WHERE GUIDANCE MATTERS MOST: SPATIALLY ADAPTIVE CFG FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

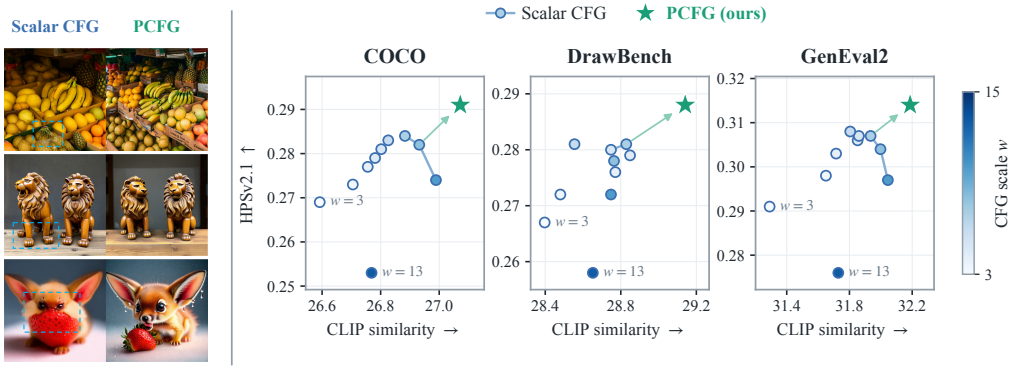


Figure 1: **PCFG on SD3.5 Medium. Left:** Scalar CFG ($w = 7$) versus PCFG at 10, 20, and 30 steps (top to bottom); blue dashed boxes mark visible artifacts. **Right:** At 20 steps, PCFG exceeds all ten scanned scalar-CFG settings in CLIP similarity and HPSv2.1 on COCO, DrawBench, and GenEval2.

ABSTRACT

Classifier-free guidance (CFG) improves prompt adherence by extrapolating beyond conditional predictions, but applies a uniform guidance scale to spatial regions that may respond differently to conditioning. We introduce pattern-based classifier-free guidance (PCFG), a training-free method that adaptively regulates further extrapolation by measuring how conditioning alters local spatial patterns. PCFG represents small patches by ordered prediction amplitudes and constructs neighborhood-level pattern distributions from their occurrence frequencies, capturing local spatial relationships beyond pointwise residual magnitudes. It then compares the corresponding conditional and unconditional distributions using JS divergence to quantify local pattern discrepancy. Finally, a bounded power mapping converts this discrepancy into spatially varying guidance scales, suppressing further extrapolation in regions with pronounced pattern changes while allowing stronger amplification where the two distributions remain similar. At each sampling step, PCFG updates the guidance map from the current conditional and unconditional predictions without additional denoiser evaluations. Experiments on SD3.5 Medium across COCO, DrawBench, and GenEval2 show that PCFG achieves the highest reported ImageReward and HPSv2.1 means in all nine benchmark–budget settings against standard CFG and seven competing guidance methods. It surpasses the metric-specific best-scanned scalar CFG in 41 of 45 comparisons; moreover, its 20-step results outperform the best-scanned 30-step scalar CFG on both ImageReward and HPSv2.1 across all three benchmarks. These results demonstrate that local pattern discrepancy provides an effective signal for spatially adaptive guidance.

1 INTRODUCTION

Classifier-free guidance (CFG) is a central inference mechanism for text-to-image diffusion and flow-matching models (Rombach et al., 2022; Esser et al., 2024). It strengthens prompt adherence by extrapolating along the difference between conditional and unconditional predictions, with

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.

a guidance scale controlling the trade-off between sample quality and diversity (Ho & Salimans, 2022). Standard fixed-scale CFG uses one weight across all spatial locations and sampling steps. Viewed spatially, this is a flat guidance map that remains unchanged throughout the sampling trajectory. Yet regions can respond differently to conditioning, so stronger global guidance can amplify some predictions unnecessarily, producing oversaturation and structural artifacts (Sadat et al., 2025a; Chung et al., 2025). This motivates adapting guidance to the evolving generation state and choosing feedback that can inform how much further each region should extrapolate.

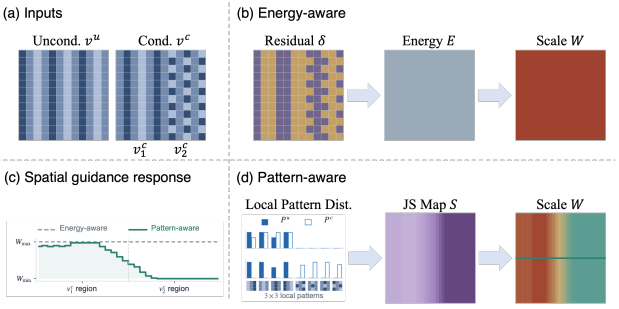


Figure 2: **Equal update strength can mask different local pattern changes.** (a) Synthetic fields v^u and v^c share a fixed unit direction; colors show amplitudes. (b) Equal-magnitude amplitude residuals yield constant channel-mean energy and uniform energy-based guidance. (c) Scale profiles along the horizontal line in (d). (d) Local pattern distributions distinguish coherent shifts (left) from alternating row-wise shifts (right); their JS divergence drives spatial guidance. This example isolates feedback discriminability.

Existing methods adapt guidance through selected noise intervals (Kynkäänniemi et al., 2024), state-dependent or online evaluation feedback (Koulischer et al., 2025; Papalampidi et al., 2026), and spatially varying scales (Shen et al., 2024; Li et al., 2026). These advances raise a central question: *what local information should determine guidance strength?* The two prediction branches provide an immediately available source of such feedback, without requiring an auxiliary evaluator or access to internal attention maps. A conservative strategy uses smaller scales where conditioning produces larger relative changes and permits stronger extrapolation where the two branches remain compatible (Zhu et al., 2025; Li et al., 2026). For example, SAMG uses the channel-mean squared conditional residual as pointwise energy feedback, attenuating guidance in high-energy regions (Li et al., 2026). Such feedback measures how much conditioning changes each location, but does not describe how neighboring prediction values are organized. Even when resolved spatially, residual energy can therefore overlook differences in local pattern changes. Effective spatial guidance calls for examining what information survives this reduction, alongside the rule that converts the resulting feedback into scales.

Figure 2 isolates this distinction in a controlled synthetic velocity field. Starting from a shared unconditional prediction, we form the conditional field by coherently shifting stripes on the left and alternating row-wise phase shifts on the right. The coherent shift largely relocates existing amplitude patterns, whereas the alternating shifts change inter-row relationships and pattern composition. Although the signed residuals have different spatial arrangements, their magnitude and energy are identical at every location. An energy-only rule with the same pointwise mapping consequently assigns uniform scales across both regions; normalizing the constant energy map cannot restore the missing distinction. Local pattern distributions remain similar on the left and differ on the right, making these changes distinguishable. The example exposes a limitation of the feedback representation: *equal update strength can mask different local pattern changes*. It motivates using changes in local organization to inform spatial guidance.

We propose pattern-based classifier-free guidance (**PCFG**), which uses discrepancies between local amplitude-pattern distributions as spatial guidance feedback. Because the conditional prediction already incorporates the prompt, PCFG regulates how far guidance extrapolates beyond it. The two branches are first mapped to a shared vocabulary of quantized amplitudes. Ordered patches define local patterns, and their occurrence frequencies within each neighborhood form a distribution. This representation retains spatial relationships inside each patch while summarizing pattern composition over its surroundings. Neighborhood frequencies describe which patterns occur without requiring their patch positions to match exactly. Comparing the two distributions with Jensen–Shannon (JS) divergence yields a bounded measure of local pattern discrepancy. We use this signal in a conservative policy: similar distributions permit stronger extrapolation, while larger discrepancies receive smaller scales through a bounded mapping. The discrepancy measures observable changes in prediction patterns, rather than semantic correctness. The resulting map preserves the conditional residual

direction at each location and limits only the additional amplification beyond the conditional prediction. Recomputed from the current prediction pair at every sampling step, it adapts both the spatial allocation and the average strength of guidance as generation evolves. PCFG reuses CFG's two existing branch predictions, requiring neither additional training nor extra denoiser evaluations.

We evaluate PCFG on SD3.5 Medium across COCO, DrawBench, and GenEval2 at 10, 20, and 30 sampling steps. With one configuration across all nine settings, PCFG exceeds the metric-specific best-scanned scalar CFG in 41 of 45 mean-score comparisons over five automatic metrics. Each scalar reference selects its scale for the same benchmark, budget, and metric. Against the eight compared guidance baselines, PCFG achieves the highest ImageReward and HPSv2.1 means in all nine settings. Its 20-step results also exceed the best-scanned 30-step scalar CFG on both metrics across all three benchmarks, demonstrating higher scores with fewer sampling steps. Complementary SDXL experiments, with a separately specified configuration, extend the evaluation to noise prediction, with gains in CLIP similarity and ImageReward across all three benchmarks over the available baselines.

Our contributions are threefold:

- We identify a limitation of energy-only spatial feedback: equal pointwise update energy can conceal different changes in local pattern composition, motivating richer feedback for guidance allocation.
- We introduce PCFG, which compares local distributions of ordered amplitude patterns to adapt spatial guidance at each sampling step, without additional training or denoiser evaluations.
- We demonstrate gains over scale-swept CFG and competing guidance methods on SD3.5, extend the evaluation to SDXL, and validate the pattern representation and guidance mapping through ablations.

2 Related Work

2.1 Calibrating Classifier-Free Guidance

Classifier-free guidance (CFG) extrapolates beyond the conditional prediction along the conditional–unconditional difference to improve prompt alignment, but a fixed scale can reduce diversity and introduce artifacts (Ho & Salimans, 2022). One line of work recalibrates prediction magnitude or initializes guidance conservatively to mitigate excessive amplification (Lin et al., 2024; Fan et al., 2025). Another revises the geometry of the update, constraining it to the data manifold or separating its parallel and orthogonal components (Chung et al., 2025; Sadat et al., 2025a). Frequency-aware variants target imbalanced amplification across spectral bands (Sadat et al., 2025b; Song & Lai, 2025). Guidance strength can also vary over the sampling trajectory: prescribed intervals and schedules exploit noise level (Kynkäänniemi et al., 2024; Wang et al., 2024), whereas state-dependent or evaluator-based feedback adapts to the evolving sample (Koulischer et al., 2025; Papalampidi et al., 2026; Wang et al., 2026). Prediction statistics such as relative update magnitude or velocity alignment provide further adaptive signals (Zhu et al., 2025; Luo et al., 2026). Together, these methods address the amount and timing of guidance, without directly specifying which local changes between the two prediction fields warrant different extrapolation.

2.2 Spatial Control and Local Structural Feedback

Spatial control can instead be imposed through layout constraints (Bar-Tal et al., 2023) or attention interventions that encourage subject coverage, perturb attention maps, or rectify attention entropy (Chefer et al., 2023; Ahn et al., 2024; Berrada et al., 2025). These mechanisms use specified layouts or internal attention signals. Closer to adaptive CFG, spatial scales have been defined over attention-derived regions (Shen et al., 2024) or attenuated according to conditional–unconditional delta-score energy (Li et al., 2026). Region assignments and pointwise residual energy capture different aspects of a prediction field, but neither directly describes how conditioning changes local arrangements: equal residual energy need not imply equal changes in neighboring patterns. Two-dimensional dispersion entropy uses ordered patch patterns to characterize image irregularity (Azami et al., 2019). We retain the local pattern distributions and compare them across CFG's two prediction branches

with Jensen–Shannon divergence, using the discrepancy to modulate bounded extrapolation along the existing conditional residual.

3 METHOD

3.1 PRELIMINARIES: CLASSIFIER-FREE GUIDANCE

Given a text condition y, sampling starts from a noisy latent $x_T \sim \mathcal{N}(0, I)$ and progressively refines it toward x_0. At step t, the model produces conditional and unconditional predictions $f_t^y = f_\theta(x_t, t; y)$ and $f_t^u = f_\theta(x_t, t; \emptyset)$ in $\mathbb{R}^{C\times H\times W}$. Here f denotes the native prediction: velocity for SD3.5 and noise for SDXL. With $\Delta f_t = f_t^y - f_t^u$, classifier-free guidance (Ho & Salimans, 2022) uses

$$f_t^{\mathrm{CFG}} = f_t^u + w\,\Delta f_t = f_t^y + (w-1)\,\Delta f_t, \qquad w \geq 1. \tag{1}$$

Thus $w = 1$ recovers ordinary conditional sampling, while $w > 1$ amplifies the condition-dependent component to favor prompt adherence without an external classifier. The second form isolates the additional extrapolation $(w-1)\Delta f_t$ beyond the conditional prediction.

Instead of passing f_t^y directly to the sampler, CFG supplies f_t^{CFG} without changing the sampling rule. Writing $\mathcal{U}_t$ for its update to the next scheduled time t^- and $\mathcal{D}$ for the latent decoder gives

$$x_{t^-} = \mathcal{U}_t\big(x_t, f_t^{\mathrm{CFG}}\big), \qquad \widehat{I} = \mathcal{D}(x_0). \tag{2}$$

The prediction pair is recomputed after each update until the final latent is decoded. Standard CFG uses the same w everywhere; PCFG asks how much further each region should extrapolate, given its current local pattern discrepancy.

3.2 LOCAL PATTERN DISCREPANCY

Residual magnitude measures change at a single location, as in residual-based spatial guidance (Li et al., 2026), but does not describe how nearby values are arranged. A smooth transition and scattered fluctuations can contain similar amplitude levels yet form different patterns. PCFG compares these arrangements: it expresses both prediction fields in a shared vocabulary of small spatial patterns, then measures how their local frequencies differ. This construction draws on dispersion-pattern representations (Li et al., 2021).

A shared amplitude vocabulary. For branch $b \in \{u, y\}$, let $A_t^b(x) = \|f_t^b(:, x)\|_2$ be the prediction amplitude at latent-grid location x. Both full maps share a pooled mean μ_t and population variance σ_t^2. We map their amplitudes to c discrete states using the standard Gaussian CDF Φ, with $\epsilon = 10^{-6}$ for numerical stability:

$$Q_t^b(x) = \min\left\{c-1,\ \left\lfloor c\,\Phi\left(\frac{A_t^b(x)-\mu_t}{\sqrt{\sigma_t^2+\epsilon}}\right)\right\rfloor\right\}. \tag{3}$$

The parameter c sets *amplitude resolution*: with $c = 3$, states represent low, intermediate, and high amplitudes on a common scale. Shared thresholds make labels comparable across branches, rather than normalizing away their differences independently. The CDF bounds the calibration; it does not assume Gaussian prediction fields. This representation captures amplitude arrangements, but not changes in vector direction that leave the amplitudes unchanged.

Patterns and their context. An ordered $p \times p$ patch defines one pattern. If δ_ℓ enumerates its offsets in row-major order, the patch starting at q receives the label

$$R_t^b(q) = \sum_{\ell=0}^{p^2-1} c^{p^2-1-\ell} Q_t^b(q+\delta_\ell), \qquad R_t^b(q) \in \{0, \ldots, c^{p^2}-1\}. \tag{4}$$

Here p sets the *pattern scale*: $p = 1$ retains only individual amplitude states, whereas larger patches distinguish their spatial arrangements. Equal state counts need not give the same ordered code. Around each location x, a $k \times k$ neighborhood supplies the *statistical context*. Let $\mathcal{A}_{p,k}(x)$ contain

the starting positions of its fully enclosed patches, with $k \geq p$. Their empirical pattern distribution is

$$P_t^b(r \mid x) = \frac{1}{(k-p+1)^2} \sum_{q \in \mathcal{A}_{p,k}(x)} \mathbf{1}[R_t^b(q) = r]. \tag{5}$$

Thus c and p determine a dictionary of c^{p^2} patterns, while k determines both the spatial extent and the $(k-p+1)^2$ observations used to estimate their frequencies. These roles explain the representation and context choices examined in our ablations. The histogram retains order within each patch, but not the exact placement of patches within the context.

Comparing the two distributions. With $M_t(r \mid x) = \frac{1}{2}[P_t^u(r \mid x) + P_t^y(r \mid x)]$, we use the base-2 Jensen–Shannon divergence

$$S_t(x) = \frac{1}{2} \sum_{b \in \{u,y\}} \sum_r P_t^b(r \mid x) \log_2 \frac{P_t^b(r \mid x)}{M_t(r \mid x)}. \tag{6}$$

Zero-frequency terms contribute zero. This symmetric comparison remains finite when a pattern occurs in only one branch and lies in $[0, 1]$: identical distributions yield zero, and disjoint supports yield one. These properties make JS suitable for comparing sparse pattern histograms and provide a bounded input to the scale mapping. It measures a change in local amplitude-pattern composition, rather than residual magnitude or semantic correctness.

3.3 From Pattern Discrepancy to Spatial Guidance

The conditional prediction already incorporates the conditioning-induced change; the remaining decision is how far to extrapolate beyond it. PCFG uses pattern discrepancy as a conservative control signal for this additional amplification, not as evidence that the conditional prediction is incorrect. Larger discrepancies receive smaller coefficients, while similar pattern distributions permit larger ones. We express this policy through the bounded power map

$$W_t(x) = w_{\min} + (w_{\max} - w_{\min})[1 - S_t(x)]^\gamma, \qquad w_{\max} > w_{\min} \geq 1, \quad \gamma > 0. \tag{7}$$

The bounds fix the allowable guidance range: $S_t = 0$ yields $w_{\max}$ and $S_t = 1$ yields $w_{\min}$. The exponent γ controls the falloff; increasing it lowers weights at intermediate discrepancies.

Replacing the scalar w in Equation 1 with $W_t \in \mathbb{R}^{1 \times H \times W}$ gives PCFG's guided prediction:

$$f_t^{\mathrm{PCFG}} = f_t^u + W_t \odot \Delta f_t = f_t^y + (W_t - 1) \odot \Delta f_t. \tag{8}$$

Weights are broadcast across channels, preserving the residual direction at each location. Since $W_t \geq 1$, reducing a weight limits additional extrapolation rather than removing conditioning. Its spatial mean is not fixed, so both allocation and effective guidance strength can change. At every step, this prediction replaces f_t^{CFG} in Equation 2. PCFG requires no training or additional denoiser evaluations; pattern statistics add computation and storage, whose implementation details are given in the supplementary material.

4 Experiments

4.1 Experimental Setup

Backbones and benchmarks. We use SD3.5 Medium for the main study and conduct a separate SDXL evaluation (Podell et al., 2023) to assess the method beyond velocity prediction. SD3.5 generates 1024×1024 images at 10, 20, and 30 sampling steps on COCO (Lin et al., 2014), DrawBench (Saharia et al., 2022), and GenEval2 (Kamath et al., 2025), containing 5,000, 200, and 800 prompts. These sets cover natural-image captions, challenging text-to-image prompts, and compositional descriptions. Here GenEval2 supplies the prompt set; the reported results use the five metrics below, rather than its dedicated compositional evaluator. All methods share the recorded prompt lists, resolution, seed 4416, and sampling budget, while retaining their own sampling updates. SDXL uses a separate 50-step deterministic DDIM protocol at the same resolution.

Table 1: **Quantitative comparison on SD3.5 across three sampling budgets and three benchmarks.** IR: ImageReward; HPS: HPSv2.1; Pick: PickScore; Aest.: Aesthetic Score. Higher is better. Bold: highest reported mean, including ties.

	COCO					DrawBench					GenEval2				
Method	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.
10 sampling steps															
CFG(w=7)	26.965	0.597	0.252	0.833	5.298	28.313	0.655	0.260	0.844	5.301	31.592	1.205	0.290	0.868	5.498
Rescale	27.008	0.753	0.262	0.841	5.331	28.544	0.722	0.266	0.850	5.324	31.731	1.277	0.295	0.871	5.533
CFG++	25.917	0.122	0.221	0.816	5.224	26.237	-0.073	0.221	0.821	5.241	27.733	0.031	0.232	0.833	5.268
CFG-Zero*	20.117	-1.879	0.142	0.699	4.324	19.659	-2.038	0.142	0.711	3.989	25.338	-1.729	0.181	0.762	4.346
APG	26.788	0.730	0.252	0.845	5.247	28.294	0.571	0.250	0.848	5.317	31.144	1.017	0.274	0.865	5.504
CFG-Ctrl	25.981	0.407	0.239	0.816	5.236	26.683	0.232	0.236	0.818	5.190	29.830	0.490	0.258	0.833	5.291
VAGS	27.082	0.831	0.264	0.848	5.313	28.412	0.699	0.264	0.852	5.376	31.807	1.241	0.293	0.873	5.552
SAMG	**27.087**	0.803	0.264	0.845	5.339	28.552	0.756	0.266	0.851	5.366	31.715	1.265	0.296	0.872	5.540
PCFG	27.077	**0.983**	**0.278**	**0.853**	**5.365**	**28.981**	**0.864**	**0.277**	**0.855**	**5.396**	**32.038**	**1.364**	**0.304**	**0.876**	**5.586**
20 sampling steps															
CFG(w=7)	26.883	0.990	0.284	0.861	5.379	28.829	0.892	0.281	0.862	5.388	31.935	1.372	0.307	0.882	5.549
Rescale	26.840	1.026	0.287	0.864	5.380	28.769	0.914	0.284	**0.865**	5.405	32.020	1.392	0.311	0.884	5.583
CFG++	26.768	0.776	0.267	0.850	5.350	28.306	0.691	0.265	0.853	5.333	31.307	1.147	0.292	0.873	5.482
CFG-Zero*	26.503	0.529	0.263	0.831	5.323	28.009	0.367	0.249	0.826	5.063	31.513	0.945	0.272	0.849	5.166
APG	26.586	0.918	0.273	0.860	5.315	28.374	0.760	0.271	0.860	5.404	31.566	1.242	0.295	0.877	5.569
CFG-Ctrl	26.373	0.869	0.275	0.851	5.349	28.264	0.716	0.267	0.848	5.349	30.946	1.045	0.292	0.865	5.508
VAGS	26.857	0.986	0.280	0.863	5.342	28.908	0.897	0.280	0.863	5.368	32.049	1.385	0.307	0.884	5.580
SAMG	26.902	1.026	0.286	**0.865**	5.382	28.789	0.919	0.283	0.862	5.394	31.932	1.390	0.310	0.883	5.574
PCFG	**27.072**	**1.100**	**0.291**	**0.865**	**5.415**	**29.141**	**0.963**	**0.288**	0.862	**5.438**	**32.187**	**1.457**	**0.314**	**0.885**	**5.613**
30 sampling steps															
CFG(w=7)	26.779	1.057	0.290	0.868	5.381	28.838	0.931	0.285	0.867	5.365	31.990	1.363	0.309	0.884	5.537
Rescale	26.746	1.075	0.291	**0.869**	5.381	28.880	0.917	0.287	**0.868**	5.391	32.089	1.396	0.312	0.886	5.563
CFG++	26.701	0.814	0.273	0.850	5.384	28.330	0.677	0.267	0.850	5.287	31.414	1.072	0.286	0.867	5.381
CFG-Zero*	26.614	0.973	0.285	0.860	5.393	28.659	0.848	0.282	0.859	5.368	31.900	1.330	0.308	0.876	**5.613**
APG	26.491	0.969	0.279	0.864	5.328	28.573	0.849	0.278	0.865	5.402	31.722	1.291	0.301	0.880	5.584
CFG-Ctrl	26.405	0.931	0.282	0.858	5.376	28.342	0.787	0.278	0.857	5.346	31.043	1.140	0.301	0.873	5.545
VAGS	26.762	1.019	0.284	0.866	5.343	28.965	0.905	0.283	0.867	5.391	32.129	1.389	0.309	0.885	5.566
SAMG	26.776	1.069	0.290	**0.869**	5.390	28.973	0.893	0.286	0.866	5.402	31.988	1.402	0.312	0.886	5.564
PCFG	**26.970**	**1.113**	**0.294**	0.867	**5.430**	**29.306**	**0.984**	**0.291**	0.865	**5.407**	**32.328**	**1.474**	**0.314**	**0.887**	5.592

Compared methods and guidance settings. The SD3.5 comparison includes Standard CFG, CFG-Rescale, CFG++, CFG-Zero*, APG, CFG-Ctrl, VAGS, and SAMG, covering magnitude recalibration, geometric correction, adaptive control, and energy-based spatial guidance. PCFG fixes $(c, p, k, w_{\min}, w_{\max}, \gamma) = (3, 3, 21, 3, 18, 2)$ across all nine SD3.5 benchmark–budget settings, following discrete search and joint consideration of the metrics (Appendix E.1). All nonstandard methods retain their recorded parameters across these settings; Appendix C lists the configurations. In addition to the CFG-7 operating point in the main table, we scan scalar CFG over $\{3, 3.5, 4, 4.5, 5, 6, 7, 8, 10, 13\}$. For each benchmark, budget, and metric, the largest mean forms a *best-scanned scalar* reference. This deliberately strong reference can select a different scale for each comparison, whereas PCFG keeps the same parameters throughout. For SDXL, Table 2 reports results for Standard CFG, CFG-Rescale, CFG++, APG, and SAMG. PCFG retains its separately evaluated SDXL configuration $c = 3$, $p = 3$, $k = 13$, $w_{\min} = 3$, $w_{\max} = 18$, and $\gamma = 2$ across all three benchmarks.

Evaluation metrics. We report CLIP image–text similarity (Radford et al., 2021), ImageReward (IR) (Xu et al., 2023), HPSv2.1 (HPS) (Wu et al., 2023), PickScore (Kirstain et al., 2023), and Aesthetic Score. Together they measure text alignment, learned preference, and visual appeal; higher values are better. The SD3.5 main table reports all five metrics; Appendix D provides the complete scalar sweeps and all benchmark–budget–metric envelope comparisons. Scores are run means: CLIP is scaled by 100, HPS is in raw units, and PickScore is the evaluator logit divided by 26. Bold values mark the highest reported mean within each comparison, including ties. Differences quoted in the text are calculated from the reported means. Additional details are provided in the appendix.

4.2 Main Results

Comparison with existing guidance methods. Table 1 compares nine guidance methods on SD3.5 across three benchmarks, three sampling budgets, and all five metrics. With one configuration across all benchmark–budget settings, PCFG achieves the highest reported IR and HPS means in all nine settings—18 of 18 cells. Its gains therefore extend beyond a single dataset or a low-step operating point. At 10 steps, for example, COCO IR increases from 0.831 for the strongest listed alternative to 0.983, and DrawBench IR from 0.756 to 0.864. The advantage remains at 30 steps: GenEval2 IR reaches 1.474, compared with 1.402 for the strongest listed alternative. These results support local pattern feedback as a useful basis for guidance alongside magnitude-, geometry-, and control-based strategies. Across the full results, PCFG also attains the highest CLIP and Aesthetic means in eight of nine settings each. For PickScore, it leads in five settings and ties for the highest reported mean in one more. The remaining cases expose metric-dependent trade-offs, particularly for PickScore at larger budgets.

Gains beyond scalar-scale tuning. We next compare against a hindsight envelope over ten scanned standard-CFG scales, allowing each metric to select its own best result. Figure 3 illustrates this comparison on COCO at 20 steps: the same PCFG configuration exceeds the scanned CFG curve on all five metrics. The gains over the corresponding scalar maxima are 0.083 in CLIP, 0.108 in IR, 0.007 in HPS, 0.001 in PickScore, and 0.036 in Aesthetic Score. Across all three benchmarks and budgets, PCFG improves 41 of 45 means over these metric-specific scalar references. Its gains therefore persist beyond the choice of a single fixed CFG scale. Appendix D reports all 45 comparisons and the underlying scalar sweeps, including the four non-winning cases.

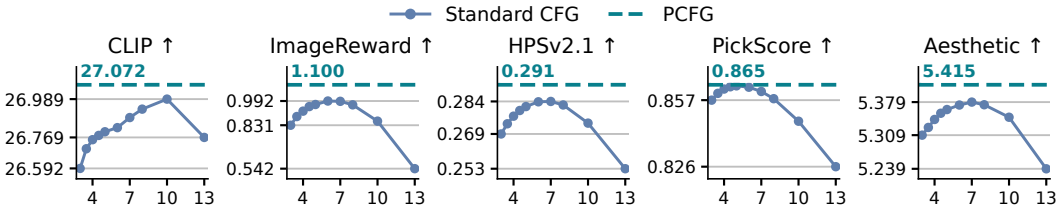


Figure 3: **Scalar CFG sweep on COCO at 20 steps.** PCFG uses one configuration and exceeds the scanned standard-CFG results on all five metrics. The dashed line is its measured mean; thin grey lines mark the left endpoint, peak, and right endpoint of the scanned CFG curve, with their values printed on the y axis.

Fewer sampling steps for competitive quality. Figure 4 compares 20-step PCFG with the best-scanned 30-step CFG. PCFG wins all six shown IR/HPS comparisons and exceeds or matches CFG in 13 of 15 comparisons across all five metrics at the reported precision; the two exceptions are PickScore on COCO and DrawBench. Thus, PCFG attains higher measured quality with fewer sampling steps in most settings; wall-clock gains also depend on guidance-map computation.

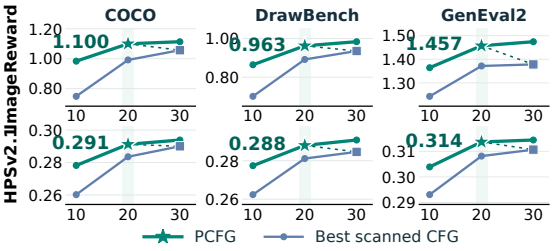


Figure 4: **Quality with fewer steps.** 20-step PCFG (stars) exceeds best-scanned 30-step CFG (squares) in all six shown IR/HPS comparisons.

Qualitative comparison. Figure 5 compares three examples under matched prompts and sampling budgets. The examples show differences in object structure, appearance, and text rendering across methods, complementing the aggregate scores in Table 1.

Extension to SDXL. To assess applicability beyond velocity prediction, we extend the evaluation to SDXL-base-1.0, applying the same local-pattern construction and guidance rule to its native noise

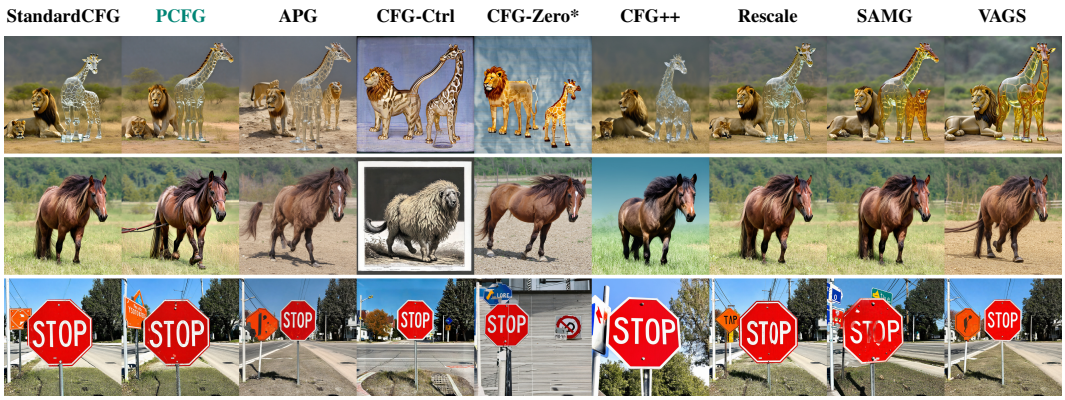


Figure 5: **SD3.5 qualitative comparison at 10 sampling steps.** Each row uses a common prompt and recorded seed across the nine methods in Table 1; columns follow the method labels above.

Table 2: **Quantitative comparison on SDXL across three benchmarks.** 50-step deterministic DDIM. Bold: highest reported mean, including ties.

	COCO					DrawBench					GenEval2				
Method	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.
CFG (w=13)	26.898	0.953	0.295	0.874	5.628	28.442	0.814	0.286	0.865	5.543	31.135	0.839	**0.307**	0.881	5.712
CFG-Rescale	26.672	0.909	0.289	0.874	5.616	28.177	0.757	0.279	0.865	5.572	30.803	0.772	0.301	0.881	5.768
CFG++	26.736	0.896	0.288	0.872	5.629	28.104	0.732	0.277	0.862	**5.578**	30.821	0.761	0.301	0.880	**5.784**
APG	26.779	0.942	0.294	**0.875**	5.622	28.413	0.823	0.285	**0.867**	5.539	31.034	0.820	0.306	**0.882**	5.729
SAMG	26.804	0.940	0.294	0.874	5.629	28.185	0.819	0.285	0.865	5.556	31.067	0.819	0.306	**0.882**	5.739
PCFG	**26.933**	**0.968**	**0.297**	0.873	**5.641**	**28.583**	**0.859**	**0.287**	0.865	5.547	**31.193**	**0.847**	**0.307**	0.881	5.694

predictions. This complementary experiment uses 50-step deterministic DDIM sampling on COCO, DrawBench, and GenEval2. As shown in Table 2, PCFG uses one configuration across all three benchmarks. Relative to the five baselines, it achieves the highest CLIP similarity and ImageReward on every benchmark. For HPSv2.1, it leads on COCO and DrawBench and ties with CFG at 0.307 on GenEval2 at the reported precision. It therefore matches or exceeds the strongest reported baseline in all nine CLIP/IR/HPS comparisons and obtains the highest reported mean, including ties, in 10 of the 15 metric–benchmark pairs overall. Its ImageReward gains over the strongest reported baseline are 0.015, 0.036, and 0.008 on COCO, DrawBench, and GenEval2, respectively. PickScore and Aesthetic Score show remaining trade-offs: PCFG does not lead PickScore and leads Aesthetic Score only on COCO.

4.3 Ablation Study

Our ablations progress from local-pattern resolution and guidance mapping (Table 3, groups A and B) to parameter sensitivity (Figure 6). The design and sensitivity studies use the same balanced set of 192 prompts, with 64 each from COCO, DrawBench, and GenEval2, on SD3.5 at 10 steps with seed 4416. The design and sensitivity studies vary one factor from the search reference $(c, p, k, w_{\min}, w_{\max}, \gamma) = (3, 3, 13, 3, 18, 2)$. Appendix E.1 expands the search evidence and explains the joint selection of the final configuration.

Table 3: **Representation and mapping choices.** Search reference: $k = 13$.

Configuration	IR ↑	HPS ↑	CLIP ↑	Pick ↑	Aest. ↑
CFG-7	0.607	0.251	26.688	0.831	5.292
PCFG ($c = 3, p = 3$)	**1.018**	**0.279**	**27.027**	0.851	**5.345**
A. Pattern representation (fixed power map)					
Smaller patches ($p = 2$)	-0.633	0.190	24.913	0.765	4.790
Coarser states ($c = 2$)	0.781	0.261	26.897	0.831	5.296
B. Guidance mapping (fixed $c = 3, p = 3$)					
Exponential ($\lambda = 3$)	1.004	0.275	26.962	**0.853**	5.325

Local-pattern resolution. With the reference power map fixed, group A of Table 3 tests amplitude resolution c and local patch size p. Coarsening the amplitude states from $c = 3$ to $c = 2$ lowers IR from 1.018 to 0.781, with corresponding reductions in HPS and CLIP. Reducing the patch from 3×3 to 2×2 yields a larger decline across all five metrics, with IR falling to -0.633. These comparisons support $(c, p) = (3, 3)$ among the tested representations. They measure representation choices within the configured policy: $p = 2$ still encodes spatial patterns, and changing c or p also changes the discrepancy distribution presented to the fixed mapping.

How discrepancy becomes guidance. Group B of Table 3 compares our power map (Equation 7) with an exponential alternative. The latter uses $W_t(x) = 1 + (w_{\max} - 1)e^{-\lambda S_t(x)}$, where λ controls decay. Both retain $(c, p, k) = (3, 3, 13)$ and $w_{\max} = 18$; Power uses $w_{\min} = 3$ and $\gamma = 2$, while Exponential uses $\lambda = 3$. Exponential reaches IR 1.004, close to Power's 1.018, showing that useful performance is not restricted to one functional form. Power is higher on IR, HPS, CLIP, and Aesthetic, while Exponential has a slightly higher PickScore (0.853 versus 0.851). This comparison supports the choice of the power policy while retaining the observed metric trade-off. Sensitivity within the power family is examined next.

Effects of statistical context. The first column of Figure 6 varies the statistical context from $k = 7$ to 25 and shows a broad high-performing range. We select $k = 21$ within this range; $k = 13$ remains the reference for the earlier design and mapping sweeps.

The remaining columns examine $w_{\min}$, $w_{\max}$, and γ. Floors $w_{\min} \in \{1, 2, 3\}$ keep IR near 1.00, whereas raising the floor to 5 lowers it to 0.922. Increasing $w_{\max}$ from 18 to 27 does not consistently improve IR. Among $\gamma \in \{1, 2, 3\}$, $\gamma = 2$ gives the highest IR and HPS in this sweep. Balancing these metric responses, we retain $(w_{\min}, w_{\max}, \gamma) = (3, 18, 2)$ and combine them with $k = 21$ for the fixed SD3.5 configuration in Section 4.1.

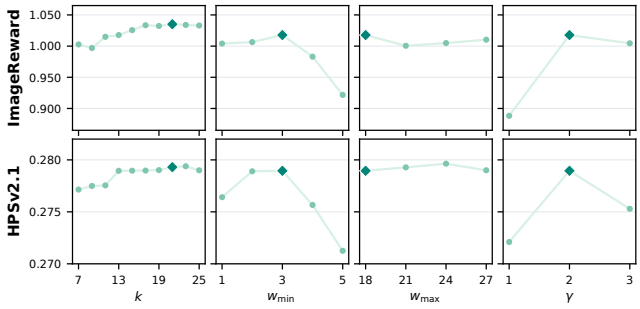


Figure 6: **Sensitivity of statistical context and guidance mapping.** Diamonds mark $(k, w_{\min}, w_{\max}, \gamma) = (21, 3, 18, 2)$ in the respective sweeps. The mapping sweeps retain $k = 13$.

5 Conclusion

We presented PCFG, a training-free approach that adapts classifier-free guidance using discrepancies between local pattern distributions of conditional and unconditional predictions. This representation captures spatial relationships that pointwise residual magnitudes can overlook, providing feedback for spatially varying guidance at each sampling step without additional denoiser evaluations. On SD3.5 Medium, one fixed configuration exceeds the metric-specific best-scanned scalar CFG in 41 of 45 comparisons across three benchmarks and three sampling budgets. Its 20-step results also surpass the best-scanned 30-step scalar CFG in ImageReward and HPSv2.1 across all three benchmarks. Complementary SDXL results extend the evidence to noise prediction, while ablations support the chosen pattern representation and guidance mapping. These findings support local pattern discrepancy as a useful signal for adaptive guidance. Improving consistency across metrics and reducing the cost of local statistics remain directions for further study.

AI Use Statement

We used AI tools solely for language polishing and debugging plotting code. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final content.

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*, 2024.

Hamed Azami, Luiz Eduardo Virgilio da Silva, Ana Carolina Mieko Omoto, and Anne Humeau-Heurtier. Two-dimensional dispersion entropy: An information-theoretic method for irregularity analysis of images. *Signal Processing: Image Communication*, 75:178–187, 2019. doi: 10.1016/j.image.2019.04.013.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 2023.

Tariq Berrada, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 45750–45769, 2025. doi: 10.52202/085713-1525.

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*, 42(4), 2023. doi: 10.1145/3592116.

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*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. CFG-Zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

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*, 2021. URL https://arxiv.org/abs/2104.08718.

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

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

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. *arXiv preprint arXiv:2305.01569*, 2023.

Felix Koulischer, Florian Handke, Johannes Deleu, Thomas Demeester, and Luca Ambrogioni. Feedback guidance of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Haosen Li, Wenshuo Chen, Lei Wang, Shaofeng Liang, Bowen Tian, Soning Lai, and Yutao Yue. Delta score matters! spatial adaptive multi guidance in diffusion models. *arXiv preprint arXiv:2604.26503*, 2026.

Yuxing Li, Shangbin Jiao, Bo Geng, and Yuan Zhou. Research on feature extraction of ship-radiated noise based on multi-scale reverse dispersion entropy. *Applied Acoustics*, 173:107737, 2021. doi: 10.1016/j.apacoust.2020.107737.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Yan Luo, Ahmadou Aidara, Jingyi Lu, Jeremy Moebel, Kai Han, and Mengyu Wang. VAGS: Velocity adaptive guidance scale for image editing and generation. *arXiv preprint arXiv:2605.15661*, 2026.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajić, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. In *International Conference on Learning Representations*, 2026.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

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*, pp. 10684–10695, 2022.

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

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

Christoph Schuhmann. Improved aesthetic predictor. `https://github.com/christophschuhmann/improved-aesthetic-predictor`, 2022.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9370–9379, 2024.

Kaiyu Song and Hanjiang Lai. Rethinking oversaturation in classifier-free guidance via low frequency. *arXiv preprint arXiv:2506.21452*, 2025.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. CFG-Ctrl: Control-based classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Shangwen Zhu, Qianyu Peng, Zhilei Shu, Yuting Hu, Zhantao Yang, Han Zhang, Zhao Pu, Andy Zheng, Xinyu Cui, Jian Zhao, Ruili Feng, and Fan Cheng. MAMBO-G: Magnitude-aware mitigation for boosted guidance. *arXiv preprint arXiv:2508.03442*, 2025.

Where Guidance Matters Most: Spatially Adaptive CFG for Diffusion Models

Supplementary Material

A PCFG Implementation Details

Shared calibration and quantization. All statistics are computed independently for each sample and sampling step. For a grid Ω of size $H \times W$, the channel-L2 amplitudes A_t^u, A_t^y give

$$\mu_t = \frac{1}{2HW} \sum_{b\in\{u,y\}} \sum_{x\in\Omega} A_t^b(x), \qquad \sigma_t^2 = \frac{1}{2HW} \sum_{b\in\{u,y\}} \sum_{x\in\Omega} \big(A_t^b(x) - \mu_t\big)^2. \tag{9}$$

The variance is the population variance, with no degrees-of-freedom correction. Calibration uses the unpadded fields. The implementation forms $z = (A_t^b(x) - \mu_t)/\sqrt{\sigma_t^2 + \epsilon}$ and $\Phi(z) = (1 + \mathrm{erf}(z/\sqrt{2}))/2$, then evaluates $\mathrm{clamp}(\lfloor c\Phi(z) + 1 \rfloor, 1, c) - 1$. In exact arithmetic this is the capped floor quantizer used in the main text. We use $\epsilon = 10^{-6}$ for numerical protection. These are shared amplitude thresholds, not independent branch-wise rank transforms or a guarantee of invariance to rescaling one branch.

Ordered patterns and boundary handling. Let δ_ℓ enumerate the $p \times p$ offsets in row-major order. The label of the patch beginning at q is

$$R_t^b(q) = \sum_{\ell=0}^{p^2-1} c^{p^2-1-\ell} Q_t^b(q + \delta_\ell). \tag{10}$$

This assigns a distinct label to every ordered discrete patch in the c^{p^2}-element dictionary. The implementation uses an odd context size $k \geq p$ and reflects each quantized field by $(k-1)/2$ pixels on every side. It then extracts stride-one $p \times p$ patches and average-pools one-hot codes with a $(k-p+1) \times (k-p+1)$ window. The result has shape $[B, c^{p^2}, H, W]$ and sums to one over the dictionary at each location. Reflection supplies the boundary neighborhoods; no padded value enters the calibration statistics. Local histograms discard the precise arrangement of patch labels within a context, although each label retains its internal pixel order.

Numerical evaluation of JS divergence. Using the midpoint distribution $M_t(r \mid x) = \frac{1}{2}[P_t^u(r \mid x) + P_t^y(r \mid x)]$, the finite-precision implementation evaluates

$$\begin{aligned} \widetilde{S}_t(x) &= \frac{1}{2} \sum_{b\in\{u,y\}} \sum_{r:P_t^b(r|x)>0} P_t^b(r \mid x) \log_2 \left(\frac{P_t^b(r \mid x) + \epsilon}{M_t(r \mid x) + \epsilon} \right), \\ S_t(x) &= \mathrm{clamp}\big(\widetilde{S}_t(x), 0, 1\big). \end{aligned} \tag{11}$$

This is an ϵ-regularized evaluation of the standard JS expression, not a different entropy objective. Zero-frequency terms contribute zero. Norms, calibration, histograms, and divergence are computed in FP32. After the power map is applied, W_t is cast to the input prediction dtype before the guided prediction is formed. A diagnostic S_t map, when requested, remains FP32.

Algorithm 1 One sampling step of PCFG

Input: Latent x_t, time t, condition y, model f_θ, scheduler update $\mathcal{U}_t$.
Input: Parameters $(c, p, k, w_{\min}, w_{\max}, \gamma)$; odd $k \geq p$; $\epsilon = 10^{-6}$.
Output: Updated latent x_{t^-} at the next scheduled time t^-.

Process each sample independently; spatial operations apply at every $x \in \Omega$.

$f_t^u \leftarrow f_\theta(x_t, t; \emptyset)$; $\quad f_t^y \leftarrow f_\theta(x_t, t; y)$

$\Delta f_t \leftarrow f_t^y - f_t^u$

$A_t^b(x) \leftarrow \|f_t^b(:, x)\|_2$ for $b \in \{u, y\}$

Compute (μ_t, σ_t^2) from the unpadded A_t^u, A_t^y using Equation 9.

for $b \in \{u, y\}$ **do**

$$Q_t^b(x) \leftarrow \min\left\{c-1, \left\lfloor c\,\Phi\left(\frac{A_t^b(x) - \mu_t}{\sqrt{\sigma_t^2 + \epsilon}}\right)\right\rfloor\right\}$$

Reflect-pad Q_t^b by $(k-1)/2$ pixels on each side.

$$R_t^b(q) \leftarrow \sum_{\ell=0}^{p^2-1} c^{p^2-1-\ell} Q_t^b(q + \delta_\ell)$$ ▷ Row-major patch offsets

$$P_t^b(r \mid x) \leftarrow \frac{1}{(k-p+1)^2} \sum_{q \in \mathcal{A}_{p,k}(x)} \mathbf{1}[R_t^b(q) = r]$$

end for

$M_t(r \mid x) \leftarrow \frac{1}{2}\left[P_t^u(r \mid x) + P_t^y(r \mid x)\right]$

$$\widetilde{S}_t(x) \leftarrow \frac{1}{2} \sum_{b \in \{u,y\}} \sum_{r : P_t^b(r|x) > 0} P_t^b(r \mid x) \log_2 \frac{P_t^b(r \mid x) + \epsilon}{M_t(r \mid x) + \epsilon}$$

$S_t(x) \leftarrow \mathrm{clamp}\big(\widetilde{S}_t(x), 0, 1\big)$

$W_t(x) \leftarrow w_{\min} + (w_{\max} - w_{\min})[1 - S_t(x)]^\gamma$

Cast W_t to the prediction dtype.

$f_t^{\mathrm{PCFG}} \leftarrow f_t^u + W_t \odot \Delta f_t$ ▷ Broadcast W_t across channels

$x_{t^-} \leftarrow \mathcal{U}_t\big(x_t, f_t^{\mathrm{PCFG}}\big)$

return x_{t^-}

B Evaluation Metrics and Compared Guidance Methods

This section briefly introduces the automatic metrics and guidance baselines used in our experiments. Additional experimental details are provided in Appendix C.

B.1 Evaluation Metrics

CLIP similarity. CLIP learns a joint image–text representation by contrastive pre-training on large-scale image–caption pairs (Radford et al., 2021). This representation also supports reference-free image–text evaluation, as in CLIPScore (Hessel et al., 2021). For each generated image and its prompt, we report 100 times the cosine similarity between normalized CLIP ViT-L/14 image and text embeddings, labeled CLIP in the tables. Thus, the score primarily reflects whether the visual content is semantically consistent with the conditioning prompt.

ImageReward. ImageReward is a learned text-to-image reward model trained from human preference annotations (Xu et al., 2023). Its training data contain 137K expert comparisons between images generated for the same prompts, and the model maps a prompt–image pair to a scalar reward. In contrast to a fixed embedding similarity, this supervision allows the score to incorporate the preferences expressed by the annotators when comparing generated images. We use the released ImageReward-v1.0 checkpoint.

HPSv2.1. Human Preference Score v2 (HPSv2) is obtained by fine-tuning a CLIP-based model on the Human Preference Dataset v2 (Wu et al., 2023). The dataset contains approximately 798K human preference choices over images collected from a range of text-to-image systems. Given

an image and its prompt, HPS produces a paired image–text similarity score learned from these preferences. Our experiments use the subsequently released HPSv2.1 checkpoint.

PickScore. PickScore is a CLIP-based preference model trained on Pick-a-Pic, an open dataset collected from real users who compared two generated images under the same prompt (Kirstain et al., 2023). The model embeds the prompt and candidate images in a shared space and assigns higher compatibility to the image predicted to be preferred. We use the released PickScore-v1 model to score each generated image against its conditioning prompt.

Aesthetic Score. The Aesthetic Score is an image-only predictor built from a frozen CLIP ViT-L/14 image encoder and a multilayer perceptron trained to regress average aesthetic ratings (Schuhmann, 2022). Because it does not take the prompt as input, it evaluates the visual presentation of an image rather than prompt–image compatibility. We use the publicly released `sac+logos+ava1-l14-linearMSE` checkpoint.

B.2 Compared Guidance Methods

Let f_t^y and f_t^u denote the conditional and unconditional predictions at sampling step t, respectively, and let $\Delta f_t = f_t^y - f_t^u$. The methods below modify either the magnitude, direction, temporal schedule, spatial distribution, or sampler use of this conditional residual.

Classifier-free guidance (CFG). Classifier-free guidance combines conditional and unconditional predictions from a jointly trained generative model, avoiding a separately trained classifier (Ho & Salimans, 2022). At inference time, it forms $f_t^{\mathrm{CFG}} = f_t^u + w\Delta f_t$, where the scalar w is kept fixed over both sampling time and spatial position. Increasing w strengthens the effect of the condition, while also changing the fidelity–diversity trade-off of the sampler.

CFG-Rescale. CFG-Rescale was introduced together with corrections to commonly used diffusion noise schedules (Lin et al., 2024). After forming the usual CFG prediction, it matches the prediction's standard deviation to that of the conditional prediction and interpolates between the original and rescaled outputs. The interpolation coefficient controls how strongly this magnitude correction is applied. The method therefore retains the CFG direction while calibrating the scale of the guided prediction to reduce over-exposure at large guidance values.

CFG++. CFG++ derives text guidance from a manifold-constrained interpretation of the sampling process (Chung et al., 2025). Its update separates the two roles that are coupled in conventional CFG: the guided prediction is used to form the denoised estimate, whereas the unconditional prediction supplies the renoising direction toward the next noise level. This modification changes the sampler update rather than only rescaling the conditional residual, and uses an interpolation parameter whose numerical range is distinct from the usual CFG scale.

Adaptive projected guidance (APG). APG decomposes the CFG update into components parallel and orthogonal to the conditional prediction (Sadat et al., 2025a). Motivated by the observation that the parallel component is chiefly associated with oversaturation, APG down-weights this component while retaining the orthogonal component. It additionally constrains the norm of the guidance update and applies reverse momentum across sampling steps. These operations act directly on the guidance vector and require no additional model evaluation.

CFG-Zero*. CFG-Zero* targets flow-matching models and combines optimized-scale guidance with zero initialization (Fan et al., 2025). At each guided step, the optimized scale is computed by projecting the conditional velocity onto the unconditional velocity and is used to rescale the unconditional term before extrapolation. Zero initialization suppresses the model output during the first few ODE-solver steps, where the initial velocity estimates can be unreliable. The two components are applied at inference time without retraining the generator.

CFG-Ctrl. CFG-Ctrl interprets classifier-free guidance as feedback control over the continuous-time generative flow (Wang et al., 2026). Under this view, standard CFG is a proportional controller whose error signal is the conditional–unconditional discrepancy. The evaluated SMC-CFG variant

Table 4: **Method configurations for the SD3.5 comparison.** Each configuration is kept fixed across COCO, DrawBench, and GenEval2 and across the three sampling budgets.

Method	Configuration
Standard CFG	Guidance scale 7; additional scan over $3, 3.5, 4, 4.5, 5, 6, 7, 8, 10, 13$.
CFG-Rescale	Guidance scale 7; rescale coefficient 0.3.
CFG++	Interpolation parameter 0.15.
CFG-Zero*	Guidance scale 7; default initialization settings enabled.
APG	Guidance scale 7; $\eta = 0$, norm threshold 15, momentum -0.5.
CFG-Ctrl	Guidance scale 7; $\lambda = 5$, $k = 0.2$, SMC enabled.
VAGS	Guidance scale 7; $\kappa = 0.5$.
SAMG	Guidance bounds $[3, 7.5]$.
PCFG	$c = 3$, $p = 3$, $k = 21$; power mapping with bounds $[3, 18]$ and $\gamma = 2$.

constructs an exponential sliding surface from this semantic error and adds a nonlinear switching term to the guided velocity. Its control gains govern the correction applied as the prediction error evolves along the trajectory.

Velocity-adaptive guidance scaling (VAGS). VAGS replaces the fixed CFG coefficient with a step-dependent scale derived from the current flow dynamics (Luo et al., 2026). For text-to-image generation, it uses the cosine alignment between the conditional and unconditional velocity fields together with the current signal level. The nominal guidance scale is multiplied by a bounded temporal-alignment factor, and standard CFG is recovered when the modulation strength is set to zero. The resulting scale is global within a step but varies along the sampling trajectory.

Spatial adaptive multi guidance (SAMG). SAMG assigns different guidance strengths to different spatial locations using the conditional–unconditional delta score (Li et al., 2026). It computes a point-wise energy by averaging the squared residual across channels, normalizes the energy within each sample and timestep, and maps it to a bounded spatial guidance field. High-energy locations receive the lower guidance value, whereas low-energy locations receive the upper value. The resulting spatial map modulates the existing CFG residual without an auxiliary network or additional denoiser evaluations.

C Additional Experimental Protocol

C.1 Method Implementations and Configurations

We evaluate PCFG against representative CFG variants using their official implementations whenever available; otherwise, we adopt established public reproductions that follow the corresponding method specifications. Method-specific parameters are set according to the respective implementations and kept fixed throughout the main comparisons. This protocol avoids benchmark-specific adjustment and ensures that each method is evaluated under a consistent configuration.

Table 4 summarizes the configurations used for the SD3.5 experiments. Standard CFG uses a guidance scale of 7 in the main comparison, while an additional scalar-scale scan is reported separately to characterize the effect of global guidance strength. The configuration of PCFG is determined through the ablation studies in Appendix E.1 and subsequently fixed across all nine main SD3.5 benchmark–budget settings. Since different methods employ distinct parameterizations, identical numerical values should not be interpreted as equivalent effective guidance strengths.

C.2 Evaluation Protocol

All methods are evaluated under matched experimental conditions. Within each benchmark and sampling budget, we use the same prompt set, random seed, image resolution, and number of sampling steps across methods. All images are generated at 1024×1024 resolution, and identical seed initialization is used to control stochastic variation across compared methods.

For SD3.5, experiments are conducted at 10, 20, and 30 sampling steps to evaluate the behavior of different guidance strategies under multiple sampling budgets. The same method configuration is retained across all three budgets and across COCO, DrawBench, and GenEval2. In particular, PCFG uses a single fixed configuration without benchmark- or budget-specific retuning.

We additionally evaluate PCFG on SDXL under a fixed 50-step deterministic DDIM protocol to examine its applicability to a noise-prediction backbone. Its separately evaluated configuration is given in Section 4.1. The SDXL comparison includes methods compatible with this evaluation setting, resulting in a more focused set of baselines than in the SD3.5 experiments. All compared methods follow the same prompt, resolution, seed, and sampling-step settings within this evaluation.

All ablations in the main paper and supplementary material use the same balanced set of 192 prompts, with 64 each from COCO, DrawBench, and GenEval2, at 10 sampling steps with seed 4416. Their reported means give equal weight to each benchmark.

Generated images are evaluated consistently using ImageReward, HPSv2.1, PickScore, CLIP similarity, and Aesthetic Score, with the same metric implementations and preprocessing procedures applied to all methods. Reported values are arithmetic means over the corresponding evaluation sets. The additional scalar-CFG scan characterizes sensitivity to the global guidance scale and supplies the best-scanned scalar references; it does not alter the fixed configurations used in the main method comparisons.

D Extended Comparisons with Scalar CFG

The main comparison uses standard CFG with a fixed guidance scale of $w = 7$. To provide a stronger scalar-CFG reference, we additionally select, for each benchmark, sampling budget, and metric, the best mean performance among ten evaluated guidance scales $\{3, 3.5, 4, 4.5, 5, 6, 7, 8, 10, 13\}$. This comparison is deliberately favorable to scalar CFG, since different metrics within the same experimental setting may correspond to different guidance scales and therefore do not constitute a single fixed configuration. Despite this advantage, PCFG achieves higher performance in 41 of the 45 benchmark–budget–metric comparisons while using one fixed configuration throughout. Figure 7 reports the signed differences for all comparisons, including the four cases in which scalar CFG performs better.

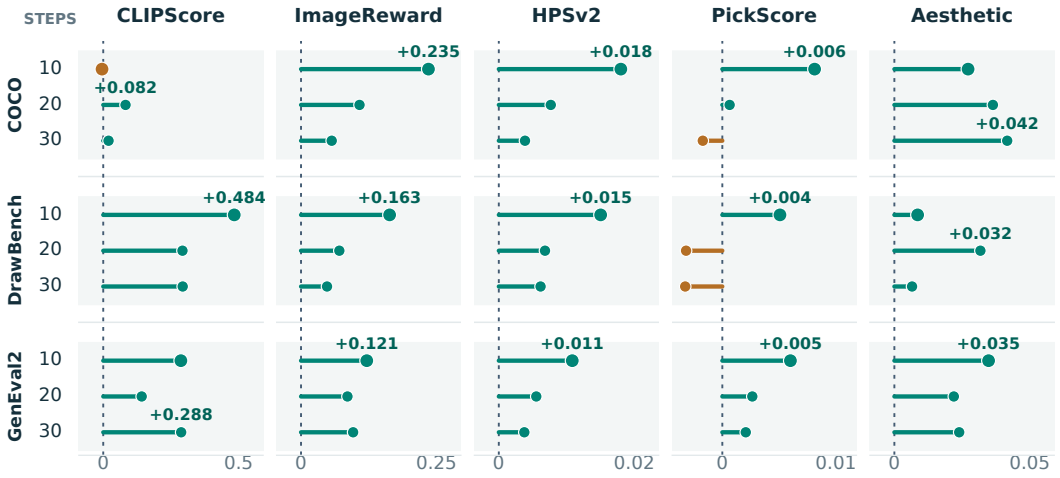


Figure 7: **Comparison with the best scanned scalar CFG.** For each benchmark, sampling budget, and metric, we report the signed difference between PCFG and the best scalar-CFG mean among the evaluated guidance scales. Positive values favor PCFG, while negative values favor scalar CFG. PCFG uses a single fixed configuration throughout, whereas the scalar-CFG reference is selected independently for each comparison. Green indicates improvements of PCFG over scalar CFG, whereas orange indicates the reverse.

Sensitivity to scalar guidance scale. Tables 5–7 provide a systematic comparison with scalar CFG across COCO, DrawBench, and GenEval2, respectively. For each benchmark, we evaluate ten fixed guidance scales w under 10-, 20-, and 30-step sampling budgets, covering a broad range of guidance strengths. PCFG is included as a fixed reference using the same configuration across all benchmarks and sampling budgets. The highest reported mean for each metric and budget, including ties, is highlighted in bold, while the PCFG row is shaded for clarity. These results characterize the sensitivity of scalar CFG to scale selection and provide a comprehensive reference for comparison with the adaptive spatial guidance of PCFG.

Table 5: **Comparison with scalar CFG under different sampling budgets on COCO.**

	10 sampling steps					20 sampling steps					30 sampling steps				
Setting	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.
$w = 3$	26.927	0.640	0.249	0.842	5.295	26.592	0.831	0.269	0.857	5.309	26.494	0.882	0.273	0.861	5.310
$w = 3.5$	27.003	0.707	0.254	0.845	5.313	26.705	0.889	0.273	0.861	5.326	26.565	0.930	0.278	0.864	5.330
$w = 4$	27.058	0.738	0.258	0.846	5.328	26.756	0.924	0.277	0.863	5.342	26.608	0.959	0.281	0.866	5.343
$w = 4.5$	27.073	0.748	0.260	0.846	5.334	26.781	0.955	0.279	0.864	5.355	26.668	0.988	0.283	0.867	5.351
$w = 5$	**27.081**	0.745	0.260	0.845	5.338	26.802	0.969	0.281	0.864	5.363	26.699	1.014	0.285	**0.868**	5.361
$w = 6$	27.033	0.701	0.258	0.840	5.330	26.826	0.992	0.283	0.863	5.373	26.745	1.044	0.288	**0.868**	5.378
$w = 7$	26.965	0.597	0.252	0.833	5.298	26.883	0.990	0.284	0.861	5.379	26.779	1.057	0.290	**0.868**	5.381
$w = 8$	26.778	0.443	0.243	0.823	5.238	26.932	0.966	0.282	0.858	5.373	26.802	1.056	0.290	0.866	5.388
$w = 10$	26.274	0.065	0.223	0.803	5.102	26.989	0.858	0.274	0.847	5.347	26.890	1.041	0.288	0.861	5.387
$w = 13$	25.147	-0.538	0.195	0.775	4.866	26.769	0.542	0.253	0.826	5.239	26.950	0.908	0.278	0.849	5.355
PCFG	27.077	**0.983**	**0.278**	**0.853**	**5.365**	**27.072**	**1.100**	**0.291**	**0.865**	**5.415**	**26.970**	**1.113**	**0.294**	0.867	**5.430**

Table 6: **Comparison with scalar CFG under different sampling budgets on DrawBench.**

	10 sampling steps					20 sampling steps					30 sampling steps				
Setting	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.
$w = 3$	28.374	0.510	0.250	0.847	5.346	28.397	0.702	0.267	0.858	5.369	28.395	0.752	0.273	0.861	5.378
$w = 3.5$	28.191	0.554	0.254	0.848	5.382	28.480	0.750	0.272	0.861	5.406	28.639	0.817	0.277	0.866	5.392
$w = 4$	28.157	0.605	0.258	0.851	5.386	28.771	0.822	0.276	0.863	5.379	28.844	0.846	0.280	0.866	5.393
$w = 4.5$	28.482	0.654	0.261	0.851	5.388	28.849	0.875	0.279	0.863	5.388	28.965	0.876	0.281	0.867	5.394
$w = 5$	28.498	0.701	0.262	0.851	5.381	28.748	0.886	0.280	**0.865**	5.404	28.894	0.901	0.283	**0.868**	5.400
$w = 6$	28.351	0.671	0.262	0.849	5.341	28.555	0.887	0.281	0.864	5.402	28.740	0.884	0.283	0.866	5.394
$w = 7$	28.313	0.655	0.260	0.844	5.301	28.829	0.892	0.281	0.862	5.388	28.838	0.931	0.285	0.867	5.365
$w = 8$	28.282	0.593	0.253	0.837	5.261	28.764	0.874	0.278	0.858	5.369	28.884	0.936	0.284	0.865	5.384
$w = 10$	27.560	0.345	0.239	0.825	5.144	28.748	0.862	0.272	0.850	5.317	28.902	0.904	0.281	0.859	5.341
$w = 13$	27.125	-0.100	0.218	0.804	4.892	28.652	0.666	0.258	0.837	5.194	29.013	0.780	0.270	0.848	5.235
PCFG	**28.981**	**0.864**	**0.277**	**0.855**	**5.396**	**29.141**	**0.963**	**0.288**	0.862	**5.438**	**29.306**	**0.984**	**0.291**	0.865	**5.407**

Table 7: **Comparison with scalar CFG under different sampling budgets on GenEval2.**

	10 sampling steps					20 sampling steps					30 sampling steps				
Setting	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.	CLIP	IR	HPS	Pick	Aest.
$w = 3$	30.839	0.916	0.272	0.863	5.474	31.291	1.151	0.291	0.874	5.535	31.348	1.197	0.296	0.878	5.524
$w = 3.5$	31.212	1.043	0.280	0.867	5.513	31.650	1.243	0.298	0.878	5.549	31.702	1.270	0.302	0.882	5.548
$w = 4$	31.435	1.124	0.285	0.870	5.524	31.715	1.303	0.303	0.881	5.577	31.940	1.334	0.306	0.884	5.569
$w = 4.5$	31.534	1.190	0.289	0.871	5.538	31.852	1.338	0.306	0.883	5.586	31.821	1.344	0.307	0.884	5.565
$w = 5$	31.568	1.221	0.291	0.871	5.551	31.860	1.352	0.307	0.883	5.591	31.827	1.367	0.309	0.885	5.566
$w = 6$	31.751	1.243	0.293	0.871	5.531	31.805	1.369	0.308	0.883	5.566	32.012	1.378	0.311	0.885	5.555
$w = 7$	31.592	1.205	0.290	0.868	5.498	31.935	1.372	0.307	0.882	5.549	31.990	1.363	0.309	0.884	5.537
$w = 8$	31.550	1.122	0.286	0.863	5.453	31.998	1.350	0.304	0.879	5.523	32.021	1.354	0.307	0.882	5.512
$w = 10$	31.101	0.825	0.269	0.850	5.284	32.045	1.291	0.297	0.873	5.423	32.041	1.254	0.299	0.876	5.437
$w = 13$	29.999	0.187	0.240	0.826	4.967	31.729	0.957	0.276	0.856	5.234	32.002	1.087	0.286	0.866	5.306
PCFG	**32.038**	**1.364**	**0.304**	**0.876**	**5.586**	**32.187**	**1.457**	**0.314**	**0.885**	**5.613**	**32.328**	**1.474**	**0.314**	**0.887**	**5.592**

Qualitative comparison across guidance scales and sampling budgets. Figure 8 complements the scalar sweeps with two examples at each sampling budget. The leftmost column shows PCFG with its fixed configuration; the remaining columns show scalar CFG at the ten scanned guidance scales. The three two-row groups use 10, 20, and 30 sampling steps, respectively.

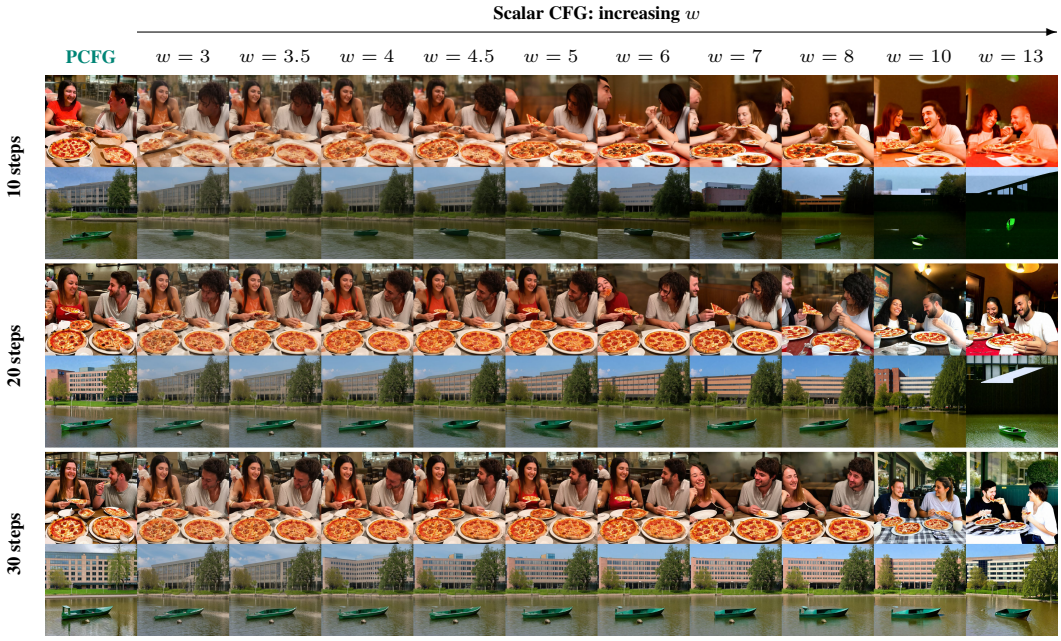


Figure 8: **Qualitative comparison with scalar CFG across sampling budgets.** Each two-row group repeats the same two examples, with PCFG in the first column and scalar CFG at different guidance scales in the remaining columns. Groups from top to bottom use 10, 20, and 30 sampling steps.

E Additional Ablations and Configuration Selection

To provide a more comprehensive analysis beyond the compact ablations reported in the main paper, we conduct additional controlled experiments on SD3.5 under a unified 10-step sampling setting. We use the same balanced set of 192 prompts as the main-paper ablations, with 64 each from COCO, DrawBench, and GenEval2 and seed 4416. These experiments systematically examine the design space of PCFG, covering the key factors that govern local-pattern construction, contextual comparison, and discrepancy-to-guidance conversion. Rather than isolating a single metric or configuration, we evaluate candidate designs jointly from the perspectives of text alignment, human preference, and perceptual quality, providing a more complete characterization of the robustness and sensitivity of the proposed policy.

E.1 Configuration Selection and Statistical Context

This subsection explains the selection of $(c, p, k, w_{\min}, w_{\max}, \gamma) = (3, 3, 21, 3, 18, 2)$. The $k = 13$ setting in Section 4.3 is an intermediate search reference; the final configuration is fixed for the main SD3.5 comparisons.

Local-pattern resolution across statistical contexts. Table 8 extends Table 3(A) to eight structures at four statistical contexts on the balanced 192-prompt set, with $(w_{\min}, w_{\max}, \gamma) = (3, 18, 2)$ fixed. The $(c, p) = (3, 3)$ representation leads all five metrics at $k = 19, 21, 25$ and four at $k = 17$, where $(2, 3)$ leads CLIP similarity, supporting its use beyond a single statistical context.

Table 8: **Effect of local-pattern resolution across statistical contexts.** We compare different amplitude resolutions c and patch sizes p under varying statistical contexts k. The selected $(c, p) = (3, 3)$ configuration is shaded. Bold: highest reported mean within each context, including ties.

Statistical context	c	p	IR ↑	HPS ↑	CLIP ↑	Pick ↑	Aest. ↑
$k = 17$	2	1	−0.927	0.182	25.104	0.769	4.541
	2	2	−0.865	0.185	25.551	0.772	4.584
	2	3	0.903	0.271	**29.015**	0.842	5.306
	3	1	−0.938	0.182	25.264	0.769	4.555
	3	2	−0.658	0.197	26.316	0.782	4.671
	3	3	**1.031**	**0.285**	28.888	**0.859**	**5.452**
	4	1	−0.942	0.182	25.298	0.769	4.529
	4	2	0.534	0.251	28.767	0.825	5.155
$k = 19$	2	1	−0.933	0.182	25.203	0.769	4.534
	2	2	−0.936	0.185	25.412	0.772	4.580
	2	3	0.815	0.266	28.982	0.837	5.261
	3	1	−0.915	0.182	25.264	0.769	4.548
	3	2	−0.741	0.194	25.981	0.779	4.640
	3	3	**1.039**	**0.285**	**28.990**	**0.859**	**5.441**
	4	1	−0.936	0.183	25.144	0.770	4.558
	4	2	0.248	0.238	28.308	0.815	5.045
$k = 21$	2	1	−0.964	0.182	25.182	0.769	4.518
	2	2	−0.935	0.185	25.336	0.771	4.566
	2	3	0.620	0.258	28.713	0.829	5.186
	3	1	−0.938	0.182	25.409	0.769	4.529
	3	2	−0.753	0.192	25.830	0.777	4.635
	3	3	**1.032**	**0.286**	**28.946**	**0.858**	**5.455**
	4	1	−0.924	0.183	25.257	0.770	4.550
	4	2	0.026	0.227	27.758	0.807	4.977
$k = 25$	2	1	−0.931	0.183	25.158	0.770	4.546
	2	2	−0.949	0.184	25.184	0.771	4.553
	2	3	0.387	0.244	28.516	0.818	5.101
	3	1	−0.967	0.182	25.245	0.769	4.514
	3	2	−0.813	0.190	25.748	0.775	4.605
	3	3	**1.034**	**0.287**	**28.824**	**0.859**	**5.455**
	4	1	−0.928	0.182	25.178	0.769	4.518
	4	2	−0.278	0.212	27.203	0.795	4.823

Guidance mapping and statistical context. Figure 9 complements the one-factor sensitivity analysis by examining the joint behavior of the guidance bounds and the robustness of the statistical context across different mappings. The upper row jointly varies $w_{\min}$ and $w_{\max}$ while keeping the remaining parameters fixed. Both ImageReward and HPSv2.1 exhibit broad high-performing regions rather than a common isolated optimum. Although larger $w_{\max}$ values can further improve some automatic metrics, we observe that overly strong guidance is more prone to visually undesirable effects such as over-saturation. We therefore adopt the more conservative setting $(w_{\min}, w_{\max}) = (3, 18)$, which maintains strong quantitative performance while avoiding excessive amplification.

The lower row examines the choice of statistical context k under five representative guidance mappings. Despite differences in their absolute performance, the mappings exhibit broadly consistent responses to context width: small contexts are generally less effective, while intermediate-to-large contexts yield stronger and more stable performance. In particular, $k = 21$ consistently lies within the high-performing regime for both ImageReward and HPSv2.1. We therefore select $k = 21$ as a robust context setting across different guidance configurations, rather than optimizing it for a single mapping.

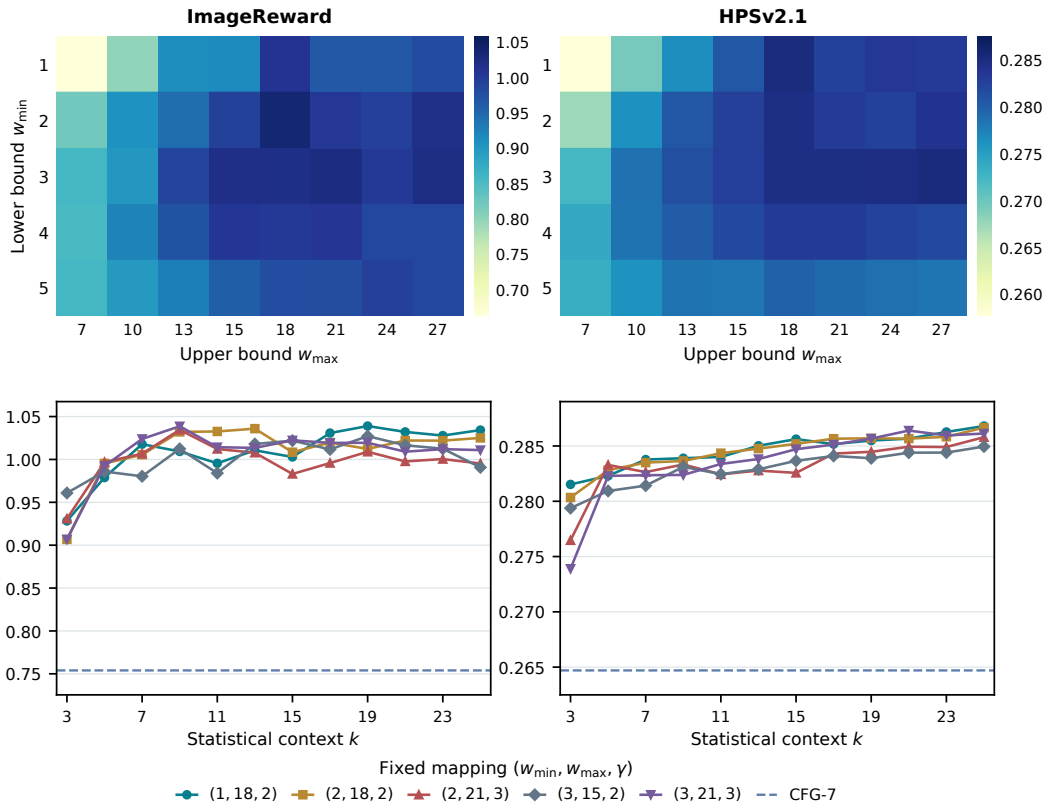


Figure 9: **Effects of guidance bounds and statistical context.** The upper row reports ImageReward and HPSv2.1 under joint variation of $w_{\min}$ and $w_{\max}$, with $(c, p, k, \gamma) = (3, 3, 21, 2)$ fixed. The lower row retains $(c, p) = (3, 3)$ and evaluates the statistical context k under the five fixed mappings listed in the legend. Dashed lines indicate the CFG-7 baseline evaluated on the same prompt set.

Final configuration. Section 4.3 reports the ablations on local-pattern representation and guidance mapping in compact form; here we extend them with the joint study of the guidance bounds and the statistical context. Combining both lines of evidence, we adopt $(c, p, k, w_{\min}, w_{\max}, \gamma) = (3, 3, 21, 3, 18, 2)$ for all main SD3.5 experiments. The configuration is chosen to balance quantitative performance and visual quality rather than to match the maximum of any individual automatic metric. Once selected, it is fixed across all benchmarks and sampling budgets without further retuning.

F ADDITIONAL QUALITATIVE COMPARISONS

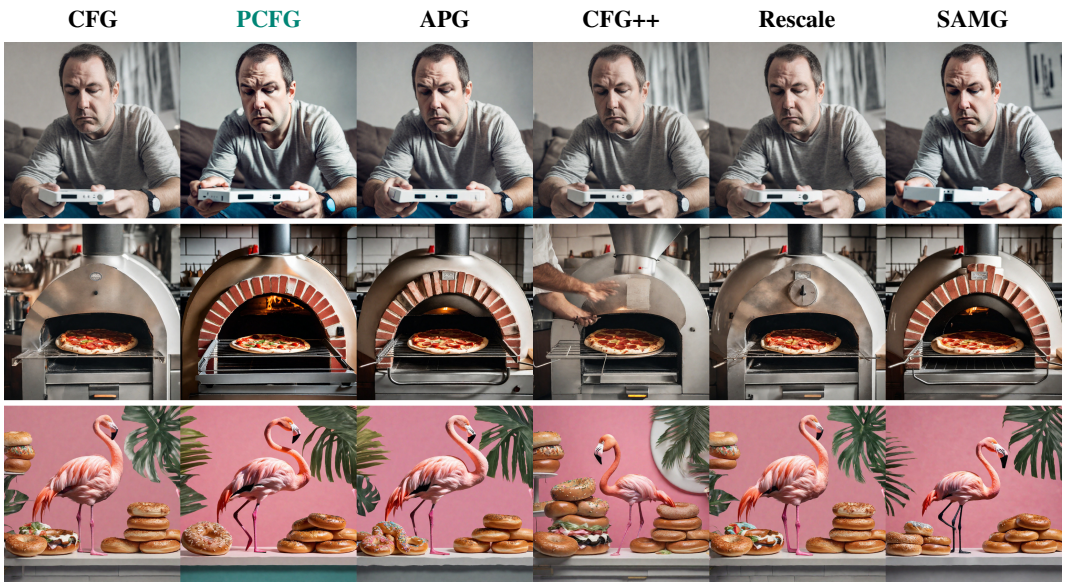


Figure 10: **SDXL qualitative comparison.** Each row shows images generated from the same prompt by the six methods labeled above.