiny Papers Track at ICLR 2024*, 2024. 1

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 1

[8] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1, 2

[10] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 1, 2

[11] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024. 3

[12] Pengyu Cheng, Yong Dai, Tianhao Hu, Han Xu, Zhisong Zhang, Lei Han, Nan Du, and Xiaolong Li. Self-playing adversarial language game enhances llm reasoning. *Advances in Neural Information Processing Systems*, 37:126515–126543, 2024. 3

[13] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 2

[14] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 1, 2, 5, 6

[15] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. 2

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 2, 3, 5, 7, 9

[17] Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with wasserstein regularization. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[18] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023. 2

[19] Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024. 2

[20] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023. 6

[21] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 1, 2, 5, 6, 7, 8, 4

[22] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun

Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025. 1, 6

[23] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022. 6

[24] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[25] Shashank Gupta, Chaitanya Ahuja, Tsung-Yu Lin, Sreya Dutta Roy, Harrie Oosterhuis, Maarten de Rijke, and Satya Narayan Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*, 2025. 2

[26] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 1

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[29] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 1

[30] Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *arXiv preprint arXiv:2501.07730*, 2025. 2

[31] Seungwook Kim, Kejie Li, Xueqing Deng, Yichun Shi, Minsu Cho, and Peng Wang. Enhancing 3d fidelity of text-to-3d using cross-view correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10649–10658, 2024. 1

[32] Seungwook Kim, Yichun Shi, Kejie Li, Minsu Cho, and Peng Wang. Multi-view image prompted multi-view diffusion for improved 3d generation. *arXiv preprint arXiv:2404.17419*, 2024. 1

[33] Seungwook Kim, Seunghyeon Lee, and Minsu Cho. Freeaction: Training-free techniques for enhanced fidelity of trajectory-to-video generation. *arXiv preprint arXiv:2509.24241*, 2025. 1

[34] Seungwook Kim, Yichun Shi, Kejie Li, Minsu Cho, and Peng Wang. Rapidmv: Leveraging spatio-angular latent space for efficient and consistent text-to-multi-view synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1674–1684, 2026. 1

[35] 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. *Advances in neural information processing systems*, 36:36652–36663, 2023. 1, 2, 6, 7, 8, 4

[36] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 3

[37] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 2

[38] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023. 2

[39] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13199–13208, 2025. 2

[40] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[41] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 1, 2, 3, 4, 6, 8, 9

[42] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 2

[43] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8009–8019, 2025. 2

[44] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[46] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 2

[47] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan.

PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022. 6

[48] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024. 2

[49] OpenAI. Hello gpt-4o, 2024. 6

[50] Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with LLM guidance. In *International Conference on Learning Representations*, 2025. 2, 4

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 2

[52] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. 2

[53] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2

[54] Gabriel Poesia, David Broman, Nick Haber, and Noah Goodman. Learning formal mathematics from intrinsic motivation. *Advances in Neural Information Processing Systems*, 37:43032–43057, 2024. 3

[55] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2, 3

[56] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 2

[57] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 2

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 6

[59] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 2

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2, 5, 6, 7, 4

[62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 6

[63] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[64] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3, 4

[65] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 1

[66] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2, 3

[67] Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024. 1

[68] Inkyu Shin, Chenglin Yang, and Liang-Chieh Chen. Deeply supervised flow-based generative models. *arXiv preprint arXiv:2503.14494*, 2025. 2

[69] Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023. 1

[70] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 1

[71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[72] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025. 2

[73] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model

beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[74] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 1, 2

[75] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3

[76] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1

[77] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 2, 6, 7

[78] Yinong Oliver Wang, Younjoon Chung, Chen Henry Wu, and Fernando De la Torre. Domain gap embeddings for generative dataset augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28684–28694, 2024. 1

[79] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36: 8406–8441, 2023. 1, 2

[80] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1, 2, 6, 7

[81] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 1

[82] Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Qiushi Sun, Kanzhi Cheng, Junxian He, Jun Liu, and Zhiyong Wu. Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning. *arXiv preprint arXiv:2504.08672*, 2025. 3

[83] 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 1, 2, 6, 7

[84] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 2

[85] Zuhao Yang, Fangneng Zhan, Kunhao Liu, Muyu Xu, and Shijian Lu. Ai-generated images as data source: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023. 1

[86] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2, 6

[87] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024. 2

[88] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *Advances in Neural Information Processing Systems*, 37:73366–73398, 2024. 2

[89] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024. 3

[90] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022. 3

[91] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159*, 2024. 2

[92] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025. 1

[93] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025. 3

[94] Hanyang Zhao, Haoxian Chen, Ji Zhang, David D Yao, and Wenpin Tang. Score as action: Fine-tuning diffusion generative models by continuous-time reinforcement learning. *arXiv preprint arXiv:2502.01819*, 2025. 2

[95] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025. 3

[96] 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. *arXiv preprint arXiv:2509.16117*, 2025. 2

[97] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1

[98] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025. 3