ALIGN AND RECOVER: TRAINING-FREE GUIDANCE FOR TEXT-TO-IMAGE DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free guidance improves text-to-image synthesis by adjusting the sampling process of a pretrained diffusion model, although stronger prompt adherence can come at the expense of local visual quality. In self-guidance, contrasting original and perturbed predictions provides a sampling direction, but amplifying the entire difference couples structural adjustments with fine-scale changes. To reconcile these effects, we propose Align–Recover Guidance (ARG), which uses the perturbation response for structural alignment and the denoised estimate for appearance refinement. Spectral–Semantic Alignment combines a low-pass prediction contrast with concentrated text attention, while local contrast left unrecovered by this restriction is addressed through Denoised-Content Recovery, which enhances the estimate's high-pass residual. As noise decreases, Trajectory Harmonization progressively relaxes alignment, allowing content recovery to continue under weaker structural guidance. ARG requires no retraining, backpropagation, or external reward model. On SDXL, ARG reduces FID by 24.2% on COCO-2014 30k and 16.4% on COCO-2017 5k relative to published SSG.

1 INTRODUCTION

Text-to-image synthesis aims to generate visually coherent images that faithfully express natural-language descriptions. Diffusion methods generate images by reversing a gradual noising process, while latent diffusion makes high-resolution synthesis practical by performing this process in a compressed image space (Ho et al., 2020; Song et al., 2020b; Rombach et al., 2022; Podell et al., 2024). Sampling guidance provides additional control over image quality and prompt adherence at inference time. Classifier guidance uses classifier gradients to steer generation, whereas classifier-free guidance (CFG) uses the difference between conditional and unconditional predictions (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). Self-guidance obtains a reference prediction from the generator itself through input or attention perturbations, avoiding the need for a separately trained guiding network (Hong et al., 2023; Ahn et al., 2024; Hong, 2024). For instance, Self-Swap Guidance (SSG) constructs this reference by selectively exchanging spatial and channel tokens (Zhang et al., 2026). Although these methods refine the reference prediction, the resulting prediction difference still affects both coarse structure and fine detail. Increasing the guidance strength amplifies these effects together, making it difficult to control structural alignment and local appearance separately.

Separating structural guidance from appearance refinement requires different controls over the perturbation response and the estimated image content. Adjusting the overall guidance strength alone offers limited control over the quality–alignment tradeoff (Sadat et al., 2024; Kynkäänniemi et al., 2024). Low-pass filtering can restrict fine-scale changes in the perturbation response, but it does not directly enhance local contrast in the generated image. The distinction lies in what the high-frequency components represent: those in the prediction difference reflect changes caused by perturbation, whereas those in the denoised estimate capture local variation in the image itself. The denoised estimate therefore provides a basis for enhancing local content without reintroducing the filtered perturbation response. Meanwhile, filtering controls the spatial frequencies used for guidance but does not explicitly strengthen text conditioning, which calls for a complementary adjustment through cross-attention. Separating these operations also allows their strengths to vary during sampling. Structural guidance and attention sharpening can gradually weaken near the end of denoising, while local content recovery continues.

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

To address the coupled amplification of structural and fine-scale changes, we propose Align–Recover Guidance (ARG), a training-free framework that separates alignment guidance from appearance recovery and coordinates their effects throughout the sampling trajectory (Fig. 1). Specifically, Spectral–Semantic Alignment (SSA) applies low-pass filtering to the self-swap prediction difference and sharpens cross-attention. The former suppresses fine-scale changes induced by perturbation, while the latter reinforces the correspondence between generated content and the text prompt. As filtering the perturbation response does not directly recover local image contrast, Denoised-Content Recovery (DCR) enhances the high-pass residual of the denoised estimate $\widehat{\mathbf{x}}_0$ during the second half of sampling. DCR operates on the estimated image content rather than the perturbation-induced prediction difference, thereby improving local appearance without restoring the filtered response. To accommodate these different information sources over time, Trajectory Harmonization (TH) jointly relaxes structural guidance and attention sharpening late in sampling while retaining content recovery, rather than attenuating both forms of guidance together. ARG requires neither backpropagation nor an external reward model and can be directly applied to a pretrained generator. Compared with the previous SOTA method SSG, experimental results on SDXL show that ARG reduces FID by 24.2% on COCO-2014 30k and by 16.4% on COCO-2017 5k. ImageReward scores also increase by 88.0% and 103.6%, respectively, complementing the FID reductions with gains in model-assessed preference. Our contributions are:

- We propose ARG, a training-free guidance framework that separately controls the perturbation response and the denoised image content. Spectral–Semantic Alignment combines low-pass self-swap guidance with cross-attention sharpening to control spatial frequency and text conditioning.
- We introduce Denoised-Content Recovery to enhance local variation in the denoised estimate without reintroducing the filtered perturbation response. Trajectory Harmonization coordinates these distinct signals through asymmetric temporal control, relaxing alignment late in sampling while retaining recovery.
- We evaluate ARG on SDXL using COCO-2014 30k and COCO-2017 5k. Relative to published SSG results, ARG achieves FID reductions of 24.2% and 16.4%, respectively, while increasing ImageReward scores by 88.0% and 103.6% on the corresponding benchmarks.

2 Related work

2.1 Diffusion models and sampling guidance

Diffusion models generate images through successive denoising steps, with DDPM learning a probabilistic reverse process (Ho et al., 2020) and score-based models describing generation through stochastic differential equations (Song et al., 2020b). To reduce sampling cost, DDIM permits non-Markovian sampling, including a deterministic trajectory, under the same training objective (Song et al., 2020a). For high-resolution synthesis, LDMs reduce the computational cost of each denoising step by operating in a learned, compressed image space (Rombach et al., 2022). Our experiments use the pretrained latent diffusion model SDXL (Podell et al., 2024). Sampling guidance further improves generation by adjusting denoising predictions at inference time. Classifier guidance uses gradients from an auxiliary classifier to favor a target class (Dhariwal & Nichol, 2021), whereas classifier-free guidance (CFG) constructs a guidance direction from conditional and unconditional predictions (Ho & Salimans, 2022). For more explicit content control, Composable Diffusion combines concept-conditioned distributions (Liu et al., 2022), while SEGA guides generation along semantic directions (Brack et al., 2023). Conventional CFG, however, requires an unconditional prediction obtained through joint training or condition dropout. Independent-condition guidance avoids this specialized training procedure by constructing a reference prediction from an independently sampled condition (Sadat et al., 2025a). Autoguidance instead uses a smaller, less-trained model under the same conditioning to improve the balance between quality and variation, at the cost of an additional network (Karras et al., 2024).

Recent self-guidance methods obtain a reference prediction directly from the generator by perturbing its inference computation. For example, SAG blurs image regions identified by self-attention (Hong et al., 2023), while timestep guidance (TSG) perturbs the timestep embedding (Sadat et al., 2025a). Within the attention layers, PAG replaces selected self-attention maps with identity maps (Ahn et al., 2024), whereas SEG smooths the energy landscape of self-attention (Hong, 2024). Other

approaches perturb intermediate representations or network computation: TPG shuffles intermediate token representations (Rajabi et al., 2026), and stochastic self-guidance randomly drops network blocks (Chen et al., 2026). SSG makes the perturbation more selective by exchanging dissimilar token pairs along the spatial or channel dimension (Zhang et al., 2026). Although this design provides finer control over the construction of the reference prediction, the difference between the original and perturbed predictions is still applied as a single guidance term. Consequently, its coarse structural response and fine-scale perturbation response remain controlled by the same guidance strength. Our work retains the token-swapped reference of SSG and further introduces frequency-selective alignment, denoised-content recovery, and trajectory-dependent coordination.

2.2 Frequency-aware and time-dependent quality control

To reduce artifacts under strong guidance, CFG++ keeps the sampling update closer to the data manifold (Chung et al., 2025), while Adaptive Projected Guidance adjusts the parallel and orthogonal components of the guidance direction (Sadat et al., 2024). Frequency-Decoupled Guidance more directly controls spectral content by weighting the low- and high-frequency components of the CFG residual separately (Sadat et al., 2025b). Within attention, OSEG selectively smooths orthogonal token components to preserve image details (Fahim et al., 2026). Related feature-level adjustments include backbone and skip-feature reweighting in FreeU (Si et al., 2024) and the use of massive activations for local detail synthesis in Detail Guidance (Gan et al., 2026). Guidance-weight schedules (Wang et al., 2024), limited-interval guidance (Kynkäänniemi et al., 2024), and stage-wise dynamics (Jin et al., 2026) further consider how guidance should vary during denoising. Additional work on attention control and compositional semantics is discussed in Appendix B. Unlike methods that reweight the frequency components of a single CFG residual, ARG uses a low-pass perturbation response for alignment and the high-pass residual of the denoised estimate for appearance recovery, while relaxing alignment late in sampling and retaining content recovery.

3 Methodology

To reconcile structural guidance with local appearance, ARG distinguishes the prediction response induced by perturbation from the content represented in the denoised estimate. SSA constrains the spectrum and text dependence of the former, whereas DCR refines the latter without re-amplifying the structural guidance response. TH progressively relaxes alignment while retaining recovery along the trajectory. Fig. 1 summarizes this separate control, without implying independence.

3.1 Preliminaries

Let $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ denote the noise prediction for latent $\mathbf{x}_t$ at noise level σ_t under prompt $\mathbf{c}$. For the i-th ($i \in \{0, \dots, N-1\}$) denoiser evaluation, normalized sampling progress is $p = i/(N-1)$. We use an Euler solver (Karras et al., 2022) and adopt Self-Swap Guidance (SSG) as the base guidance mechanism (Zhang et al., 2026). SSG evaluates two prompt-conditioned branches that differ only in whether dissimilar spatial and channel tokens are exchanged within the first six transformer blocks of the U-Net upsampling path. Their prediction difference defines the contrast used for guidance:

$$\mathbf{d}_t = \boldsymbol{\epsilon}_t^o - \boldsymbol{\epsilon}_t^s, \qquad \boldsymbol{\epsilon}_t^{SSG} = \boldsymbol{\epsilon}_t^o + s\mathbf{d}_t, \tag{1}$$

where o and s denote the original and swapped branches, respectively. Because both branches use the same prompt, the swapped prediction requires neither a null condition nor an auxiliary network.

3.2 Spectral–Semantic Alignment

Token swapping produces a prediction contrast containing both coarse spatial changes and fine-scale perturbation responses, which a single guidance coefficient cannot regulate separately. SSA therefore attenuates fine-scale variation in the contrast while retaining its coarse-scale contribution. Because this spectral constraint does not determine how text tokens contribute to each image query, SSA also sharpens cross-attention with a temperature shared by both branches, preserving a common conditioning rule. This temperature is applied in every cross-attention layer:

$$\mathrm{Attn}_{\tau(p)}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}\,\tau(p)}\right)\mathbf{V}, \tag{2}$$

where d is the key dimension and $\tau_0 = 0.5$ is the base temperature. For fixed logits, decreasing the temperature concentrates attention on tokens with higher compatibility scores, whereas $\tau = 1$ recovers the standard attention rule. Because both branches use the same temperature, their prediction difference continues to arise from token swapping under a common text-conditioning rule.

Let $\boldsymbol{\epsilon}_t^o(\tau(p))$ and $\boldsymbol{\epsilon}_t^s(\tau(p))$ denote the predictions of the original and swapped branches under the shared temperature, with their contrast given by $\mathbf{d}_t = \boldsymbol{\epsilon}_t^o(\tau(p)) - \boldsymbol{\epsilon}_t^s(\tau(p))$. To suppress fine-scale perturbation responses without smoothing the original-branch prediction, we apply a normalized separable Gaussian kernel G_{σ_g} only to this contrast. The filter operates independently on each channel over the spatial dimensions of the latent tensor, with $\sigma_g = 3$ and $*$ denoting spatial convolution:

$$\mathbf{d}_t^L = G_{\sigma_g} * \mathbf{d}_t, \qquad \bar{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_t^o(\tau(p)) + s(p)\mathbf{d}_t^L. \tag{3}$$

The original-branch prediction remains unfiltered because the Gaussian acts only on the added contrast, yielding a soft frequency response rather than a hard cutoff. Unlike latent-gradient attention optimization (Chefer et al., 2023), this shared-temperature adjustment requires no backward pass, although it cannot correct an already strong but erroneous association.

3.3 Denoised-Content Recovery

Suppressing fine-scale perturbation responses does not itself restore local variation in the generated content, while directly reintroducing the removed contrast would partly reverse SSA because that signal describes the response to token swapping. DCR instead derives a recovery signal from the current denoised estimate. Since this estimate is less reliable for local refinement early in sampling, DCR begins in the second half of the trajectory, at $p_0 = 0.5$. For $p \geq p_0$, the estimate is

$$\widehat{\mathbf{x}}_0 = \mathbf{x}_t - \sigma_t \bar{\boldsymbol{\epsilon}}_t, \qquad p_0 = 0.5, \tag{4}$$

where $\bar{\boldsymbol{\epsilon}}_t$ is the SSA prediction defined in Eq. (3). We use a smoothed version of $\widehat{\mathbf{x}}_0$ as a local reference and extract the corresponding high-pass residual, which is added back before the refined estimate is converted to the noise parameterization:

$$\mathbf{r}_t^H = \widehat{\mathbf{x}}_0 - G_{\sigma_u} * \widehat{\mathbf{x}}_0, \quad \widehat{\mathbf{x}}_0^R = \widehat{\mathbf{x}}_0 + \lambda \mathbf{r}_t^H, \quad \boldsymbol{\epsilon}_t^R = \frac{\mathbf{x}_t - \widehat{\mathbf{x}}_0^R}{\sigma_t}, \tag{5}$$

where $\lambda = 0.1$ controls recovery and $\sigma_u = 2$ sets the smoothing scale. The resulting residual enhances variation already represented in the denoised latent rather than reconstructing absent content. Thus SSA filters the perturbation contrast, whereas DCR refines the estimated content; converting the recovered estimate back to the noise parameterization lets this refinement affect subsequent denoising steps. Although separately adjustable, λ and $s(p)$ affect the same latent.

3.4 Trajectory Harmonization

Because SSA guides through perturbation contrasts while DCR refines estimated content, their relative influence need not remain fixed throughout sampling. Persistent structural extrapolation near the low-noise end may alter established spatial relations, whereas reducing both signals would also weaken local-contrast recovery. TH therefore coordinates alignment and recovery asymmetrically, jointly relaxing the contrast scale and attention concentration while retaining content refinement. The initial alignment strength is preserved until p_d, after which it decays gradually. Using $p = i/(N-1)$ as defined in Section 3.1, the schedules are written as

$$\tau(p) = \begin{cases} \tau_0, & p < p_d, \\ \tau_0 + (1-\tau_0)\dfrac{p - p_d}{1 - p_d}, & p \geq p_d, \end{cases} \qquad s(p) = \begin{cases} s_0, & p < p_d, \\ s_0 \dfrac{1-p}{1-p_d}, & p \geq p_d, \end{cases} \tag{6}$$

where $p_d = 0.7$ and $s_0 = 6$. Over the final 30% of sampling, $s(p)$ gradually approaches zero while $\tau(p)$ returns to one. Because DCR becomes active at $p_0 = 0.5$, recovery begins before alignment is progressively relaxed, producing a controlled overlap between the two mechanisms instead of switching alignment off at a single sampling step.

For a compact formulation of the complete method, we define the recovery operator as

$$\mathcal{R}(\mathbf{z}) = \mathbf{z} + \lambda\big(\mathbf{z} - G_{\sigma_u} * \mathbf{z}\big). \tag{7}$$

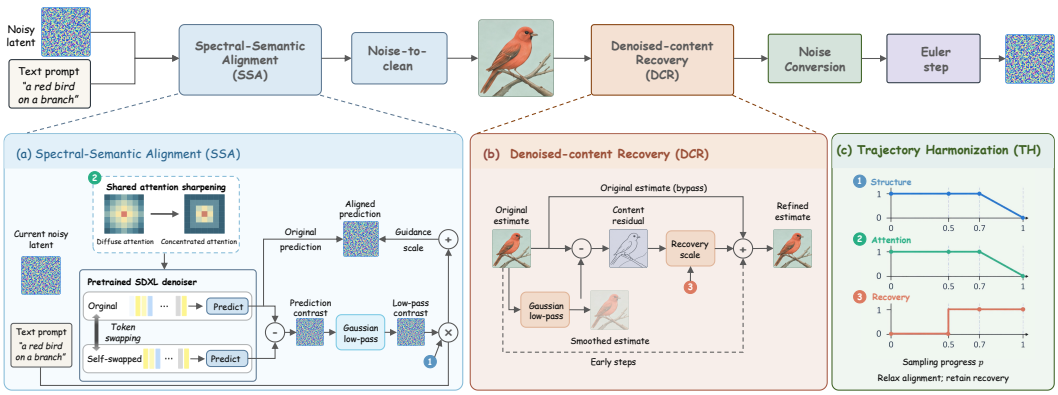

Fig. 1: Overview of Align–Recover Guidance (ARG). The upper pipeline combines alignment and denoised-content recovery before converting the refined estimate back to a noise prediction for the Euler update. (a) SSA sharpens cross-attention in both denoiser branches and adds a scaled low-pass self-swap contrast to the original prediction. (b) DCR enhances the high-pass residual of the denoised estimate, with the dashed path bypassing recovery during early steps. (c) TH activates recovery at $p = 0.5$ and progressively relaxes structural guidance and attention sharpening after $p = 0.7$. Markers 1–3 connect the corresponding strengths in (a,b) to their normalized schedules in (c), where attention denotes concentration rather than temperature.

Combining this operator with the SSA prediction $\bar{\epsilon}_t$ from Eq. (3), the final ARG prediction becomes

$$\epsilon_t^{ARG} = \begin{cases} \bar{\epsilon}_t, & p < p_0, \\ \dfrac{\mathbf{x}_t - \mathcal{R}(\mathbf{x}_t - \sigma_t \bar{\epsilon}_t)}{\sigma_t}, & p \geq p_0. \end{cases} \tag{8}$$

The Euler update then uses ϵ_t^{ARG}, while DCR and TH require no denoiser evaluations, trainable parameters, or reward queries beyond those already introduced by the two SSG branches. Appendix C.2 derives the exact single-step Euler displacement induced by DCR and explains how the solver step size offsets the apparent division by σ_t, while the evolving latent keeps alignment and recovery coupled throughout the complete sampling trajectory.

4 Experiments

4.1 Setup

Following SSG (Zhang et al., 2026), we evaluate on both the 30,000-prompt MS-COCO-2014 validation dataset (Lin et al., 2014) and the 5,000-prompt MS-COCO-2017 validation dataset. Because their validation images overlap, we treat them as two evaluation protocols rather than independent domains (Lin et al., 2014). Each run generates one 1024×1024 image per prompt with SDXL-base-1.0 (Podell et al., 2024), 50 Euler steps, no CFG, and prompt-index seed $1000i$. We report FID, IS, CLIP score, precision, recall, AES, ImageReward, and PickScore, and compare ARG with unguided SDXL and the training-free SAG, PAG, SEG, and SSG results reported by Zhang et al. (2026). Paired component studies hold the sampler, initial scale, prompts, and seeds fixed. Appendix A provides all metric references, generation settings, and evaluation details.

4.2 Main results

Tables 1 and 2 show that ARG consistently outperforms SSG results across both evaluation datasets. Relative to published SSG on COCO-2014, ARG reduces FID by 24.2% and increases CLIP, IS, ImageReward, and PickScore by 1.9%, 10.7%, 88.0%, and 0.14%, respectively. On COCO-2017, the corresponding improvements are a 16.4% FID reduction and increases of 2.2%, 5.8%, 103.6%, and 0.27% in the four scores. These gains come with a modest aesthetic tradeoff, as AES decreases by 0.7% and 0.5% on the two datasets, respectively, indicating that ARG improves the overall balance

Table 1: Quantitative comparison of conditional image generation by SDXL on MS-COCO 2014 (30k prompts, 50 steps). Baseline results are reported by Zhang et al. (2026); ARG is evaluated here.

Method	#Steps	FID↓	CLIP↑	IS↑	AES↑	PickScore↑	IR↑
w/o guidance	50	45.09	0.281	21.31	5.671	20.20	−0.847
SAG (Hong et al., 2023)	50	34.14	0.295	22.95	5.745	20.64	−0.487
SEG (Hong, 2024)	50	28.55	0.302	27.16	5.894	21.38	−0.016
PAG (Ahn et al., 2024)	50	26.55	0.306	29.70	5.820	21.56	−0.003
SSG (Zhang et al., 2026)	50	21.73	0.313	34.63	**5.902**	22.17	0.276
ARG (ours)	50	**16.47**	**0.319**	**38.34**	5.860	**22.20**	**0.519**

Table 2: Quantitative comparison of conditional image generation by SDXL on MS-COCO 2017 (5k prompts, 50 steps). Baseline results are reported by Zhang et al. (2026); ARG is evaluated here.

Method	#Steps	FID↓	CLIP↑	IS↑	AES↑	PickScore↑	IR↑
w/o guidance	50	54.93	0.280	21.00	5.660	20.17	−0.870
SAG (Hong et al., 2023)	50	43.76	0.293	23.08	5.745	20.60	−0.515
SEG (Hong, 2024)	50	38.66	0.301	27.70	5.881	21.35	−0.058
PAG (Ahn et al., 2024)	50	36.94	0.306	29.03	5.817	21.54	−0.015
SSG (Zhang et al., 2026)	50	31.92	0.312	34.15	**5.890**	22.14	0.253
ARG (ours)	50	**26.67**	**0.319**	**36.13**	5.860	**22.20**	**0.515**

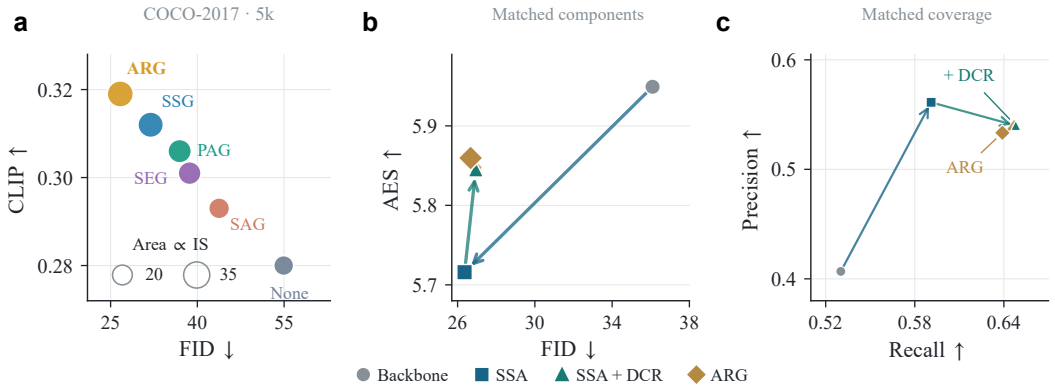

Fig. 2: Comparison of guidance methods and ARG components on COCO-2017. (a) FID and CLIP of the published methods and ARG, with bubble area proportional to IS. (b,c) FID–AES and precision–recall for four matched component settings; arrows indicate successive additions.

among the measured properties without outperforming SSG on every individual metric. The concurrent gains in CLIP and ImageReward complement the distributional improvements indicated by FID, showing that ARG improves both semantic alignment and preference-based generation quality. Appendix Table 6 reports the corresponding precision and recall results, while the matched controls presented below further examine the contributions of the proposed components.

4.3 From Structural Gain to Recovered Detail

Table 3 compares the components under shared seeds, with Backbone denoting SSG at scale 6 implemented in the same inference pipeline as ARG. Relative to Backbone, SSA reduces FID by 27.0% and increases CLIP and ImageReward by 0.62% and 17.0%, while AES decreases by 3.9%. Adding DCR to fixed SSA increases AES, ImageReward, and recall by 2.3%, 27.0%, and 9.6%, respectively, although FID increases by 2.2% and PickScore decreases by 0.07%. Relative to SSA+DCR, TH reduces FID by 1.0% and increases AES by 0.24%, with small decreases in recall and PickScore. As illustrated in Fig. 2, these successive changes trace a non-monotonic relationship between distributional quality and aesthetic scores, with fixed SSA retaining the lowest FID and Backbone the

Table 3: Ablation of SSA, DCR, and TH on COCO-2017 5k. All rows use the same prompts and seeds, and each row adds the named component to the preceding setting.

Variant	FID↓	CLIP↑	IS↑	AES↑	IR↑
Backbone	36.09	0.3140	30.72	**5.949**	0.341
+ SSA	**26.36**	0.3159	**37.49**	5.716	0.399
+ DCR	26.94	0.3186	36.74	5.845	0.507
+ TH	26.67	**0.3191**	36.13	5.860	**0.515**

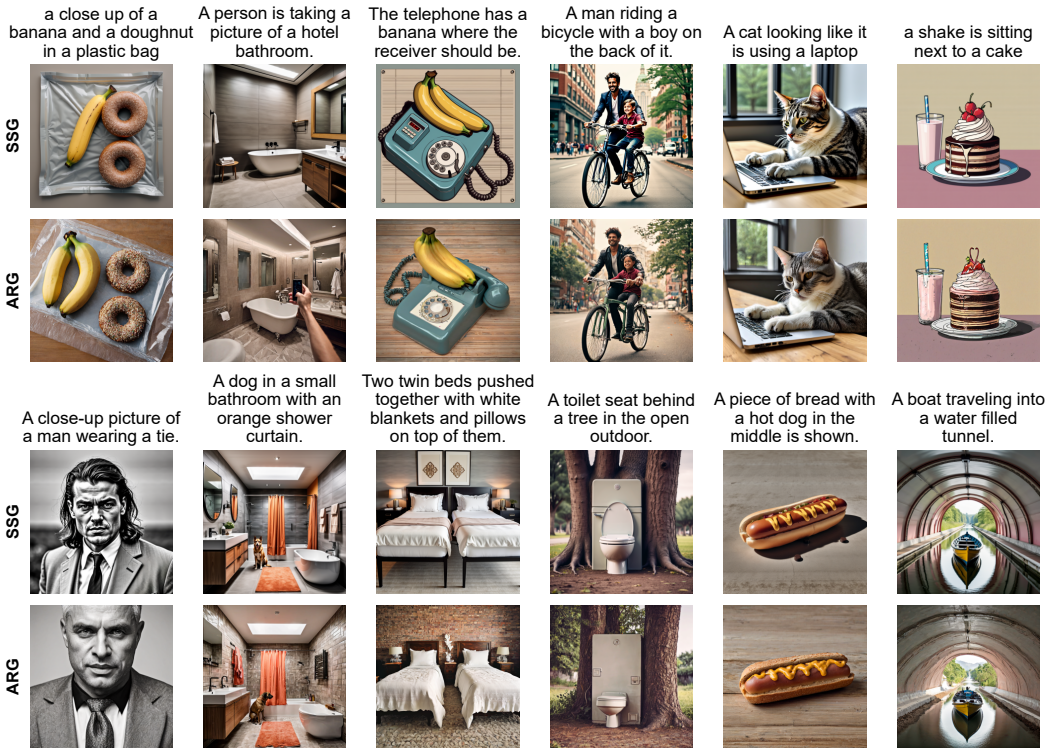

Fig. 3: Qualitative comparison of SSG and ARG with SDXL. The twelve prompt–seed pairs are illustrative cases selected from a 25-prompt archive. SSG denotes our matched scale-6 control rather than its published setting.

highest AES. DCR thereby recovers 55.5% of SSA's AES loss while preserving a 25.3% FID reduction over Backbone, retaining distributional gains alongside aesthetic recovery. Appendix D.2 reports eight metrics for matched rather than historical settings. Across these comparisons, alignment and recovery retain the spectral stage's FID benefit while offsetting much of its aesthetic loss. Relative to the controlled Backbone, full ARG increases precision by 31.1% and recall by 20.5% (Appendix D.2), improving fidelity and coverage.

4.4 Qualitative comparisons

Under the same experimental setting, the visualized comparisons in Fig. 3 make ARG's appearance gains visible beyond aggregate scores. Relative to previous SOTA method (SSG), the banana–doughnut and hot-dog images show more distinct food textures and material shading, while the bedding, cat, portrait, and bathroom examples exhibit clearer fabric, fur, facial, and room-surface detail. The five-stage controls in Appendix D.7 show that these local details become more pronounced when DCR is added to SSA, with the principal objects and broad scene layout remaining recognizable in the displayed cases. Together, these images provide qualitative support for the AES

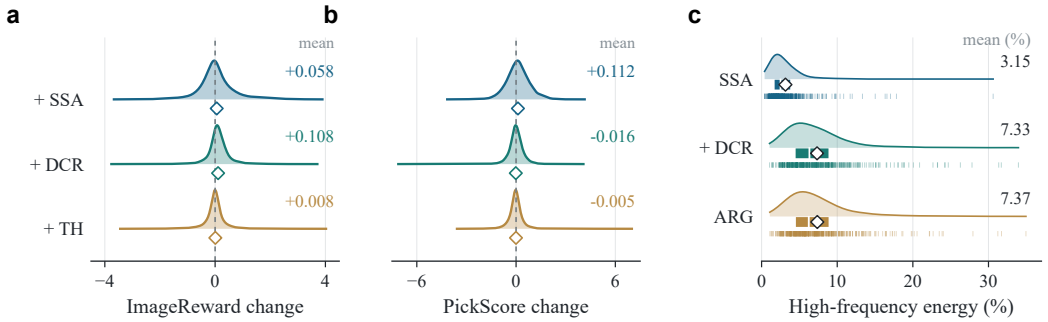

Fig. 4: Paired score changes and high-frequency energy across ARG components. (a,b) ImageReward and PickScore changes for Backbone→SSA, SSA→SSA+DCR, and SSA+DCR→ARG on 5000 matched prompts. (c) High-frequency energy on 500 matched prompts; diamonds mark means.

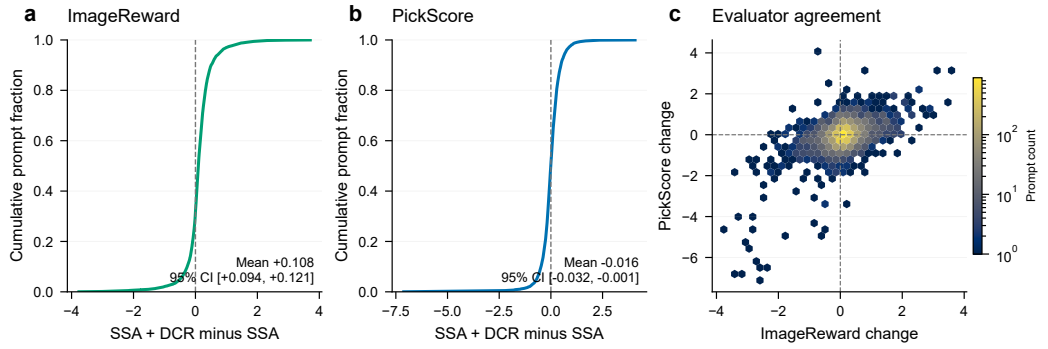

Fig. 5: Effect of DCR on paired reward scores for 5000 COCO-2017 prompts. (a,b) ImageReward and PickScore changes from SSA to SSA+DCR, with 95% intervals for the mean differences. (c) Joint distribution of the two changes; dashed lines mark zero.

and ImageReward gains in Table 3, while the score-ranked cases in Appendix D.7 document how the effect varies across prompts.

4.5 Paired effects and spectral evidence

Figure 4 provides paired evidence for the complementary contributions of the three components under the same experimental setting. Relative to the Backbone, SSA increases mean ImageReward by 17.0% and PickScore by 0.51%, showing that alignment improves both reward-based quality and learned-preference scores. Adding DCR further increases ImageReward by 27.0% relative to SSA, demonstrating that local appearance recovery provides a substantial additional gain beyond SSA, while TH retains a similar mean score, with a 1.6% increase relative to SSA+DCR whose paired interval includes zero (Appendix Table 7). The spectral analysis further supports the effect of DCR on local contrast, with mean high-frequency energy increasing by 132.7% relative to SSA (from 3.15% to 7.33%), followed by only a 0.4% relative increase under TH. Together with the visual comparisons, these results support complementary effects of SSA on alignment and DCR on local contrast, while TH largely preserves the added high-frequency energy as it balances the measured properties.

The distribution in Fig. 5 confirms that the recovery gain extends broadly across the prompt set rather than being driven by a few isolated cases. With SSA fixed, DCR improves ImageReward on 68.44% of all 5000 prompts, yielding a 27.0% mean-score increase with a paired 95% score-difference interval of $[0.0944, 0.1211]$. PickScore decreases by 0.07% with a score-difference interval of $[-0.0315, -0.0008]$, while still improving on 49.20% of the prompts, and both evaluators improve simultaneously on 1889 prompts or 37.78% of the complete set. These results show that

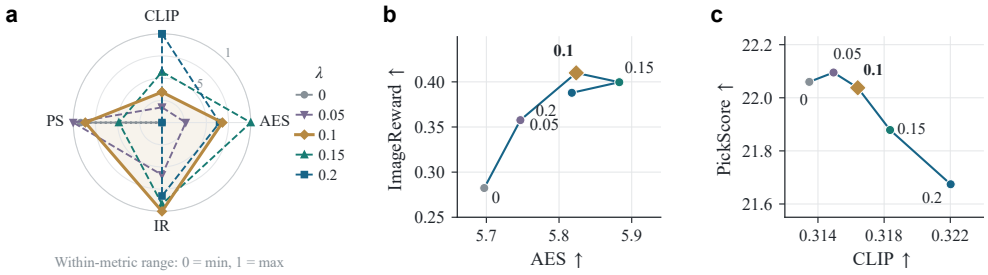

Fig. 6: Effect of recovery strength on image quality and prompt alignment. Results use 500 matched COCO-2017 prompts. (a) Values are scaled within each metric for display; (b,c) AES–ImageReward and CLIP–PickScore use raw scores. Gold marks the reference setting $\lambda = 0.1$.

DCR produces a broad improvement in ImageReward rather than an average gain dominated by a small number of prompts, while retaining competitive PickScore performance and achieving consistent gains across both evaluators for a substantial portion of the dataset.

When extreme scores are downweighted, DCR retains an ImageReward median gain of 0.092 and a 5% trimmed mean of 0.110. ARG also improves on Backbone in median ImageReward (0.090) and PickScore (0.070), supporting central gains, though not independent-seed stability (Appendix D.3).

4.6 Sensitivity and inference cost

Figure 6 examines the sensitivity to λ while holding SSA, TH, prompts, and seeds fixed, with the complete setup provided in Appendix D.4. Across $\{0, 0.05, 0.10, 0.15, 0.20\}$, ImageReward reaches its highest value at 0.10, AES at 0.15, and PickScore at 0.05, while CLIP continues to increase beyond the optima of the preference metrics. Therefore, we use $\lambda = 0.1$, which maximizes ImageReward while providing a balanced operating point across semantic alignment, aesthetic quality, and preference-based evaluation. Within this 500-prompt study, doubling λ from 0.1 to 0.2 increases CLIP by 1.8% but decreases ImageReward and PickScore by 5.4% and 1.7%, respectively. This divergence favors moderate recovery, since stronger amplification need not improve preference scores. ARG also introduces negligible method-specific computational cost. Interleaved measurements on one A800 using 20 shared prompts and two warmup runs yield 5.794 seconds per image for ARG and 5.787 seconds for the Backbone with the same attention processor, corresponding to an overhead of only 0.11%. Relative to native attention, the overhead is 5.68%, indicating that most of the remaining difference comes from the shared attention processor rather than the proposed components, while all six settings have the same peak memory usage of 8.95 GiB. ARG therefore adds negligible latency and no additional peak memory beyond the unified processing pipeline, with the complete measurement protocol and runtime distributions reported in Appendix D.5.

5 Conclusion

Training-free guidance for text-to-image synthesis steers a pretrained generator toward a prompt by modifying its sampling predictions. Since the contrast between original and perturbed predictions contains both coarse structural responses and fine-scale changes, amplifying the entire difference can improve prompt adherence at the expense of local appearance. We introduced Align–Recover Guidance (ARG) to control structural alignment and content refinement through distinct signals. SSA combines a low-pass prediction contrast with shared cross-attention sharpening, while DCR enhances the high-pass residual of the denoised estimate rather than reintroducing the discarded perturbation response. TH progressively relaxes alignment late in sampling while retaining content recovery, without retraining, backpropagation, or external reward queries. On SDXL, ARG reduces FID by 24.2% on COCO-2014 30k and 16.4% on COCO-2017 5k relative to published SSG. Matched component comparisons further show that DCR improves ImageReward and AES by 27.0% and 2.3% over SSA alone, although FID and PickScore exhibit trade-offs. ARG thus separately controls alignment and recovery information and their relative influence over time.

AI USE STATEMENT

Generative AI tools assisted methodology and experiment design, implementation and debugging, experiment orchestration, algebraic derivation, literature research, result interpretation, figure code, and manuscript drafting and revision. Benchmark images were produced by the evaluated SDXL pipelines, and reported numerical results were computed from generated outputs or archived evaluation records using executable code, rather than supplied as AI-written estimates. The authors remain responsible for independently checking the references, mathematics, implementation, data provenance, reported results, and scientific claims, and for the final content and any required disclosures.

ETHICS STATEMENT

ARG modifies the inference of a pretrained text-to-image generator and therefore inherits its representational biases, potential copyright and privacy concerns, and risks of deceptive or harmful image generation. Improved prompt adherence is not a safety mechanism and may also make undesirable requests easier to realize. Any deployment should retain appropriate content safeguards, provenance measures, and human oversight. Our experiments evaluate image quality and text alignment using COCO captions and automated metrics, without establishing improvements in fairness, safety, factual accuracy, or consent. Learned preference scores are not a substitute for human assessment, and synthetic outputs should not be presented as evidence of real events.

REPRODUCIBILITY STATEMENT

The model, sampling rule, datasets, and evaluation protocol are specified in Sections 3 and 4.1, with default settings and explicit runtime flags in Appendix A. Appendix C provides the inference algorithm and derivations, while Appendix D documents the paired analyses, sensitivity study, and timing protocol. The accompanying materials contain the archived inference source, per-prompt reward scores, score-analysis and figure-generation scripts, and source-file hashes. Published baseline values are distinguished from locally generated matched controls, and bootstrap intervals describe prompt-level uncertainty conditional on fixed outputs. The available source snapshots do not constitute a complete environment lock or identify an immutable model revision, as discussed in Appendix A.4, so bitwise reproducibility across hardware and software configurations is not claimed.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 134462–134490, 2026.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 30824–30850, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Masud An Nur Islam Fahim, Nazmus Saqib, and Joon-Min Gil. Oseg: Improving diffusion sampling through orthogonal smoothed energy guidance. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5996–6005. IEEE, 2026.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Chaofan Gan, Zicheng Zhao, Yuanpeng Tu, Xi Chen, Ziran Qin, Tieyuan Chen, Mehrtash Harandi, and Weiyao Lin. Massive activations are the key to local detail synthesis in diffusion transformers. In *International Conference on Learning Representations*, volume 2026, pp. 104855–104883, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37:66743–66772, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7428–7437. IEEE, 2023.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 154028–154053, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

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.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022.

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. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. *Advances in Neural Information Processing Systems*, 38:67153–67175, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 76833–76858, 2025a.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M Weber. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales. *arXiv preprint arXiv:2506.19713*, 2025b.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

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.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4733–4743. IEEE, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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*, 2020b.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024.

Dian Xie, Shitong Shao, Lichen Bai, Bojun Cheng, Shuo Yang, JUN WU, Zeke Xie, et al. Guidance matters: Rethinking the evaluation pitfall for text-to-image generation. In *International Conference on Learning Representations*, volume 2026, pp. 52047–52079, 2026.

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.

Weijia Zhang, Yuehao Liu, Shanyan Guan, Wu Ran, Yanhao Ge, Wei Li, and Chao Ma. Guiding a diffusion model by swapping its tokens. *arXiv preprint arXiv:2604.08048*, 2026.

Appendix

A Experimental settings

A.1 Generation and implementation

All ARG and controlled component experiments use SDXL-base-1.0 (Podell et al., 2024) at 1024×1024 resolution, with an fp16 U-Net, a bf16 VAE, and 50 Euler steps under the epsilon parameterization (Karras et al., 2022). The two text encoders are evaluated once per prompt, after which the token embeddings, pooled embeddings, and SDXL size conditions are shared by the original and self-swap branches without an unconditional CFG branch. Initial noise is generated from seed $1000i$ for prompt index i and multiplied by the scheduler's initial noise scale, while the scheduler's own input scaling is retained at every step. The final latent is divided by the VAE scaling factor before decoding, and the decoded image is mapped to $[0, 1]$, clipped, and rounded to an 8-bit PNG. Table 4 lists the shared benchmark defaults, except for stated control settings.

Following the public SSG implementation (Zhang et al., 2026), perturbations are applied within the first six `BasicTransformerBlock` modules encountered in the U-Net up-blocks. For features with T spatial tokens and D channels, the auxiliary branch selects $\lfloor 0.1T \rfloor$ spatial pairs and then $\lfloor 0.1D \rfloor$ channel pairs from the lowest cosine similarities in the corresponding upper-triangular similarity matrices. Because these counts refer to tokens and channels rather than all possible pairs, and selected pairs may overlap, the original scatter/gather indexing is retained instead of replacing it with a disjoint permutation. Both branches receive the same prompt, allowing their prediction difference to describe the response to self-swap perturbation under a shared condition.

SSA filters this difference with a separable 13-tap Gaussian of standard deviation $\sigma_g = 3$ and applies temperature scaling to all 70 cross-attention layers, while the remaining attention normalization, residual connections, and output scaling are unchanged. DCR uses the same filter construction with $\sigma_u = 2$ on the denoised latent estimate, with both convolutions evaluated in float32 using zero padding before conversion to the prediction dtype. The default strengths are $s_0 = 6$, $\tau_0 = 0.5$, and $\lambda = 0.1$, while $p_0 = 0.5$ and $p_d = 0.7$ determine the onset of recovery and annealing. With progress $p = i/49$, recovery is active at indices 25–49 and annealing at indices 35–49, so that recovery and unreduced alignment overlap for ten steps before alignment is relaxed.

Table 4: Default SDXL and ARG settings for both COCO evaluation protocols. Only the specified factor varies in controlled studies.

Item	Default
Checkpoint / resolution	SDXL-base-1.0 / 1024×1024
UNet / VAE precision	fp16 / bf16
Sampler / steps / CFG	Euler discrete / 50 / disabled
Prompt seed	$1000i$ for prompt index i
Self-swap depth / initial scale	first six up-block transformer modules / $s_0 = 6$
Swap selection	$\lfloor 0.1T \rfloor$ spatial and $\lfloor 0.1D \rfloor$ channel pairs
SSA spectral filter	13-tap separable Gaussian, $\sigma_g = 3$
SSA attention / annealing	all 70 `attn2` layers, $\tau_0 = 0.5$, $p_d = 0.7$
DCR	$\lambda = 0.1$, $\sigma_u = 2$, $p_0 = 0.5$
Distributional evaluation resolution	256×256
ImageReward / PickScore inputs	original PNGs, evaluator-specific preprocessing

A.2 Benchmarks and evaluation

The COCO-2014 and COCO-2017 protocols contain 30,000 and 5,000 prompts, respectively, with one image generated for each caption in the fixed benchmark order. Matching validation images provide the reference distributions after resizing to 256×256, and identical preprocessing is used across ARG and the controlled local variants when computing FID (Heusel et al., 2017), IS (Salimans et al., 2016), CLIP score (Hessel et al., 2021), precision and recall (Kynkäänniemi et al., 2019), AES (Schuhmann et al., 2022), ImageReward (Xu et al., 2023), and PickScore (Kirstain et al., 2023). For ImageReward and PickScore, the original generated PNGs are passed to each

evaluator's own image preprocessing rather than a shared 256×256 input transform. As the COCO validation protocols overlap in image content, their results are not interpreted as evidence of transfer between independent domains. Distributional scores are computed over each complete output set rather than averaged over repeated generation seeds, and learned preference scores are considered alongside the other criteria because a single evaluator may favor stronger guidance despite visible artifacts (Xie et al., 2026).

For paired studies, prompt order and generation seeds are held fixed, and uncertainty in the mean score difference is estimated by 10,000 percentile bootstrap resamples of matched prompt indices. All appearances of the same comparison reuse its stored bootstrap results, including the DCR intervals in the main text and Table 7. These intervals describe prompt-level variability conditional on the generated outputs rather than uncertainty across independent seeds, hardware platforms, or human raters, and no correction for multiple-metric testing is applied. Before evaluation, filenames and run manifests are checked for missing, duplicate, or unexpected prompt indices, while published comparison rows and historical local replications remain distinguished from the matched controls because their operating points or seed rules differ.

A.3 Computational protocol

ARG retains two conditional U-Net branches per step and introduces neither additional denoiser evaluations nor trainable parameters, although its prediction filter is evaluated at every step and its recovery filter is additionally evaluated once recovery begins. The historical 30k run was deterministically sharded over eight GPUs by prompt index, with model, dataset, and metric caches on the shared filesystem, whereas the cost study in Appendix D.5 isolates inference on one A800-SXM4-80GB. That study uses 20 prompts, two warmups per variant, and a shuffled variant order within each prompt, measuring encoding, denoising, VAE decoding, and the 8-bit CPU transfer while excluding loading, processor setup, and PNG writing. Native attention and a common attention processor are reported separately so that the processor cost is not attributed to SSA, DCR, or TH, and peak allocated memory refers specifically to PyTorch allocations rather than total device usage.

A.4 Explicit run configuration and artifact provenance

Reproducing ARG requires explicit settings, since the archived script defaults disable recovery and contrast annealing. Table 5 therefore records the explicit values corresponding to ARG, rather than relying on the script defaults. In particular, `SHARP_ANNEAL_SCALE=1` enables contrast annealing and `UNSHARP_LAM=0.1` enables recovery, while the temperature endpoints and decay onset determine the attention schedule. Prompt indices must refer to the full caption list when using `--start_idx` and `--end_idx`, since the generation seed is computed from the global index rather than the position within a shard. Empty-line filtering and caption order must also remain unchanged, otherwise the nominally identical seed rule would be applied to different prompts.

Table 5: Explicit environment values for the archived ARG inference entry point. The command-line settings additionally specify 50 steps, scale 6, the model location, caption list, and output directory.

Variable	Value	Variable	Value
SHARP_TAU_HIGH	0.5	SHARP_TAU_LOW	1.0
SHARP_DECAY_START	0.7	SHARP_ANNEAL_SCALE	1
SHARP_TARGET	attn2	FLG_SIGMA	3.0
UNSHARP_LAM	0.1	UNSHARP_SIGMA	2.0
UNSHARP_START	0.5		

The accompanying source snapshot retains the inference entry point and its pipeline dependency, while the robustness analysis reads the four archived per-prompt score files without regenerating images or changing their scores. It verifies the complete index sequence before computing differences in float64, and records input-file hashes, aggregation rules, and full-precision results in a machine-readable report. The figure manifest separately preserves image-source hashes, with the supplementary qualitative grids embedded at their native 1024×1024 resolution. The compact main-text comparison uses resized display copies, which are not used for metric evaluation. These records identify the supplied artifacts but do not replace a checkpoint revision or an environment

lockfile, neither of which is established by the manuscript compilation package alone, so bitwise reproduction across software and hardware configurations is not asserted.

B Additional related work

B.1 Attention control and semantic grounding

Cross-attention has been used to connect text tokens with spatial image features in both editing and generation, although the desired intervention depends on whether an existing composition must be preserved or a neglected concept must be introduced. Prompt-to-Prompt (Hertz et al., 2022) reuses and modifies attention maps to preserve structure during text-guided editing, while Structured Diffusion Guidance (Feng et al., 2022) incorporates linguistic structure to improve attribute binding and composition. By optimizing latent variables to increase attention to neglected subjects, Attend-and-Excite (Chefer et al., 2023) addresses semantic omissions through an explicit refinement objective, which differs from changing the magnitude of a global prediction contrast.

Although attention maps can reveal word–image dependencies, as studied through DAAM (Tang et al., 2023), a more concentrated map does not by itself establish that the association is correct. ARG therefore uses temperature control as part of a joint sampling rule rather than as a standalone certificate of grounding, with the same scalar temperature applied in both conditional branches without linguistic parsing or a backward pass. Whereas the methods above impose more specific editing or grounding constraints, the present study evaluates whether concentrated conditioning, spectrally restricted guidance, and denoised-content recovery can jointly improve the generated images without changing the pretrained model.

B.2 Guidance refinement and temporal allocation

Beyond constructing a negative reference, sampling quality can be affected by how guidance is represented and when it is applied, motivating geometric refinements such as CFG++ (Chung et al., 2025), adaptive projected guidance (Sadat et al., 2024), and frequency-dependent guidance (Sadat et al., 2025b). Feature-level adjustments such as FreeU (Si et al., 2024) offer a complementary perspective on the balance between structure and detail, whereas guidance schedules (Wang et al., 2024), limited-interval guidance (Kynkäänniemi et al., 2024), and analyses of stage-dependent denoising (Jin et al., 2026) motivate varying the strength of an intervention along sampling. In ARG, this temporal dependence coordinates a perturbation-derived alignment signal with recovery from the denoised estimate, rather than treating Gaussian filtering, temperature scaling, or scheduling alone as the methodological novelty. Because the resulting changes can affect distributional fidelity, prompt alignment, and learned preferences differently, the supplementary controls report these criteria together instead of reducing the comparison to a single score.

C Theoretical analysis and algorithms

C.1 Complete sampling procedure

Algorithm 1 makes explicit how the three components enter the existing Euler sampler. The original and self-swap predictions are evaluated as a batched pair under the same conditioning, after which SSA determines the guided prediction and DCR modifies its denoised estimate only when $p \geq p_0$. The schedule in Eq. (6) is applied before the branch evaluations so that attention temperature and contrast strength refer to the same sampling progress, while the original scheduler retains responsibility for input scaling and the Euler step. Recovery is evaluated only at positive noise levels, including the last denoiser evaluation before the scheduler reaches zero.

Algorithm 1 Align–Recover Guidance with an Euler scheduler

Input Prompt c, initial noise $\mathbf{z}$, N timesteps, noise levels $\sigma_0, \ldots, \sigma_N$, scheduler initialization scale σ_{init}
Output Generated image $\mathbf{y}$

Encode c once and share all text and size conditions across both branches
$\mathbf{x} \leftarrow \sigma_{\text{init}}\mathbf{z}$
for $i = 0, \ldots, N-1$ **do**
  $p \leftarrow i/(N-1)$ and $(s, \tau) \leftarrow (s(p), \tau(p))$ from Eq. (6)
  $\mathbf{u} \leftarrow \text{ScaleInput}(\mathbf{x}, t_i)$
  $(\boldsymbol{\epsilon}^o, \boldsymbol{\epsilon}^s) \leftarrow \text{DenoisePair}(\mathbf{u}, t_i, c, \tau)$
  $\bar{\boldsymbol{\epsilon}} \leftarrow \boldsymbol{\epsilon}^o + sG_{\sigma_g} * (\boldsymbol{\epsilon}^o - \boldsymbol{\epsilon}^s)$
  $\boldsymbol{\epsilon} \leftarrow \bar{\boldsymbol{\epsilon}}$
  if $p \geq p_0$ **then**
    $\widehat{\mathbf{x}}_0 \leftarrow \mathbf{x} - \sigma_i\bar{\boldsymbol{\epsilon}}$
    $\mathbf{r} \leftarrow \widehat{\mathbf{x}}_0 - G_{\sigma_u} * \widehat{\mathbf{x}}_0$
    $\boldsymbol{\epsilon} \leftarrow [\mathbf{x} - (\widehat{\mathbf{x}}_0 + \lambda\mathbf{r})]/\sigma_i$
  end if
  $\mathbf{x} \leftarrow \mathbf{x} + (\sigma_{i+1} - \sigma_i)\boldsymbol{\epsilon}$
end for
return $\text{Decode}(\mathbf{x}/\,\text{VAEscale})$

C.2 Exact recovery displacement under Euler sampling

To isolate the influence of DCR, consider one step at a fixed latent $\mathbf{x}_i$ and a fixed SSA prediction $\bar{\boldsymbol{\epsilon}}_i$, with $\sigma_i > 0$ and $0 \leq \sigma_{i+1} \leq \sigma_i$. Writing $H = I - G_{\sigma_u}$ and $\widehat{\mathbf{x}}_0 = \mathbf{x}_i - \sigma_i\bar{\boldsymbol{\epsilon}}_i$ separates the content residual from the changing denoiser, allowing the induced displacement to be characterized without assuming that semantics and texture are independent.

Proposition 1 (Single-step recovery displacement)**.** *Let $\mathbf{x}_{i+1}^A$ and $\mathbf{x}_{i+1}^R$ denote the Euler updates without and with DCR from the same latent and SSA prediction. Then*

$$\mathbf{x}_{i+1}^R - \mathbf{x}_{i+1}^A = \lambda\eta_i H\widehat{\mathbf{x}}_0, \qquad \eta_i = \frac{\sigma_i - \sigma_{i+1}}{\sigma_i} \in [0, 1]. \tag{9}$$

Proof. Substitution of the recovered estimate into Eq. (5) gives $\boldsymbol{\epsilon}_i^R - \bar{\boldsymbol{\epsilon}}_i = -\lambda H\widehat{\mathbf{x}}_0/\sigma_i$. Multiplying this difference by the common Euler step $\sigma_{i+1} - \sigma_i$ yields Eq. (9), while the stated range of η_i follows directly from the nonnegative decreasing noise levels. □

Although the recovered prediction contains a division by σ_i, its contribution to the Euler displacement is scaled by $\lambda\eta_i$ because the solver step changes with the noise level. At the terminal step, where $\sigma_{i+1} = 0$, the next latent is exactly $\widehat{\mathbf{x}}_0 + \lambda H\widehat{\mathbf{x}}_0$ under this update, so the expression does not require evaluating a denoiser or dividing by zero at the terminal noise level. These identities hold at a fixed state, whereas the altered latent affects subsequent predictions throughout sampling.

Proposition 2 (Finite-step norm bound)**.** *For the normalized nonnegative Gaussian used in the implementation, including its zero-padded finite-image realization, and $\lambda \geq 0$,*

$$\|\mathbf{x}_{i+1}^R - \mathbf{x}_{i+1}^A\|_2 \leq \lambda\eta_i\|H\widehat{\mathbf{x}}_0\|_2 \leq 2\lambda\eta_i\|\widehat{\mathbf{x}}_0\|_2. \tag{10}$$

Proof. Zero extension and restriction do not increase the Euclidean norm, while convolution by a nonnegative kernel with unit sum has ℓ_2 operator norm at most one by Young's inequality. Consequently, $\|G_{\sigma_u}\mathbf{z}\|_2 \leq \|\mathbf{z}\|_2$ and $\|H\mathbf{z}\|_2 \leq \|\mathbf{z}\|_2 + \|G_{\sigma_u}\mathbf{z}\|_2 \leq 2\|\mathbf{z}\|_2$. Applying these inequalities to Eq. (9) gives the stated bound. □

The bound is deliberately conservative and holds in exact arithmetic for the specified linear filter, without asserting that the denoised estimate remains bounded over an arbitrary nonlinear trajectory or that contrast amplification improves image quality. Zero padding also means that constant inputs need not be preserved near image boundaries, so a global claim of exact DC preservation would not follow from kernel normalization alone.

C.3 Content recovery and attention concentration

The distinction between recovering content and restoring the discarded prediction contrast can be seen by expanding the residual before recovery. If $\mathbf{d}_i = \boldsymbol{\epsilon}_i^o - \boldsymbol{\epsilon}_i^s$, then linearity of the filters gives

$$H\widehat{\mathbf{x}}_0 = H\mathbf{x}_i - \sigma_i H\boldsymbol{\epsilon}_i^o - \sigma_i s_i HG_{\sigma_g}\mathbf{d}_i. \tag{11}$$

Because this expression also depends on the current latent and original prediction, it is not generally a fixed multiple of $H\mathbf{d}_i$, which would instead restore high-frequency variation in the perturbation response. Recovery therefore acts on the current content estimate, although it cannot reconstruct missing ground-truth detail from this algebraic relation alone.

For attention, a complementary statement can be made by holding one row of logits ℓ_j fixed and writing $a_j(\tau) = \exp(\ell_j/\tau)/Z(\tau)$. Its entropy $\mathcal{H}(\tau)$ satisfies

$$\mathcal{H}(\tau) = \log Z(\tau) - \frac{\mathbb{E}_{a(\tau)}[\ell]}{\tau}, \qquad \frac{d\mathcal{H}}{d\tau} = \frac{\mathrm{Var}_{a(\tau)}(\ell)}{\tau^3} \geq 0. \tag{12}$$

Indeed, differentiating the log partition function gives $d\log Z/d\tau = -\mathbb{E}_a[\ell]/\tau^2$, while differentiating the expectation gives $d\mathbb{E}_a[\ell]/d\tau = -\mathrm{Var}_a(\ell)/\tau^2$, from which the two expectation terms cancel. A lower temperature thus reduces entropy for fixed logits, providing a precise meaning for concentration without implying correct semantic associations or monotonic entropy during sampling.

C.4 Paired uncertainty estimation

Algorithm 2 describes the prompt-level resampling used for matched controls. Resampling the difference rather than the two score vectors independently preserves each prompt's pairing, while reusing the stored output of a fixed random stream prevents the same comparison from acquiring slightly different intervals across sections. The procedure is applied separately to each reported metric, and positive-difference fractions distinguish mean gains from improvement frequency.

Algorithm 2 Paired-prompt bootstrap for a fixed generated sample

Input Matched scores $\{a_i, b_i\}_{i=1}^M$, $B = 10{,}000$, fixed bootstrap random stream
Output Mean difference $\bar{\delta}$, percentile interval $[L, U]$, positive fraction q

$\delta_i \leftarrow b_i - a_i$ for all i
$\bar{\delta} \leftarrow M^{-1}\sum_i \delta_i$ and $q \leftarrow M^{-1}\sum_i \mathbb{1}[\delta_i > 0]$
for $b = 1, \ldots, B$ **do**
    Draw M indices $I_1, \ldots, I_M$ uniformly with replacement from $\{1, \ldots, M\}$
    $v_b \leftarrow M^{-1}\sum_{j=1}^M \delta_{I_j}$
end for
$(L, U) \leftarrow (\mathrm{Quantile}_{0.025}(v), \mathrm{Quantile}_{0.975}(v))$
Store and reuse $(\bar{\delta}, L, U, q)$ for every report of this comparison

D Additional experimental results

D.1 Spectral response and temporal coordination

Across 500 matched prompts, DCR raises the mean high-frequency energy fraction by 132.7% relative to SSA (from 3.15% to 7.33%), while ARG adds only 0.4% relative to SSA+DCR, supporting the interpretation that recovery strengthens local contrast without TH simply undoing the added high-frequency content. Fig. 7 combines radial spectra, prompt-level distributions, and matched energies so that the aggregate change can be related to its variation across images. Since sharp artifacts can also increase spectral energy, the diagnostic is considered together with aesthetic, distributional, and preference measurements rather than as an independent perceptual-quality score.

The temporal view in Fig. 8 explains how this recovery remains active while alignment is reduced. Recovery begins at step 25, followed by ten steps of overlap with unreduced alignment before annealing begins at step 35, so the final fifteen steps progressively weaken the contrast scale and attention concentration without abruptly switching between separate sampling regimes. The activity map reports the discrete schedule actually evaluated over the 50 steps, whereas the parameter-space path makes the ordering of these changes visible.

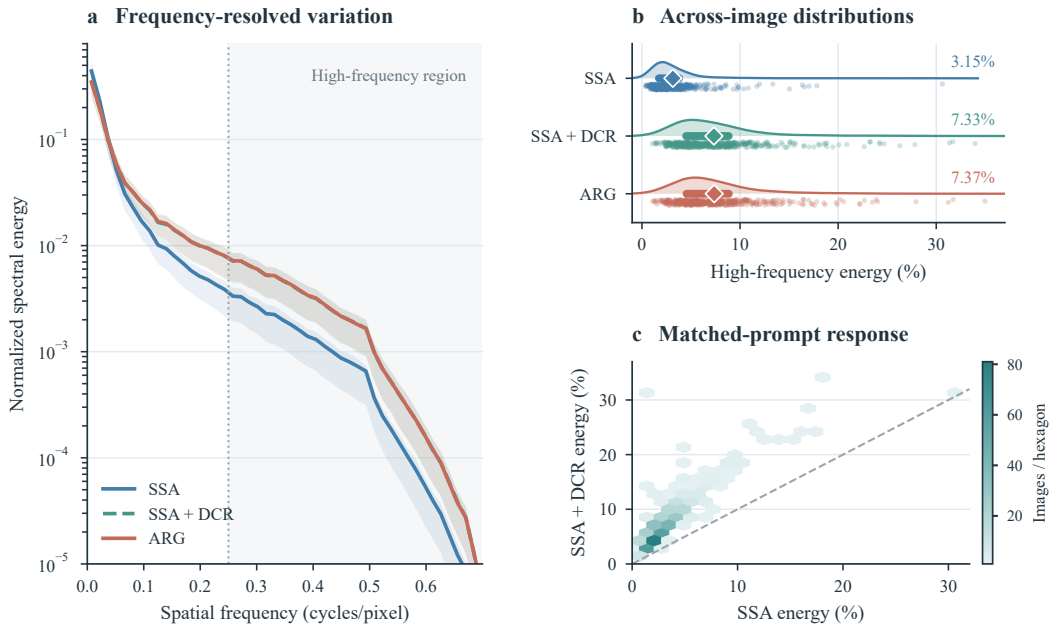

Fig. 7: Spectral changes across 500 matched COCO-2017 prompts. (a) Mean radial spectra with interquartile bands across images. (b) Half-violin densities, individual images, interquartile bars, and mean diamonds for energy above 0.25 cycles/pixel. (c) Paired SSA and SSA+DCR energies, hexagon color indicates image count and the dashed line marks equality. Spectra use mean-centered, Hann-windowed luminance images resized to 256×256 and normalized per image.

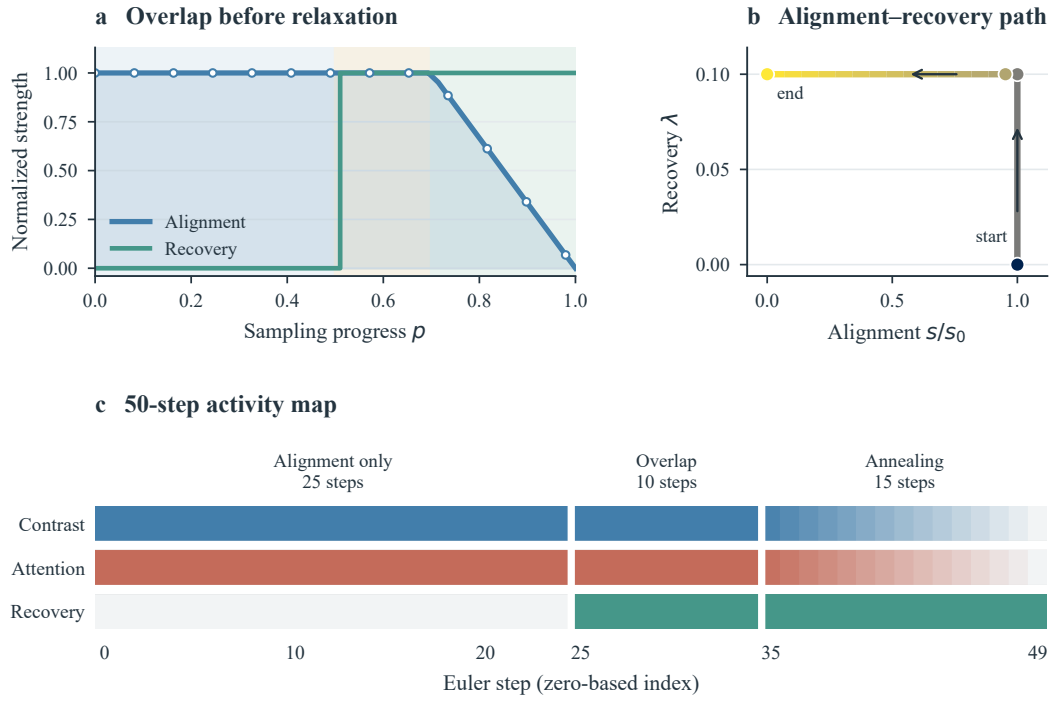

Fig. 8: Temporal coordination over 50 Euler steps. (a) Normalized alignment and binary recovery activity. (b) The trajectory in contrast-scale and recovery-strength space, colored from early (dark) to late (light). (c) Per-step activity, with contrast normalized by s_0 and attention concentration by $1 - \tau_0$. Recovery begins at index 25 and annealing at index 35, producing 25 alignment-only, 10 overlap, and 15 annealing steps.

D.2 Quantitative comparisons and component controls

Table 6 preserves the additional metrics available for the historical local replications, among which ARG achieves lower FID and higher recall on both COCO protocols. Because the historical SSG and spectral-only runs used different seed rules, these results complement the benchmark comparison but cannot identify a paired component effect, for which the matched controls below are required.

Table 6: Quantitative results for historical local SSG/SSA-S runs and ARG on COCO-2014 and COCO-2017. The historical runs use different seed rules and are not paired component controls.

Benchmark	Method	FID↓	Prec↑	Rec↑	IR↑	PS↑
COCO-2014 (30k)	SSG	21.715	0.466	0.447	0.274	22.16
	ARG	**16.472**	**0.487**	**0.546**	**0.519**	**22.20**
COCO-2017 (5k)	SSG	31.755	0.502	0.552	0.251	22.14
	SSA-S	27.449	**0.569**	0.565	0.226	22.07
	ARG	**26.673**	0.533	**0.639**	**0.515**	**22.20**

In the matched study, all five settings share the prompt seeds $1000i$ and a common attention processor, with 5000 validated outputs per setting. The supplementary spectral-only control, denoted SSA-S in Fig. 9, sets $\tau = 1$ while retaining the filtered contrast, thereby separating attention sharpening from the spectral restriction within SSA rather than introducing a fourth ARG component. The eight-metric profile shows how these conditional increments affect different criteria, while the precision–recall and AES–ImageReward views retain the original metric scales rather than implying that normalized scores are directly comparable across metrics. These sequential comparisons do not constitute a factorial analysis, and TH changes contrast strength and attention temperature together.

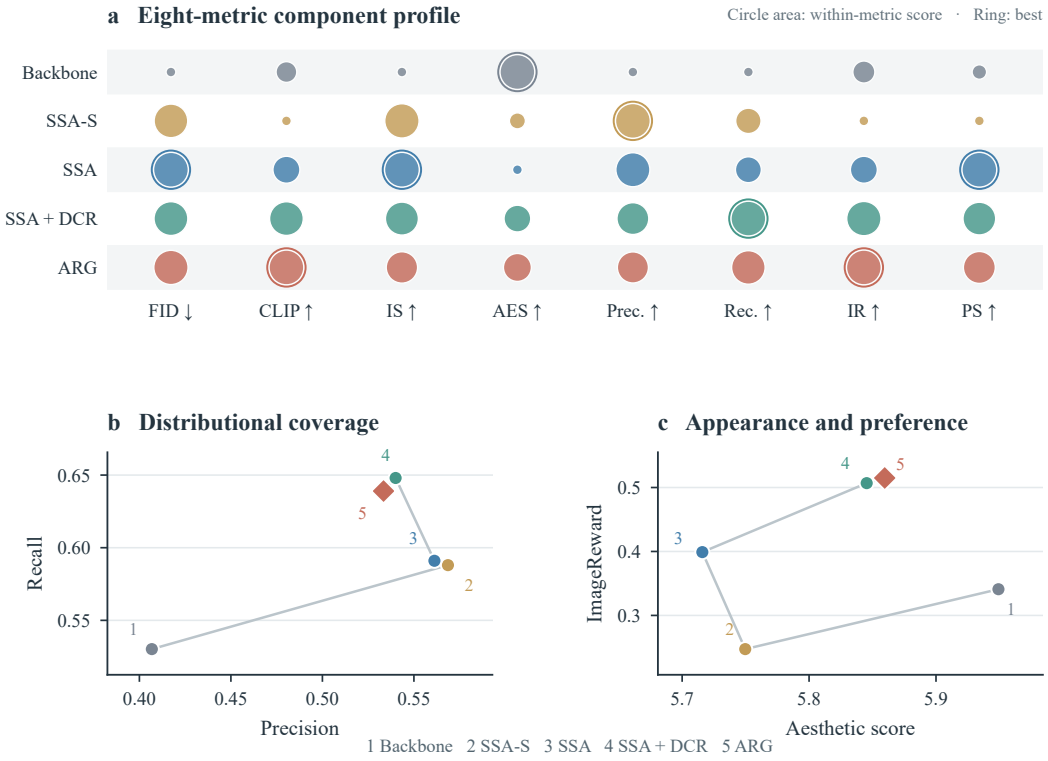

Fig. 9: Component profiles and metric tradeoffs for five matched settings. (a) Circle area increases with the within-metric normalized score (FID reversed), outer rings identify column maxima. (b) Precision–recall and (c) aesthetics–ImageReward trajectories use the original metric scales. Numbers and connecting lines follow the order in (a), without interpolation between settings.

Table 7: Paired changes in ImageReward and PickScore on 5000 COCO-2017 prompts. Intervals use 10,000 paired-prompt bootstrap resamples without multiplicity correction, the first two rows isolate the spectral-only control.

Contrast	ΔIR	95% CI	ΔPS	95% CI
Backbone → SSA-S	-0.094	$[-0.1106, -0.0767]$	-0.013	$[-0.0312, +0.0057]$
SSA-S → SSA	$+0.152$	$[+0.1344, +0.1693]$	$+0.125$	$[+0.1083, +0.1427]$
SSA → SSA+DCR	$+0.108$	$[+0.0944, +0.1211]$	-0.016	$[-0.0315, -0.0008]$
SSA+DCR → ARG	$+0.008$	$[-0.0051, +0.0215]$	-0.005	$[-0.0208, +0.0110]$

Under the shared bootstrap results in Table 7, adding DCR to SSA increases mean ImageReward by 27.0% and decreases mean PickScore by 0.07%, with absolute score-difference intervals of $[0.0944, 0.1211]$ and $[-0.0315, -0.0008]$, respectively. Positive changes occur on 68.44% and 49.20% of prompts, respectively, including 1889 prompts on which both improve, 1533 on which only ImageReward improves, 571 on which only PickScore improves, and 1007 on which neither increases. By relating the mean changes to their joint distribution in Fig. 5, these counts reveal a substantive average recovery gain under ImageReward without suggesting agreement between the two learned evaluators. TH's smaller 1.6% mean ImageReward increase has an absolute score-difference interval of $[-0.0051, 0.0215]$, which includes zero, while independent score re-evaluation of the ARG outputs reproduces the reported IR of 0.515 and PickScore of 22.195.

D.3 Central gains and sensitivity to score tails

A positive mean difference may reflect a broadly shared improvement or a smaller number of large gains, which motivates examining the center and tails of the matched score distribution rather than repeating the aggregate comparison. For each of the 5000 prompts, we subtract the earlier setting's score from the later setting's score and report the median alongside the mean, a 5% trimmed mean, and the fraction of strictly positive differences in Table 8. Trimming removes 250 observations from each tail of the sorted differences separately for each comparison and evaluator, leaving 4500 observations, while the 5th–95th percentile range describes variation among individual prompts rather than uncertainty in an estimated mean. All summaries reuse the generated outputs, adding a distributional check without new sampling or replacing bootstrap intervals.

Table 8: Robust summaries of paired score changes over 5000 COCO-2017 prompts. The trimmed mean removes the lowest and highest 5% separately. The 5th–95th percentile range describes individual prompt differences, not a confidence interval.

Comparison	Metric	Mean	Median	Trimmed	5th–95th	Positive
Backbone → SSA	IR	+0.058	-0.004	+0.041	$[-0.947, 1.394]$	49.38%
	PS	+0.112	+0.098	+0.111	$[-1.039, 1.358]$	57.10%
SSA → SSA+DCR	IR	+0.108	+0.092	+0.110	$[-0.498, 0.787]$	68.44%
	PS	-0.016	-0.005	+0.000	$[-0.689, 0.700]$	49.20%
SSA+DCR → ARG	IR	+0.008	-0.002	-0.001	$[-0.626, 0.637]$	49.08%
	PS	-0.005	-0.025	-0.025	$[-0.688, 0.657]$	45.76%
Backbone → ARG	IR	+0.174	+0.090	+0.152	$[-0.812, 1.488]$	60.60%
	PS	+0.091	+0.070	+0.083	$[-1.091, 1.305]$	54.62%

For recovery, the ImageReward median is 0.092 and the trimmed mean is 0.110, close to the untrimmed mean of 0.108, indicating that its average gain does not disappear after removing extreme differences in either direction. The complete ARG comparison with Backbone likewise remains positive under both robust summaries, with ImageReward median and trimmed mean of 0.090 and 0.152, and PickScore values of 0.070 and 0.083. Since positive changes occur on 60.60% and 54.62% of prompts, respectively, these results support gains beyond a few exceptional examples without implying that every prompt improves or equating either evaluator with human judgment.

The same analysis also distinguishes the effects of individual stages. SSA has a positive ImageReward mean but a slightly negative median, whereas DCR's PickScore trimmed mean is approximately zero despite its negative untrimmed mean. TH's ImageReward mean is slightly positive while its median and trimmed mean are slightly negative, consistent with the previously reported in-

terval containing zero. These differences caution against assigning every stage the same preference benefit, and the negative lower percentiles show why the average comparison must be accompanied by adverse as well as favorable examples.

D.4 Sensitivity to recovery strength

Table 9 varies only λ under the full ARG schedule on the first 500 COCO-2017 prompts, allowing the influence of stronger recovery to be examined without changing its timing. Relative to $\lambda = 0$, the default $\lambda = 0.1$ increases ImageReward by 45.4%, AES by 2.2%, and CLIP by 0.9%, while PickScore decreases by 0.10%. Increasing λ to 0.2 further improves CLIP but lowers both learned preference scores relative to the default, indicating why recovery strength is assessed jointly across criteria rather than selected from a single evaluator. Because this subset overlaps the main evaluation and was not reserved for independent model selection, the sweep describes local sensitivity rather than establishing an optimum on unseen prompts or other backbones.

Table 9: Effect of DCR strength λ on the first 500 COCO-2017 prompts. Only λ varies under the full ARG schedule. FID is not reported for this 500-prompt sensitivity study.

λ	CLIP↑	AES↑	IR↑	PS↑
0.00	0.3135	5.697	0.282	22.060
0.05	0.3149	5.747	0.358	**22.095**
0.10	0.3164	5.824	**0.410**	22.038
0.15	0.3184	**5.883**	0.400	21.879
0.20	**0.3220**	5.818	0.388	21.674

D.5 Inference cost

Fig. 10 reports all 120 measurements from the six-variant timing study under the protocol in Appendix A. ARG requires a mean of 5.794 seconds per image, with a mean paired overhead of 0.11% relative to the common-processor Backbone and 5.68% relative to native attention, separating the cost of the three components from that of the attention implementation. The ARG timings span 5.783–5.805 seconds, compared with 5.778–5.796 seconds for the common-processor Backbone, while peak PyTorch-allocated memory is 8.953 GiB for every record. Centered timing distributions and paired differences expose the variation within this run, although they do not provide uncertainty estimates across independently repeated runs or hardware platforms.

D.6 Interpretation of the controls and remaining scope

The component ladder identifies changes conditional on the preceding setting, which is narrower than attributing an independent effect to each component in isolation. Backbone to the spectral-only control changes the frequency content of the prediction contrast, and the subsequent SSA comparison adds attention sharpening while holding that restriction fixed. Adding DCR then tests recovery in the presence of SSA, whereas the final ARG comparison tests joint contrast and temperature annealing with recovery retained. Consequently, neither the standalone effect of DCR without SSA nor the interaction between spectral restriction and sharpening can be inferred from this sequence, and the TH comparison does not separate contrast annealing from temperature annealing. These controls therefore establish conditional effects rather than testing all component interactions.

The current evidence is restricted to SDXL, the stated Euler sampler, COCO prompts, and one fixed generation seed per prompt, although the prompt-level distributions and bootstrap analyses quantify variation within that setting. Robust trimmed means do not establish stability across independent generation seeds, transfer to another architecture, or human preference, just as the single-step bound in Appendix C.2 does not establish convergence of the nonlinear sampling trajectory. Mechanistically, recovery amplifies high-pass content already present in the estimate and may therefore strengthen an existing artifact, while concentrating attention cannot guarantee correction of a wrong token–region association. These are boundaries of the method and evidence, rather than measured failure frequencies, and further independent-seed, architecture, and human evaluations are required before broader claims can be made.

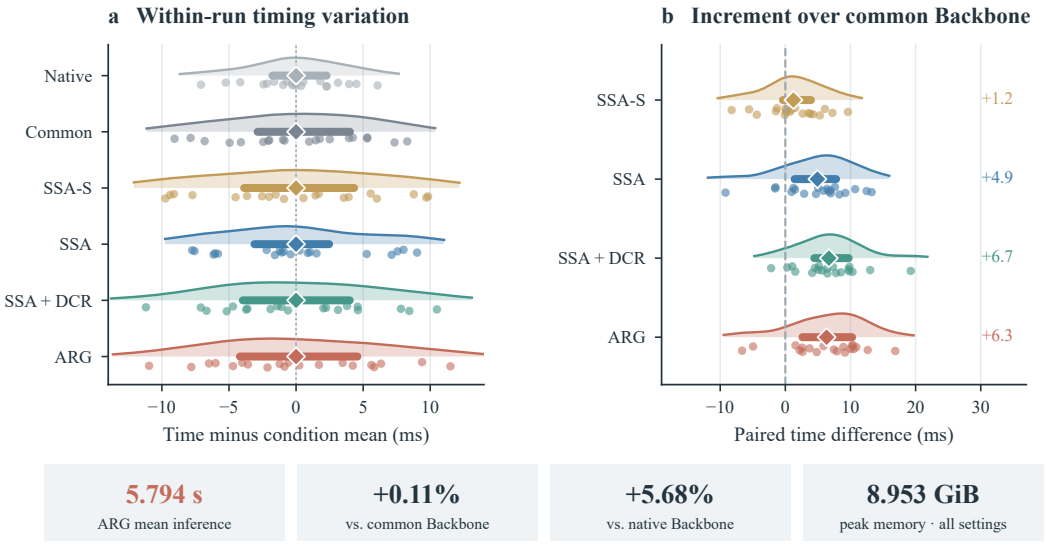

Fig. 10: Paired inference-cost measurements on one A800-SXM4-80GB, using 20 prompts per setting. (a) Timings centered at each setting's mean expose within-run variation. (b) Per-prompt time differences relative to the common-processor Backbone. Half-violins show density, dots individual prompts, thick bars interquartile ranges, and diamonds means. The summary reports ARG's absolute time, mean paired overhead, and peak PyTorch-allocated memory, identical for all 120 observations.

D.7 Ranked examples and matched component images

To complement the selected examples in the main text, Fig. 11 shows prompts nearest the 5%, 50%, and 95% ranks of the recovery-induced ImageReward difference among all 5000 matched prompts, with ties resolved by prompt index. The corresponding ARG outputs provide context but do not influence selection, allowing low, median, and high recovery gains to be inspected under the same ranking rule rather than chosen solely for visual appeal.

Fig. 12 additionally traces four matched prompts through the Backbone, spectral-only control, SSA, SSA+DCR, and ARG settings. As each row preserves both prompt and seed, changes in texture, local contrast, and composition can be related to the same sequence of interventions used in the quantitative controls, while these selected examples illustrate rather than quantify effect prevalence.

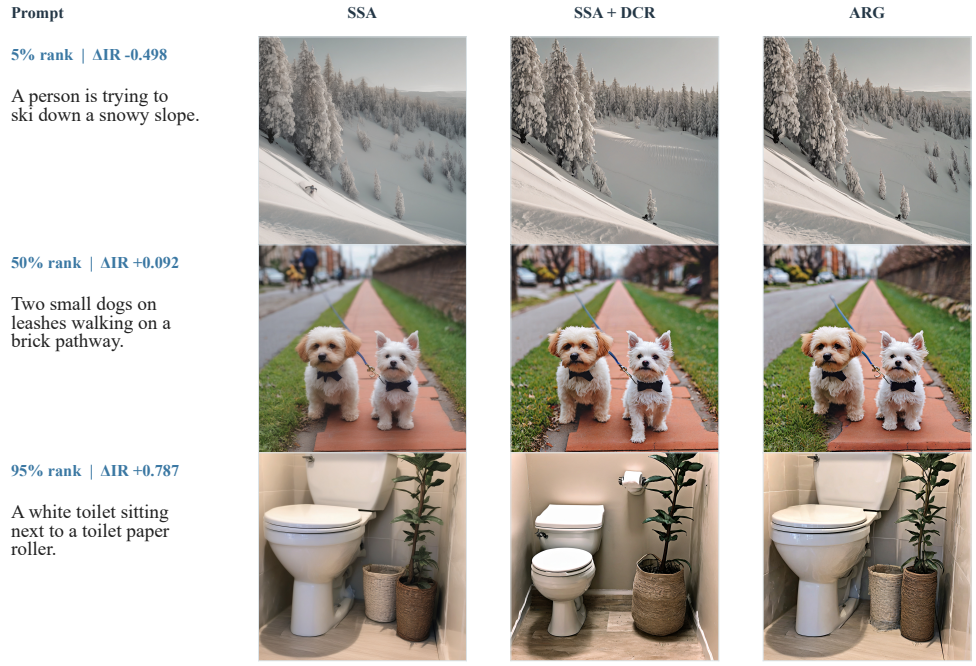

Fig. 11: Qualitative examples at low, median, and high recovery gains. Prompts are selected at the 5%, 50%, and 95% ranks of the paired ImageReward change from SSA to SSA+DCR among all 5000 prompts, with ties broken by prompt index. ARG is shown for context, not selection.

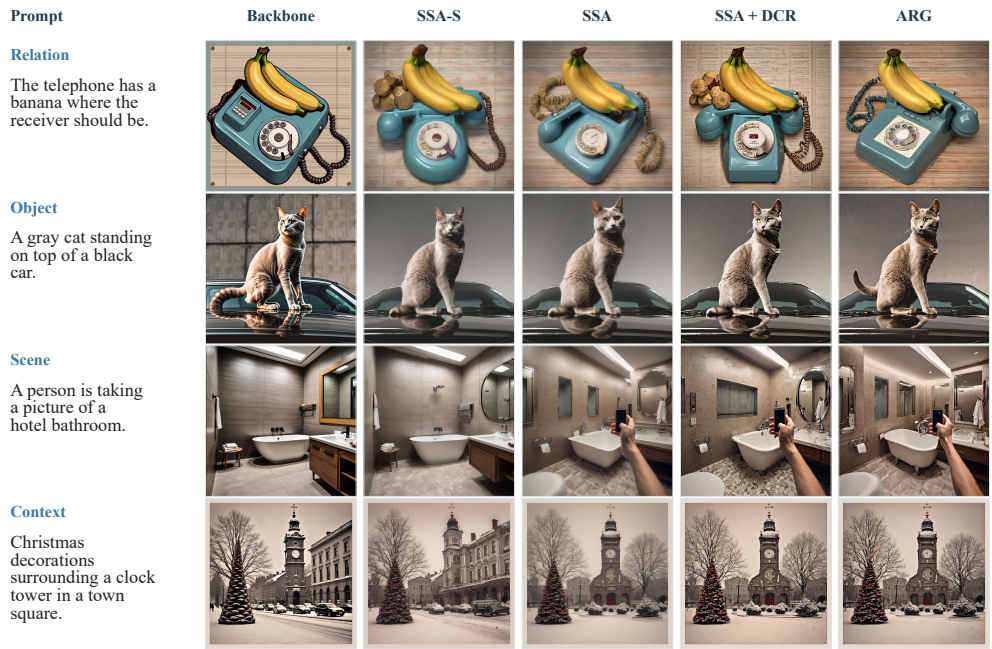

Fig. 12: Qualitative comparison of five matched component settings. The same four prompts and seeds are shown for Backbone, the spectral-only SSA-S control, SSA, SSA+DCR, and full ARG.

D.8 HISTORICAL QUALITATIVE ARCHIVE

Figs. 13 and 14 retain all 25 historical prompt pairs in index order, with full images and no retouching, replacement, or selective sharpening. As selection predates this revision and was not random, and historical SSG seeds and operating points are not verified as equivalent, these examples provide descriptive comparisons rather than same-seed ablations or failure-rate estimates. The visualization manifest preserves source hashes and rendering rules, while matched component images and quantitative analyses above provide the controlled evidence.

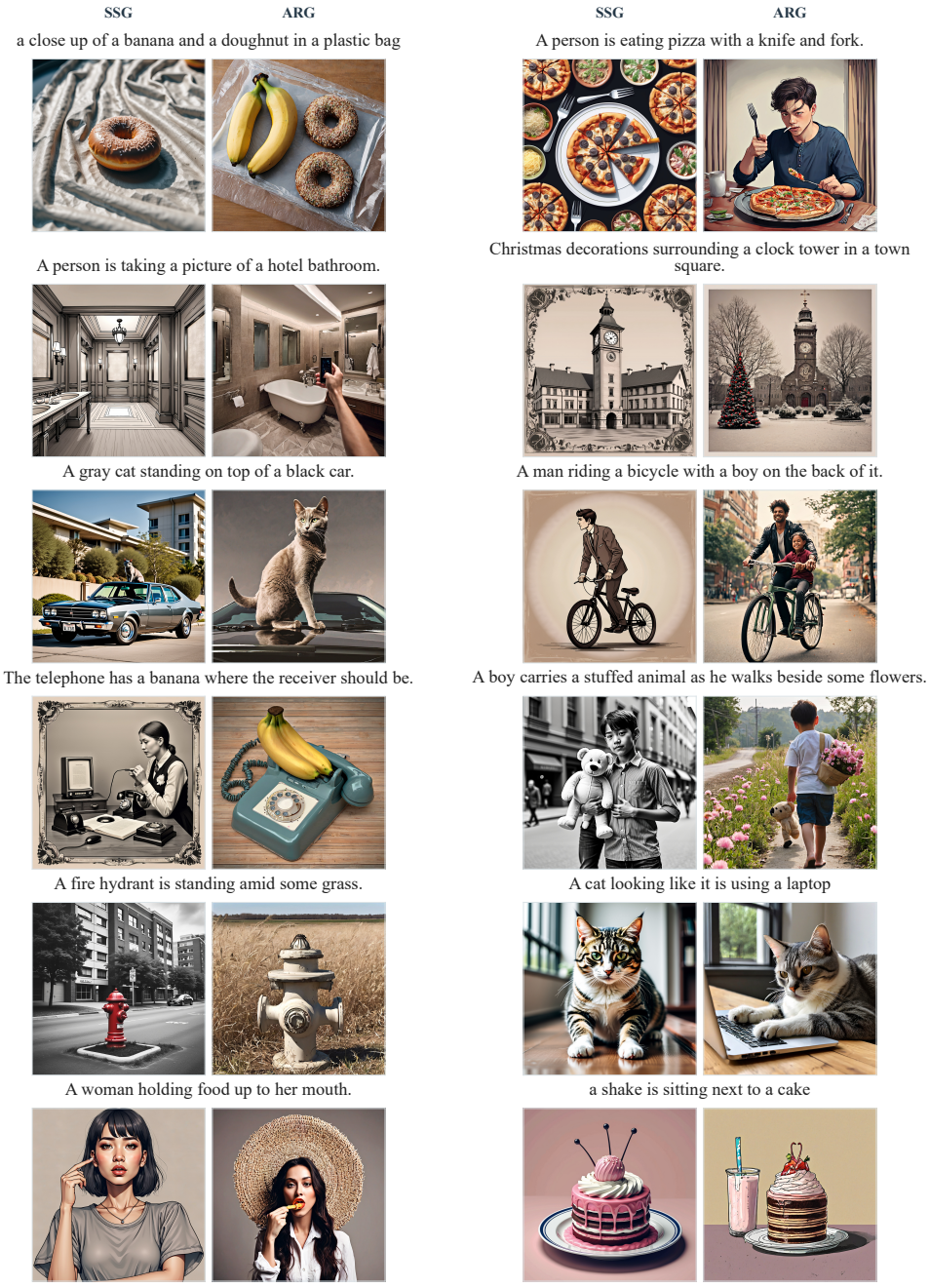

Fig. 13: Qualitative archive of historical SSG and ARG outputs, part 1. Each row contains one prompt and its two method outputs, historical seed matching has not been verified.

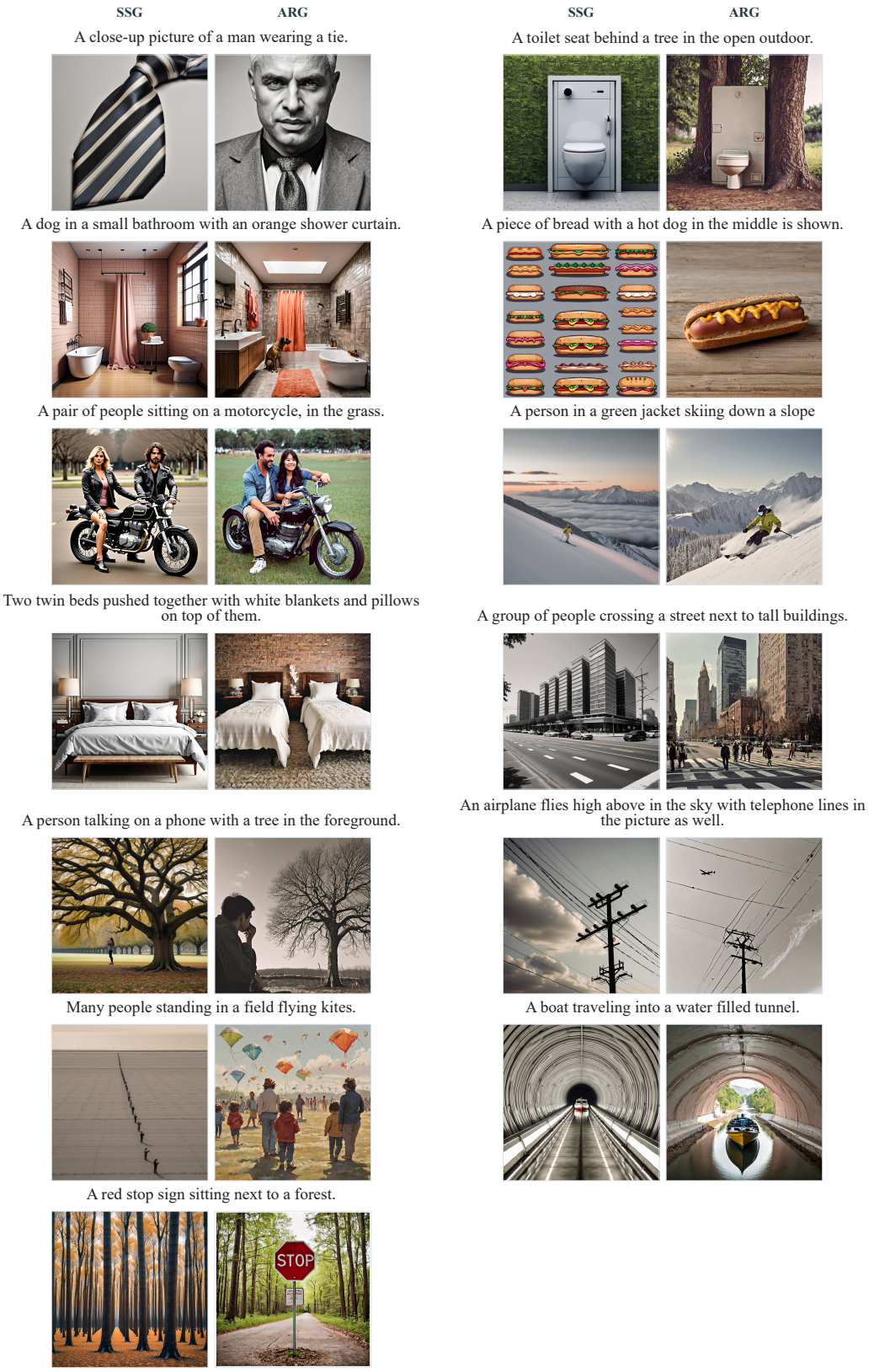

Fig. 14: Qualitative archive of historical SSG and ARG outputs, part 2. Layout and comparison protocol follow part 1, the images are descriptive rather than paired controls.