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

Supplementary Material

Supplementary Contents

7. SOLACE Post-Training on SD3.5-L

To assess scalability, we apply SOLACE to SD3.5-L [16], a larger base model than the SD3.5-M used in the main experiments. Unless otherwise noted, we reuse the same training recipe (shortened denoising horizon, suffix-only updates, shared probes, CFG-free scoring). As reported in Tab. 3, SOLACE yields *consistent gains* in compositional generation, text rendering, and text–image alignment, while remaining competitive on human-preference metrics (e.g., HPSv2, PickScore). These results suggest that SOLACE scales to higher-capacity text-to-image models without inducing reward hacking and remains effective beyond the SD3.5-M setting.

8. Applying SOLACE on FLUX.1-Dev

To test architectural generality, we apply SOLACE to **FLUX.1-Dev** [3], a flow-matching text-to-image generator with a design distinct from SD3.5. We keep the core SOLACE recipe unchanged (shortened denoising horizon, suffix-only updates, shared probes, CFG-free scoring), adapting only to the model's native scheduler and inference step count. A small deviation is the suffix window: we set $\rho = 0.5$, i.e., train on the latter half of the scheduler steps, which increased training stability in this setting. As reported in Tab. 3, SOLACE delivers *consistent gains* in compositional generation, text rendering, and text–image alignment, while remaining competitive on human-preference metrics (e.g., HPSv2, PickScore). The results indicate that SOLACE transfers effectively across architectures and remains robust on another representative flow-matching T2I model.

9. Applying SOLACE on SDXL

To verify that SOLACE is not inherently tailored to DiT-based or flow-matching architectures, we apply SOLACE to **SDXL** [53], a UNet-based latent diffusion model. Despite the architectural differences (SDXL uses a UNet backbone with DDPM-style noise scheduling rather than DiT-based flow matching), SOLACE produces consistent improvements in compositional generation (GenEval) and text rendering (OCR), as shown in Tab. 4. These results suggest that SOLACE's self-confidence reward is architecture-agnostic and can benefit UNet-based diffusion models as well.

10. SOLACE for Text-to-Video Generation

To test the applicability of SOLACE beyond text-to-image generation, we apply SOLACE to **Wan2.1-1.3B** [75], a text-to-video diffusion model. We evaluate on the VBench-1.0 [29] subset, and report the results in Tab. 5. As shown in the table, SOLACE yields improvements in subject consistency, background consistency, and dynamic degree, while maintaining competitive motion smoothness, demonstrating that SOLACE generalizes effectively to the text-to-video generation setting. Qualitative results are provided in Fig. 7. For instance, in the jellyfish example (top), SOLACE produces noticeably more stable jellyfish movements compared to the baseline. In the "bicycle gliding through a snowy field" example (bottom), the baseline generates an unnatural gliding motion where the gliding direction does not match the bicycle's orientation, whereas SOLACE produces a much more natural and coherent gliding motion.

11. Resolution Analysis

Our main experiments use 512×512 resolution for both training and evaluation, following the configuration of Flow-GRPO [41]. To verify that the improvements transfer across resolutions, we additionally train SOLACE at 1024×1024 resolution and evaluate both models at both scales.

As shown in Tab. 6, SOLACE trained at 512×512 (SOLACE_{512}) transfers well to 1024×1024 inference, yielding consistent improvements in GenEval and OCR at the higher resolution. SOLACE trained directly at 1024×1024 (SOLACE_{1024}) also shows gains, though with a slightly different trade-off profile across metrics. These results confirm that SOLACE's benefits are not resolution-specific.

12. Comparison with Closed-Source Models

To contextualize SOLACE's improvements, we evaluate two closed-source models (Gemini 2.5-Flash and GPT-image-1.5) on our benchmark suite. As shown in Tab. 7, closed-source models achieve higher absolute scores due to larger model capacities and proprietary training data. Nevertheless, SOLACE narrows the gap from the SD3.5-M baseline, particularly in compositional generation and text rendering.

13. Training Collapse Analysis

When and why collapse occurs. We monitor the batch-mean self-confidence (negative log error, averaged over probes and probed timesteps) across training iterations. Collapse is characterized by

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
SD3.5-M (2.5B)	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-L† (8.1B)	0.71	0.68	0.289	5.50	22.91	0.288	0.96	3.25
(Reproduced)	0.51	0.68	0.284	5.28	21.86	0.264	0.70	2.98
†**+ SOLACE** (Ours)	0.58	0.74	0.288	5.25	21.91	0.253	0.65	2.98
FLUX.1-Dev† (12B)	0.66	0.59	0.295	5.71	22.69	0.292	0.96	3.27
(Reproduced)	0.66	0.61	0.269	5.71	22.84	0.274	0.88	3.21
+ SOLACE (Ours)	0.66	0.65	0.271	5.67	22.69	0.292	0.90	3.23

Table 3. **Applying SOLACE to SD3.5-L [16] and FLUX.1-Dev [3].** We apply SOLACE on additional models of SD3.5-L and FLUX.1-Dev, to verify the effect of SOLACE given (1) a larger base model, and (2) a different architecture from SD3.5-M. † denotes results taken from DiffusionNFT [96]. We base our experiments on our reproduced results based on the official weights of SD3.5-L [16] and FLUX.1-Dev [3]. The results show that SOLACE consistently results in improved compositionality, text rendering and text-image alignment, while being competitive at human preference metrics.

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
SDXL [53]	0.23	0.127	0.284	5.58	22.34	0.274	0.67	2.92
+ SOLACE (Ours)	**0.25**	**0.144**	0.284	5.57	22.33	0.270	**0.70**	**2.94**

Table 4. **Applying SOLACE to SDXL [53].** SOLACE yields improvements in compositional generation and text rendering on a UNet-based diffusion model, demonstrating architecture-agnostic applicability.

	Subj. Consist.	BG Consist.	Aesth. Qual.	Motion Smooth.	Dyn. Deg.
Wan2.1-1.3B	0.94	0.96	**0.59**	**0.97**	0.47
+ SOLACE	**0.95**	**0.97**	0.58	**0.97**	**0.51**

Table 5. **Applying SOLACE to Wan2.1-1.3B for text-to-video generation.** Evaluation on VBench-1.0 subset. SOLACE improves subject consistency, background consistency, and dynamic degree while maintaining competitive motion smoothness.

a rapid, sustained surge in this score (an overconfidence spike), followed by degenerate, low-texture generations (reward hacking). Empirically, two settings precipitate this behavior: (i) training on too many timesteps ($\rho > 0.6$ in $|\mathcal{T}_{\text{train}}| = \lceil \rho |\mathcal{T}| \rceil$), which exposes early, easily exploitable steps; and (ii) sampling the G rollout candidates *without* CFG, which reduces exploration and inflates apparent self-confidence. A KL anchor alone is insufficient to prevent these modes.

Mitigations used in SOLACE. We restrict training to the latter 60% of steps ($\rho = 0.6$), keep CFG *on* during rollouts (but *off* when scoring self-confidence), and retain clipping, per-timestep weighting, and antithetic probes. These choices suppress overconfidence spikes and stabilize learning.

Why SOLACE's reward is amenable to targeted stabilization. Since the reward is a monotonic transform of denoising error (*i.e.* $r = -\log(\text{MSE} + \delta)$), the degenerate solution is concrete and diagnosable: maximizing $\mathbb{E}_{z_0 \sim \pi_\theta(\cdot|c)}[r(z_0)]$ can steer samples toward latent regimes where injected noise becomes trivially predictable (*e.g.* low-variance, textureless outputs). Because self-confidence is not a fixed black-box oracle, we can directly modify the *reward computation itself* (solver-aligned timestep probing, suffix-window training, and no-CFG scoring) to suppress these shortcut solutions, rather than relying solely on generic stabilizers (*e.g.* KL weights or reward scaling) that do not change what the reward measures.

14. Diversity and Semantic Correctness Analysis

A natural concern with self-confidence as a reward is whether it biases the model toward high-density but semantically incorrect modes, or reduces sample diversity. We address both concerns empirically.

Semantic correctness on rare compositions. Self-confidence is computed *under the same text conditioning* c in $r(x, c)$, which reduces pressure toward prompt-agnostic high-density modes. To test whether SOLACE degrades on less common compositional prompts, we evaluate on RareBench [50], a benchmark consisting of diverse and complex rare concept compositions. As shown in Tab. 8, CLIPScore on RareBench is largely preserved after SOLACE post-training, suggesting no measurable degradation on rare or out-of-distribution compositions.

Diversity preservation. We measure diversity using the mean pairwise CLIP embedding distance across 64 samples per prompt on 50 DrawBench [61] prompts. As reported in Tab. 8, the diver-

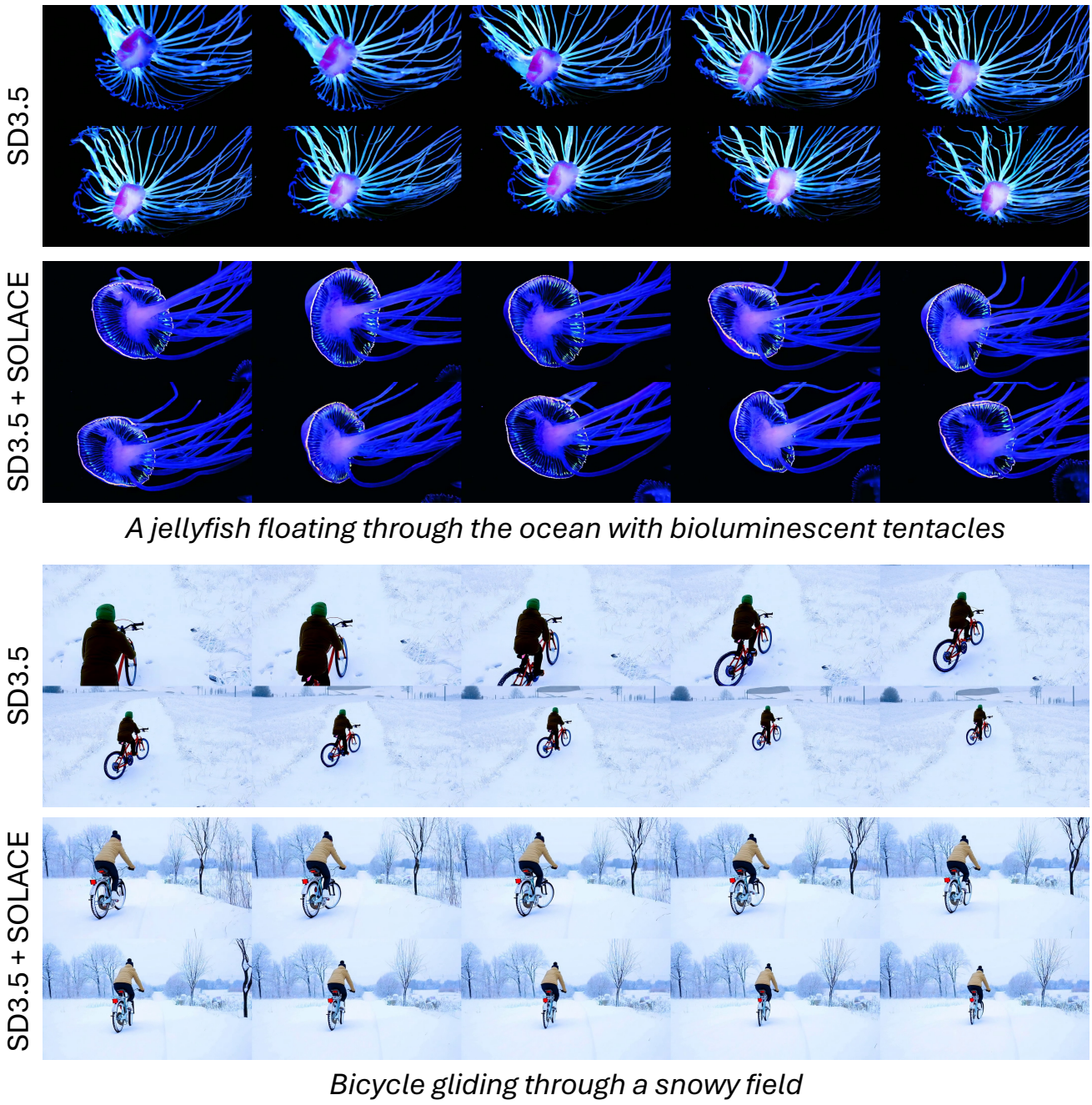

Figure 7. **Qualitative results of SOLACE on Wan2.1-1.3B.** SOLACE produces videos with improved visual quality and prompt adherence compared to the base model.

sity score is maintained (and even slightly improved) after SOLACE post-training. This is consistent with the fact that SOLACE's reward measures conditional denoising self-consistency rather than explicitly minimizing conditional entropy $H(x|c)$, and the GRPO objective with KL regularization provides sufficient diversity preservation.

15. Effect of Negative Advantages

SOLACE uses GRPO, where updates are weighted by a signed, within-group advantage: samples with below-average self-confidence receive negative advantages and are explicitly down-weighted. To verify the importance of this negative signal, we compare against a *positive-only* variant that clips advantages to be non-negative (*i.e.* $\max(\widehat{A}_t^i, 0)$), effectively removing the penalty

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
Inference at 512×512								
SD3.5-M	0.65	0.61	0.282	5.36	22.34	0.279	0.84	3.08
+ SOLACE_{512}	**0.71**	**0.67**	**0.288**	**5.39**	**22.41**	**0.284**	**0.87**	**3.10**
+ SOLACE_{1024}	0.68	0.63	0.284	**5.39**	22.39	**0.284**	**0.87**	**3.10**
Inference at 1024×1024								
SD3.5-M	0.65	0.57	**0.293**	**5.98**	21.91	**0.305**	**1.15**	**3.48**
+ SOLACE_{512}	**0.71**	**0.64**	0.292	5.95	21.68	0.283	1.00	3.41
+ SOLACE_{1024}	0.68	0.63	0.289	5.38	**22.48**	0.283	0.93	3.19

Table 6. **Resolution analysis.** SOLACE trained at 512×512 transfers effectively to 1024×1024 inference, with consistent gains in compositional generation and text rendering across resolutions.

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
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
Gemini 2.5-Flash	0.75	0.72	0.270	5.70	23.02	0.287	0.79	3.45
GPT-image-1.5	0.84	0.81	0.286	5.54	23.24	0.301	1.11	3.57

Table 7. **Comparison with closed-source models.** While closed-source models achieve higher absolute scores due to larger capacities and proprietary training, SOLACE narrows the gap from the SD3.5-M baseline, particularly in compositional generation and text rendering.

	CLIPScore↑ (RareBench)	Diversity Score↑ (DrawBench)
SD3.5-M	0.2752	0.9519
SD3.5-M + SOLACE	0.2746	0.9545

Table 8. **Semantic correctness and diversity analysis.** CLIPScore on RareBench [50] (rare compositions) and diversity score on DrawBench [61] (64 samples per prompt, 50 prompts) show that SOLACE preserves both semantic accuracy on uncommon concepts and sample diversity.

for low-confidence samples.

As shown in Tab. 9, the positive-only variant underperforms the full SOLACE objective on GenEval, OCR, and CLIPScore, confirming that negative advantages provide important learning signal. While the positive-only variant achieves higher aesthetic and some human preference scores, it sacrifices the core compositional and text-rendering gains that SOLACE targets.

16. Additional Ablation Studies

We conduct additional ablation studies and comparative experiments to validate the design choices of SOLACE. The results are summarized in Tab. 11.

16.1. Caption datasets for SOLACE

SOLACE relies on intrinsic self-confidence and thus requires only prompts (not external reward models). We compare three prompt sources: (i) *text-rendering (OCR)* prompts from Flow-GRPO [41] (our default), (ii) *PickScore* [35] prompts, and (iii) *GenEval* [21] prompts. As shown in Tab. 3, denser, more prescriptive prompts (OCR) yield the strongest gains; empirically, self-confidence is most reliable when the text condition is explicit and descriptive. We provide the descriptions and examples for each prompt dataset in Tab. 10.

16.2. Effect of group size

We clarify a typographical error in the main paper: although we stated G=24, all experiments used G=16. Varying G shows that G=16 outperforms G=8 (more within-prompt exploration improves group-relative normalization) while G=32 destabilizes training: larger groups reduce the number of distinct prompts per batch, lowering inter-prompt diversity and increasing the risk of over-optimization under relative advantages. In practice, G=16 strikes a robust compute–stability trade-off.

16.3. Stepwise vs. aggregated reward

Although SOLACE's self-confidence can be computed per step, we find that using the *aggregated* reward, *i.e.*, averaging weighted per-step scores over the probed timesteps, consistently performs better than optimizing stepwise advantages. Stepwise improvements at individual timesteps need not translate to a better final sample and tend to increase variance and solver sensitivity; ag-

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
SD3.5-M	0.65	0.61	0.282	5.36	22.34	0.279	0.84	3.08
+ SOLACE	**0.71**	**0.67**	**0.288**	5.39	**22.41**	0.278	0.87	3.11
+ SOLACE (positive-only)	0.69	0.62	0.285	**5.80**	21.57	**0.281**	**0.91**	**3.20**

Table 9. **Effect of negative advantages.** Removing negative advantages (positive-only variant) degrades compositional generation, text rendering, and text-image alignment, demonstrating that the full signed advantage is important for SOLACE's effectiveness.

(i) Text-rendering (OCR) prompts — default	
Characteristics	Dense, explicit textual content (exact strings, font/placement hints), strong conditioning for legibility and alignment.
Examples	"A postage stamp design featuring the motto "Unity in Diversity", showcasing a vibrant collage of people from various ethnic backgrounds, each holding hands in a circle, set against a backdrop of colorful, interwoven patterns symbolizing unity and cultural richness." "In a luxurious hotel lobby, an elegant digital display above the elevator reads "Now Playing". Soft, ambient elevator music fills the space, enhancing the serene and welcoming atmosphere. A plush, modern sofa and a glass coffee table are seen in the foreground, with polished marble floors reflecting the ambient light." "A sleek, modern corporate lobby featuring a large, minimalist sculpture prominently inscribed with "Innovate or Perish", reflecting the company's commitment to forward-thinking. The sculpture stands against a backdrop of glass and steel, with subtle lighting enhancing its form and the powerful message it conveys."
(ii) PickScore prompts	
Characteristics	Open-ended statements; often adds context with simple concatentation of adjectives; weaker constraints on text content/layout.
Examples	"An attractive young woman petting a cat" "(a girl in steampunk fantasy world), (ultra detailed prosthetic arm and leg), (beautifully drawn face:1.2), blueprints, (magic potions:1.4), mechanical tools, plants, (a small cat:1.1), silver hair, (full body:1.2), magic dust, books BREAK (complex ultra detailed of medieval fantasy city), (steampunk fantasy:1.2), indoors, workshop, (Steam-powered machines:1.2), (clockwork automatons:1.2), (a small wooden toy), (intricate details:1.6), lamps, colorful details, iridescent colors, BREAK illustration, ((masterpiece:1.2, best quality)), 4k, ultra detailed, solo, (photorealistic:1.2), asymmetry, looking at viewer, smile" "Cyborg cow, cyberpunk alien india, body painting, bull, star wars design, third eye, mehendi body art, yantra, cyberpunk mask, baroque style, dark fantasy, kathakali characters, high tech, detailed, spotlight, shadow color, high contrast, cyberpunk city, neon light, colorful, bright, high tech, high contrast, synthesized body, hyper realistic, 8k, epic ambient light, octane rendering, kathakali, soft ambient light, HD,"
(iii) GenEval prompts	
Characteristics	Compositional verification (objects, counts, relations), moderate specificity, minimal typography.
Examples	"a photo of a yellow bus and an orange handbag" "a photo of four surfboards" "a photo of a book left of a cat"

Table 10. **Prompt sources compared for SOLACE.** Denser, text-focused prompts (OCR) provide stronger supervision signals for intrinsic self-confidence, leading to larger gains than more open-ended (PickScore) or simple compositional (GenEval) prompts.

gregation provides a more stable, outcome-aligned signal for post-training.

17. Additional Implementation Details

In this section we summarize the main implementation choices used in our SOLACE training pipeline. **We acknowledge and correct a typographical error in the main paper**: although we stated that the group size was $G = 24$, all experiments were in fact conducted with $G = 16$. The summary of hyperparameters and configurations is illustrated in Tab. 12.

17.1. Base models and LoRA configuration

We build on the `StableDiffusion3Pipeline` from `diffusers` with the pretrained model SD3.5-M: `stabilityai/stable-diffusion-3.5-medium`. We freeze all components except the denoiser: the VAE and all text encoders are kept fixed and used only for inference. Only the main transformer (UNet-like denoiser) is updated during training, based on LoRA. We run the text encoders in mixed precision

	Task-specific		Image Quality		Human Preference			
	GenEval	OCR	ClipScore	Aesthetic	PickScore	HPSv2.1	ImageReward	UnifiedReward
Caption dataset used for SOLACE								
-	0.65	0.61	0.282	5.36	22.34	0.279	0.84	3.08
PickScore prompts	0.70	0.62	0.285	5.26	22.13	0.278	0.65	2.96
GenEval prompts	0.71	0.62	0.286	5.32	22.35	0.275	0.80	3.05
OCR prompts (Ours)	**0.71**	**0.67**	**0.288**	**5.39**	**22.41**	0.278	**0.87**	**3.11**
Group size G								
8	0.70	0.64	0.285	5.29	22.28	0.267	0.75	3.00
16 (Ours)	**0.71**	**0.67**	**0.288**	**5.39**	**22.41**	0.278	**0.87**	**3.11**
32	0.61	0.51	0.274	5.18	21.73	0.226	0.16	2.73
Step-wise reward vs. Aggregated reward								
Stepwise	0.67	0.60	0.285	**5.39**	22.36	0.277	0.83	3.07
Aggregated (Ours)	**0.71**	**0.67**	**0.288**	**5.39**	**22.41**	0.278	**0.87**	**3.11**

Table 11. **Additional ablation/comparative results.** The results show that our current design choices for the (1) Caption dataset used, (2) Group size G, and (3) Aggregated self-confidence rewrads yield the best performances.

(`fp16` in our main SOLACE runs) and keep the VAE in `fp32` for stability.

For parameter-efficient fine-tuning we apply LoRA to the transformer with

- LoRA rank $r = 32$ and scaling factor $\alpha = 64$,
- Gaussian initialization of LoRA weights,
- Target modules inside each attention block:

`attn.add_k_proj`, `attn.add_q_proj`, `attn.add_v_proj`, `attn.to_add_out`, `attn.to_k`, `attn.to_q`, `attn.to_v`, `attn.to_out.0`.

All non-LoRA base weights remain frozen.

17.2. Datasets and prompt processing

We consider two kinds of prompt datasets:

- **Plain text prompt datasets.** We store the prompts in plain text files `train.txt` and `test.txt`. Each line contains a single prompt string. (*e.g.* PickScore, Text Rendering dataset)
- **GenEval-style metadata.** For experiments on GenEval-style prompts we use JSONL files `{train,test}_metadata.jsonl`, where each line is a JSON object that contains at least a `"prompt"` field and additional metadata.

For each batch of prompts we compute text embeddings using the three SD3.5 text encoders. We also precompute embeddings for the empty prompt `""` and use them as unconditional embeddings for classifier-free guidance (CFG) during sampling and log-probability computation.

17.3. Distributed sampling and grouping

We use HuggingFace Accelerate [23] for distributed training. Let N be the number of GPUs (processes), and let B_{sample} denote the per-device sample batch size. In our main SOLACE setting we use $N = 8, B_{\text{sample}} = 8, G = 16$. Thus a single sampling batch contains $NB_{\text{sample}} = 64$ images, corresponding to $64/16 = 4$ distinct prompts, each with $G = 16$ candidate images. We train for 2,000 iterations, which takes around 30 hours on 8×NVIDIA 332 RTX PRO 6000 Blackwell GPUs.

17.4. KL regularization

Following Flow-GRPO, regularize the policy via a KL term that constrains the transition mean to stay close to a reference (the base model without LoRA):

- The SDE step module returns the current mean μ_θ and a reference variance σ_t^2.
- We compute a reference mean μ_{ref} by temporarily disabling LoRA adapters and re-evaluating the same step.
- Assuming Gaussian transitions with equal variance, the per-step KL divergence simplifies to

$$D_{\text{KL}} = \frac{1}{2\sigma_t^2} \|\mu_\theta - \mu_{\text{ref}}\|_2^2 .$$

We average this KL over spatial dimensions and the batch and add it to the policy loss with weight $\beta = 0.04$.

18. User Study Instructions and Interface

We provide the details of the instructions and interface used for the user study.

Instructions. For each text prompt, you will be shown a pair of *AI-generated* images (left and right). For every image pair, you are asked to answer the following two questions *independently*:

1. **Visual realism and appeal:** Which image do you find to be more visually realistic and appealing?
2. **Text–image alignment:** Which image better aligns with the given text description?

For each question, please select your preferred image (left or right) based solely on the specified criterion.

Interface. The user interface used in the study is illustrated in Fig. 9.

Category	Hyperparameter	Value (SOLACE, SD3.5-M)
Model	Base model	`stabilityai/stable-diffusion-3.5-medium` (SD3.5-M)
	Components trained	Transformer (denoiser) only; VAE and all text encoders frozen
LoRA	LoRA usage	`use_lora = True`
	Rank r	32
	Scaling factor α	64
	Init of LoRA weights	Gaussian
	Target modules	`attn.add_k_proj`, `attn.add_q_proj`, `attn.add_v_proj`, `attn.to_add_out`, `attn.to_k`, `attn.to_q`, `attn.to_v`, `attn.to_out.0`
Data / prompts	Train / test files	`train.txt`, `test.txt` (one prompt per line)
	Tokenization	SD3.5 tokenizers; max length 128 (embeddings), 256 (logging)
Sampling	Image resolution	512×512
	Sampler steps (train / eval)	train:10, eval:40
	Train timestep fraction	`train.timestep_fraction = 0.99` $\Rightarrow T_{\text{train}} = 9$
	Suffix proportion ρ in GRPO	0.6
	Guidance scale (train/eval)	`sample.guidance_scale = 4.5`
	Noise level (SDE step)	`sample.noise_level = 0.7`
	Train batch size / GPU (sampling)	`sample.train_batch_size = 8` images
	Test batch size / GPU	`sample.test_batch_size = 16` images
	Images per prompt (group size G)	`sample.num_image_per_prompt = 16`
	Number of GPUs	8
	Batches per epoch (sampling)	`sample.num_batches_per_epoch = 4`
	Global samples / batch	8 (bs) $\times$ 8 (GPUs) $= 64$ images
	Prompts / batch	$64/16 = 4$ prompts per sampling batch
	Same latent per prompt	`sample.same_latent = False`
Self-confidence (SOLACE)	Probes per step K	8 (antithetic pairing: $K/2$ noise, $K/2$ negated)
	Probe timesteps	Last half of used timesteps: $j = 4, \ldots, 8$ (for $T_{\text{train}} = 9$)
	Noise schedule for probe	$\lambda_t = \tau_t/1000$; $x_t = (1 - \lambda_t)x_0 + \lambda_t \epsilon$
	Per-step score	$s_t = -\log(\text{MSE}_t + 10^{-6})$, MSE between injected and predicted noise
	Normalization	Per-timestep batch-wise z-score, then mean over timesteps
	CFG inside probe	Disabled (conditional branch only)
Training (GRPO)	PPO / GRPO clip range	$\rho_{i,t}$ clipped to $[1 - \texttt{clip\_range}, 1 + \texttt{clip\_range}]$ (PPO style)
	KL regularizer weight	`train.beta = 0.04`
	KL form	$D_{\text{KL}} = \|\mu_\theta - \mu_{\text{ref}}\|_2^2/(2\sigma_t^2)$ (mean-only Gaussian)
Optimization / EMA	Optimizer	AdamW on LoRA parameters (no base-parameter updates)
	Learning rate	3×10^{-4} (constant)
	Gradient clipping	Global norm clipping at `train.max_grad_norm`
	EMA usage	`train.ema = True`
	EMA decay	0.9
	EMA update interval	Every 8 optimizer steps (`update_step_interval = 8`)
	EMA usage in eval	EMA weights used for evaluation; online weights restored afterwards
External rewards / eval	Training reward	Internal self-confidence only (no external reward in training)
	SDS-only eval	Optional SDS self-confidence evaluation on EMA model for monitoring

Table 12. Hyperparameters and key implementation details for SOLACE training on SD3.5-M.

Figure 8. **Visualization of training collapse in SOLACE.** Self-confidence (y-axis) versus training iteration under different settings. Using $\rho > 0.6$ or sampling rollouts without CFG drives a steep, short-horizon increase in self-confidence, followed by degenerate outputs—evidence of reward hacking. SOLACE's default settings (ρ=0.6 and CFG for rollouts) avoid this behavior while preserving steady improvements.

19. Additional Qualitative Results

We provide side-by-side samples for (i) PickScore–post-trained (Flow-GRPO) SD3.5–M, (ii) FLUX.1–Dev, and (iii) SD3.5–L in Fig. 10. Across diverse prompts, SOLACE yields visibly sharper text rendering, more faithful object counts and relations, and fewer artifacts, echoing the quantitative gains in compositionality, text rendering, and text–image alignment, with no obvious regressions on non-target aspects.

Figure 9. User study interface used to collect human preferences between pairs of AI-generated images.

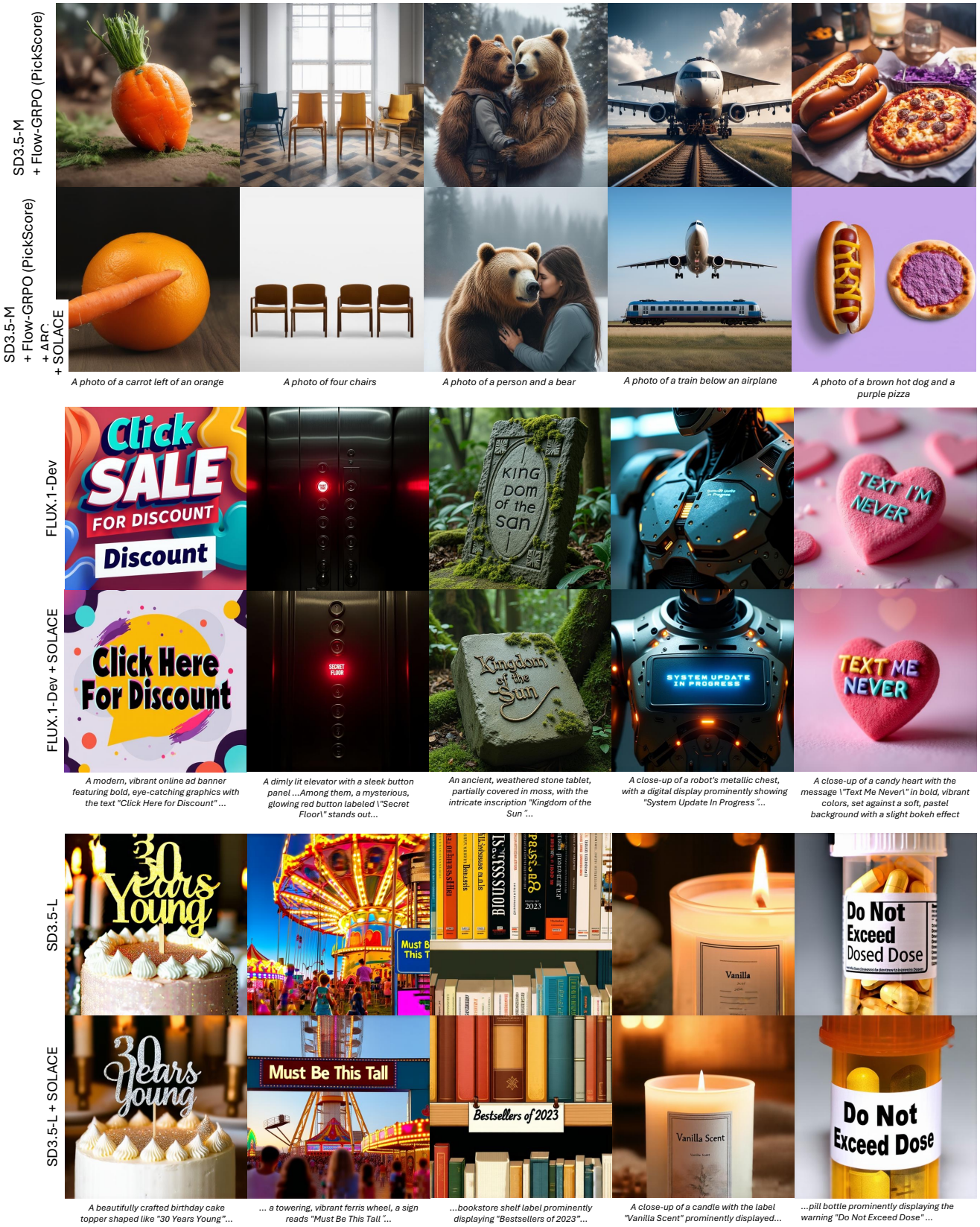

Figure 10. **Additional qualitative results of SOLACE.** We present additional qualitative results of SOLACE when applied to (1) Flow-GRPO [41] post-trained SD3.5-M [16], (2) FLUX.1-Dev [3], and (3) SD3.5-L [16]. Best viewed on electronics.

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