MORE THAN A SINGLE GUIDANCE STRENGTH: RETHINKING CLASSIFIER-FREE GUIDANCE IN NEXT-TOKEN AUTOREGRESSIVE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance (CFG) in next-token autoregressive image generation typically uses the difference between conditional and unconditional predictions as a fixed guidance direction, controlled by a single scalar strength. This raises a basic question: **Is a single guidance strength sufficient to characterize how CFG should be applied?** First, we view CFG as increasing conditional evidence while limiting its KL deviation from the model's conditional prediction, and further study this trade-off under the fisher geometry. In this geometry, we decompose the CFG direction into two components with distinct effects on token prediction. One preferentially amplifies class-specific tokens, while the other preserves tokens favored by both conditional and unconditional predictions. We then investigate how the guidance direction should be combined for a given guidance strength and observe a systematic shift in the preferred balance: weak guidance favors more class-specific amplification, while stronger guidance requires greater preservation of jointly supported tokens. We further study how guidance should be applied for a given strength and composition, and find that the update strategy also matters: successive local updates become more useful with stronger guidance and greater emphasis on class-specific amplification. Together, these findings show that a single guidance strength is insufficient to characterize CFG, as its behavior also depends on how the guidance direction is composed and how the guidance is applied. Experiments across multiple autoregressive backbones validate these findings and show consistent improvements in both class-to-image and text-to-image generation.

1 INTRODUCTION

Autoregressive (AR) visual generation has advanced rapidly in recent years, ranging from discrete-token text-to-image models to scalable next-token and next-scale image-generation frameworks (Yu et al., 2022; Tian et al., 2024; Wang et al., 2025). Classifier-free guidance (CFG) has become a key inference technique in these models to improve conditional alignment and generation quality (Ho & Salimans, 2022). At each autoregressive step, the standard CFG constructs a guidance signal from the difference between conditional and unconditional predictions and amplifies this signal with a single scalar strength, as illustrated in Fig. 1(a). The standard CFG therefore treats guidance control largely as a one-dimensional choice of guidance strength. Increasing this strength often improves conditional alignment and sample fidelity, but can reduce diversity and eventually degrade generation quality under excessively strong guidance (Sun et al., 2024).

However, reducing CFG to a single guidance strength leaves two basic assumptions largely unquestioned. First, the standard CFG uses a fixed guidance direction in different guidance strengths. This raises the question of whether the same direction should remain appropriate as the guidance strength changes. Second, the prediction distribution itself evolves as guidance is applied, raising the question of whether guidance should always be applied through a one-shot update, as is typically done in standard CFG. Taken together, these considerations motivate us to move beyond guidance strength and examine two further aspects of CFG: **how to compose the guidance direction** and **how to apply the guidance**.

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.

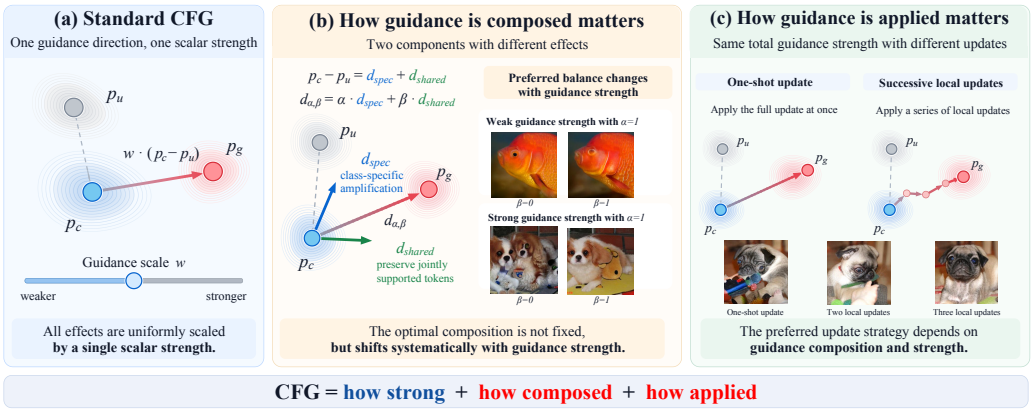


Figure 1: **More than a single guidance strength.** Standard CFG controls the full guidance signal with one scalar strength. We show that its behavior additionally depends on how the guidance direction is composed and how the guidance is applied.

We begin our analysis at a single next-token prediction step, where CFG directly modifies the model's predictive distribution over the token vocabulary. From this distributional perspective, standard CFG can be interpreted as amplifying conditional evidence while constraining the KL divergence from the model's original conditional prediction, an interpretation we refer to as the **Evidence–KL view**. This view allows us to examine how guidance direction and strength jointly determine the guided prediction, rather than viewing CFG merely as a logit-space extrapolation along a fixed direction. The KL divergence in this view also induces a local geometry on the categorical predictive distribution, characterized by the Fisher metric, which we refer to as the **Fisher Geometry view**.

Building on this local geometry, we decompose the standard CFG direction into orthogonal and parallel components and examine how each affects next-token prediction. The two components play distinct roles: the orthogonal component primarily amplifies class-specific token preferences, whereas the parallel component better preserves tokens supported by both conditional and unconditional predictions. Building on this functional distinction, we further find that the preferred composition shifts systematically with guidance strength: weak guidance favors a greater emphasis on the orthogonal component, while stronger guidance favors a larger parallel contribution, as shown in Fig. 1(b).

With the relationship between strength and composition established, we next consider how guidance should be applied once its strength and composition are specified. The Fisher geometry is defined locally by the current predictive distribution and therefore changes as the distribution evolves during the update. We therefore compare the standard one-shot update with successive local updates under the same total guidance strength and composition, as indicated in Fig. 1(c). We find that successive local updates become increasingly beneficial as guidance strength increases, particularly when the guidance composition places greater emphasis on the orthogonal component. In general, our findings show that **guidance strength alone is insufficient: CFG also depends on how the guidance direction is composed and how the guidance is applied.** Our main contributions are summarized as follows:

- **New perspectives on CFG for next-token autoregressive Generation.** We reinterpret CFG through the Evidence–KL view and its induced local Fisher geometry, enabling a functional decomposition of the guidance direction for next-token prediction.
- **More than a single guidance strength.** We show that the preferred guidance composition varies with guidance strength, and the update strategy remains important even when strength and composition are fixed.
- **Consistent improvements across multiple autoregressive backbones.** We validate these guidance mechanisms across multiple autoregressive backbones under both class and text conditioning, demonstrating consistent improvements over standard CFG.

2 Related Work

Autoregressive Visual Generation. Autoregressive image generation has evolved from pixel-wise modeling to sequence modeling over learned visual tokens, enabling scalable Transformer-based generation (van den Oord et al., 2016; 2017; Esser et al., 2021; Ramesh et al., 2021; Yu et al., 2022). With advances in tokenization and model scaling, next-token prediction has become a competitive paradigm for both class-conditional and text-conditional image generation (Sun et al., 2024). Recent work has further expanded visual autoregressive modeling beyond conventional raster-order token prediction, including coarse-to-fine next-scale prediction (Tian et al., 2024), randomized factorization orders (Yu et al., 2025), flexible prediction units (Ren et al., 2025), and alternative visual representations based on compact one-dimensional tokens (Yu et al., 2024), bitwise codes (Han et al., 2025), and continuous-valued tokens (Li et al., 2024; Han et al., 2026).

Classifier-Free Guidance (CFG). CFG was introduced in diffusion models by extrapolating conditional and unconditional predictions at sampling time, providing a simple mechanism to improve conditional fidelity (Ho & Salimans, 2022). Subsequent work has shown that CFG behavior depends on more than a fixed global scale, exploring guidance scheduling, alternative references, projection-based or manifold-aware corrections, and different interpretations of its sampling dynamics (Lin et al., 2024; Kynkäänniemi et al., 2024; Karras et al., 2024; Sadat et al., 2025; Chung et al., 2025; Kwon et al., 2025; Bradley & Nakkiran, 2025; Papalampidi et al., 2026; Chen et al., 2025a; Galashov et al., 2026). More recently, CFG in visual autoregressive generation has been studied through training-based guidance absorption (Chen et al., 2025b), improved stability during sequential decoding (Xu et al., 2026), and specialized strategies for multi-scale generation or generated-prefix consistency (Shin et al., 2026; Nguyen et al., 2026; Liao et al., 2026). In contrast, we study the mechanisms of standard CFG at the next-token prediction level, rather than changing the training objective or introducing guidance strategies tailored to the autoregressive generation.

Fisher Geometry. Information geometry views a parametric family of probability distributions as a statistical manifold whose local geometry is characterized by the Fisher–Rao metric, which arises from the second-order expansion of KL divergence and is invariant to smooth reparameterizations (Fisher, 1922; Rao, 1945; Čencov, 1982; ichi Amari, 2016; Nielsen, 2020). The Fisher geometry has been widely used to develop natural-gradient, Riemannian optimization, and KL-constrained optimization methods (Amari, 1998; Pascanu & Bengio, 2014; Ollivier, 2015; Martens, 2020; Schulman et al., 2015), and more recently for second-order optimization and generative modeling of discrete data (GOMES et al., 2025; Wu et al., 2024; Cheng et al., 2024; Davis et al., 2024). In this work, we use the local Fisher geometry to analyze CFG at the level of categorical next-token prediction.

3 Preliminaries

Evidence–KL View. We introduce our analysis in a single next-token prediction step, where the CFG directly modifies a categorical distribution over the token vocabulary. Let $z_c, z_u \in \mathbb{R}^V$ denote the conditional and unconditional logits, respectively, and define

$$p_z = \mathrm{Softmax}(z), \qquad p_c = p_{z_c}, \qquad e = z_c - z_u. \tag{1}$$

Standard CFG constructs the guided logits and sampling distribution as

$$z_w = z_c + we, \qquad q_w = p_{z_w}, \qquad w \geq 0, \tag{2}$$

where e denotes the conditional signal and w is the scalar guidance strength (Ho & Salimans, 2022; Chen et al., 2025b). For $w > 0$, standard CFG admits the equivalent formulation

$$q_w = \arg\max_{q \in \Delta^{V-1}} \left[\mathbb{E}_{v \sim q}[e(v)] - \frac{1}{w} D_{\mathrm{KL}}(q \| p_c) \right]. \tag{3}$$

Thus, the standard CFG increases conditional signal while limiting deviation from the original conditional prediction, inspired by the general principle of relative entropy-regularized and trust-region updates (Peters et al., 2010; Schulman et al., 2015). The detailed derivation is provided in Appendix A.1.

Fisher Geometry View. We further study this formulation from a fisher geometry view, with

$$F_c = \mathrm{Diag}(p_c) - p_c p_c^\top, \qquad \langle a, b \rangle_{F_c} = a^\top F_c b, \qquad \|a\|_{F_c}^2 = a^\top F_c a. \tag{4}$$

For a local perturbation wd with $q_{w,d} = \mathrm{Softmax}(z_c + wd)$ and $\mathcal{J}_e(q) = \mathbb{E}_{v\sim q}[e(v)]$, we can obtain:

$$\mathcal{J}_e(q_{w,d}) - \mathcal{J}_e(p_c) = w\langle d, e\rangle_{F_c} + \mathcal{O}(w^2), \qquad D_{\mathrm{KL}}(q_{w,d}\|p_c) = \frac{w^2}{2}\|d\|_{F_c}^2 + \mathcal{O}(w^3). \tag{5}$$

Eq. 5 shows that the evidence gain depends on Fisher alignment $\langle d, e\rangle_{F_c}$, while the KL deviation depends on the squared Fisher length $\|d\|_{F_c}^2$. Overall, these relations show that the local Evidence–KL behavior depends jointly on the guidance direction d and its strength w. Detailed derivations are provided in Appendix A.2.

4 Beyond a Single Guidance Strength

To investigate whether CFG can be fully characterized by a single guidance strength, we build on Sec. 3 and proceed in three steps, as shown in Fig. 2: (1) identifying distinct functional roles within the CFG direction; (2) investigating how their preferred composition changes with guidance strength; (3) examining how guidance should be applied once strength and composition are fixed.

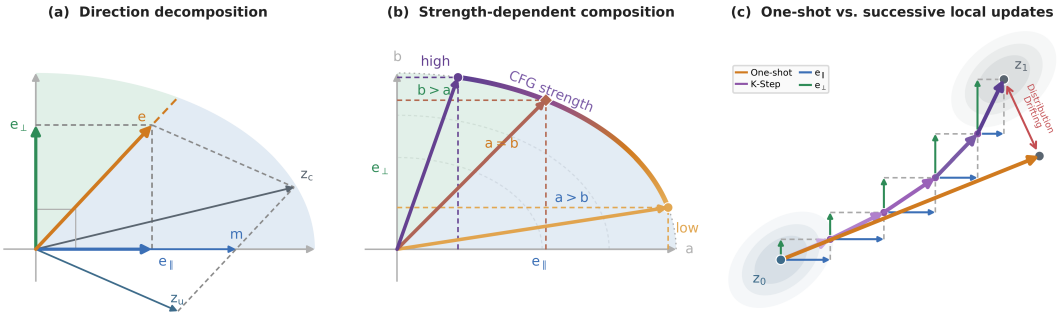


Figure 2: **Overview of guidance composition and application. (a)** The CFG direction e is decomposed into orthogonal and parallel components under the local Fisher geometry. **(b)** As guidance strength increases, the preferred coefficients move from $a > b$ to $a = b$ and eventually to $b > a$. **(c)** A one-shot update applies the decomposition defined at the starting distribution, while successive local updates recompute it along the guidance path.

4.1 Functional Roles within the CFG Direction

We first explore whether the standard CFG direction $e = z_c - z_u$ contains components with distinct roles in next-token prediction. Let $p_u = \mathrm{Softmax}(z_u)$. To construct a symmetric reference for the probability structure supported by both the conditional and unconditional predictions, we define the midpoint logits

$$m = \frac{z_c + z_u}{2}. \tag{6}$$

The corresponding categorical distribution satisfies

$$p_m(v) = \frac{\sqrt{p_c(v)p_u(v)}}{\sum_j \sqrt{p_c(j)p_u(j)}}, \tag{7}$$

so the midpoint emphasizes tokens that receive support from both predictions. We use m as a shared-support reference and decompose the CFG direction under the local Fisher geometry as

$$e_\parallel = \frac{\langle e, m\rangle_{F_c}}{\|m\|_{F_c}^2} m, \qquad e_\perp = e - e_\parallel, \tag{8}$$

with $\langle e_\perp, m\rangle_{F_c} = 0$. The standard CFG can then be written as

$$z_w = z_c + w\left(e_\perp + e_\parallel\right), \tag{9}$$

which assigns the same coefficient to the two components. Alternative references and the midpoint derivation are also discussed in Appendix B.1.

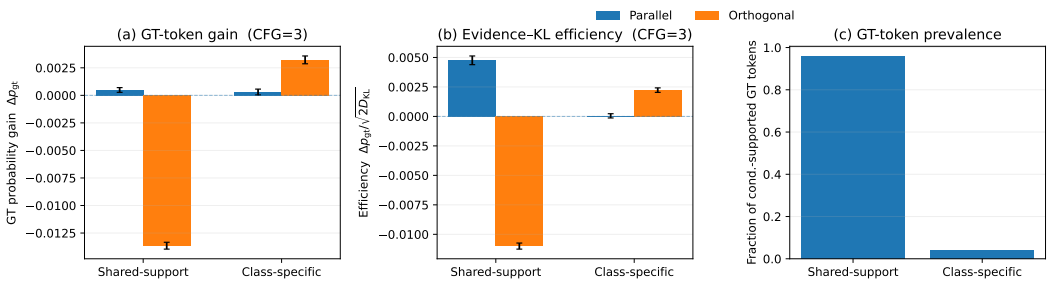


Figure 3: **Functional roles of the CFG components.** We use AliTok-L (Wu et al., 2026) on ImageNet for analysis. The parallel component better preserves shared-support GT tokens, while the orthogonal component more strongly amplifies class-specific GT tokens. **(a)** GT-token probability change. **(b)** Change normalized by KL. **(c)** Relative prevalence of the two token groups.

Token-level Functional Roles. To isolate the effect of each component at a single next-token prediction step from errors accumulated during decoding, we condition the model on the ground-truth(GT) prefix $x^{\mathrm{gt}}_{<t}$ and measure the change in probability assigned to the next GT token x^{gt}_t. We define a token as *shared-support* if

$$x^{\mathrm{gt}}_t \in \mathrm{Top}_k(p_c) \cap \mathrm{Top}_k(p_u), \tag{10}$$

and as *class-specific* if

$$x^{\mathrm{gt}}_t \in \mathrm{Top}_k(p_c) \setminus \mathrm{Top}_k(p_u). \tag{11}$$

Using Eq. 8, we apply the two components separately to the same next-token prediction. Fig. 3(a) shows that the parallel component better preserves shared-support GT tokens, while the orthogonal component more strongly favors class-specific tokens. To compare the gain in ground-truth token probability against the guidance from the original prediction p_c, we further normalize it by $\sqrt{2D_{\mathrm{KL}}(q\|p_c)}$, where q denotes the distribution obtained after applying the corresponding component. Fig. 3(b) shows that the same functional distinction remains after normalization and Fig. 3(c) shows that shared-support tokens dominate the ground-truth population. Overall, these results show that $e_\perp$ and $e_\parallel$ serve different roles despite being equally weighted in standard CFG, motivating us to investigate their relative composition next. Additional analyzes are provided in the Appendix B.1.

4.2 Guidance Composition Depends on Guidance Strength

We investigate whether the preferred composition remains optimal across different guidance strengths. We write the guided logits as

$$z_{a,b} = z_c + ae_\perp + be_\parallel, \tag{12}$$

where a and b control the strengths of the orthogonal and parallel components, respectively, and we characterize their relative composition by $\lambda = b/a$. Thus, λ controls the relative emphasis of the two components, with $\lambda = 1$ recovering standard CFG.

We first study the range $\lambda \in [0, 1]$, from pure orthogonal guidance to the standard CFG composition. At each guidance strength, we sweep λ while keeping all other sampling settings fixed. As shown in Fig. 4(a), weak guidance favors smaller λ, while the optimum progressively shifts toward a larger parallel contribution as guidance becomes stronger. Fig. 4(b) summarizes this trend, with the preferred $\lambda^\star$ increasing toward the standard CFG composition at $\lambda = 1$. As the preferred composition approaches $\lambda = 1$ under stronger guidance, we further extend the search beyond $\lambda = 1$. Fig. 4(c) shows that $\lambda^\star$ continues to increase with a and eventually exceeds 1.

These results show that the standard CFG composition is not universally optimal, as different compositions are preferred at different guidance strengths. Guidance composition should therefore be considered jointly with guidance strength, rather than treated as a fixed choice. Further analyzes and an interpretation of this strength-dependent shift are provided in Appendix B.2.

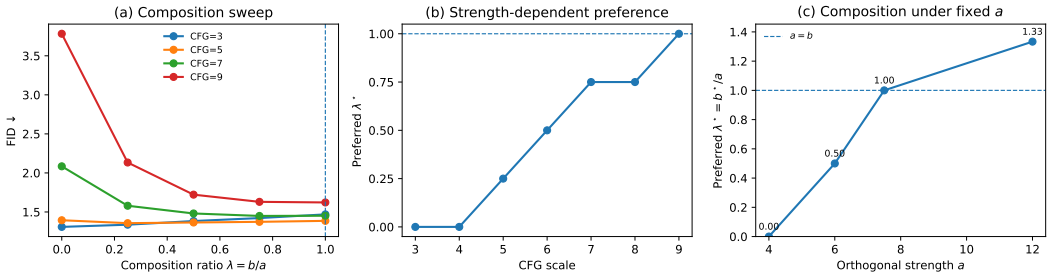


Figure 4: **Guidance composition depends on guidance strength.** Results are reported on AliTok-L. **(a)** Composition sweeps at representative strengths. **(b)** Preferred $\lambda^\star$ across strengths. **(c)** Preferred $\lambda^\star$ beyond the standard CFG composition.

4.3 How Guidance Is Applied Also Matters

Once the guidance strength and composition are specified, the remaining decision is how the update is applied. A one-shot update keeps the decomposition fixed in the starting distribution. However, the Fisher geometry changes with the prediction evolves, so the initial decomposition may not remain appropriate throughout the update. We thus divide the update into K local steps and recompute the decomposition after each step. Starting from $z^{(0)} = z_c$, we update

$$z^{(j+1)} = z^{(j)} + \frac{a}{K}\left(e_\perp^{(j)} + \lambda e_\parallel^{(j)}\right). \tag{13}$$

At each intermediate state,

$$q^{(j)} = \mathrm{Softmax}(z^{(j)}), \qquad e = e_\perp^{(j)} + e_\parallel^{(j)}, \qquad \left\langle e_\perp^{(j)}, m \right\rangle_{F(q^{(j)})} = 0. \tag{14}$$

The evidence e and reference m remain fixed within the same next-token prediction, while the current distribution $q^{(j)}$ and its Fisher geometry change along the update. Thus, $K = 1$ recovers the one-shot update in Sec. 4.2, whereas $K > 1$ refreshes the local decomposition while keeping a and λ fixed. To study how successive local updates interact with guidance strength and composition, we first isolate the effect of guidance strength at $\lambda = 0$. We then vary $\lambda \in [0, 1]$ and examine the gain from successive updates at representative strengths.

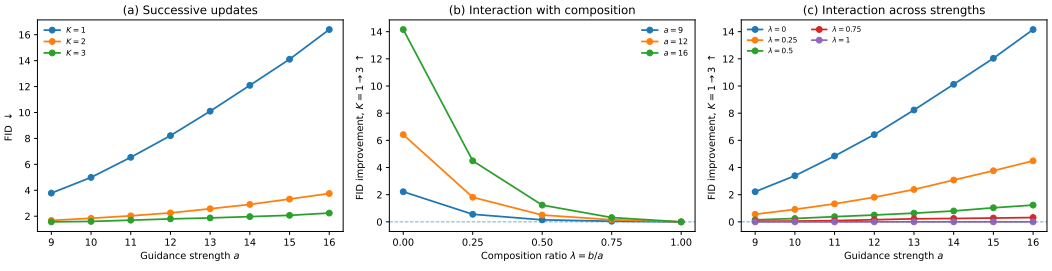


Figure 5: **Effect of successive local updates under different guidance strengths and compositions.** Results are reported on AliTok-L. **(a)** FID for $K = 1, 2, 3$ across guidance strengths under pure orthogonal guidance ($\lambda = 0$). **(b)** FID improvement from $K = 1$ to $K = 3$ as the composition ratio λ varies, shown at representative guidance strengths. **(c)** FID improvement from $K = 1$ to $K = 3$ across the evaluated strength range for different values of λ.

Under pure orthogonal guidance, successive local updates have a relatively smaller effect at lower guidance strengths, but their benefit grows substantially as a increases (Fig. 5(a)). This gain steadily decreases as λ approaches 1 at representative strengths (Fig. 5(b)). The sweep in Fig. 5(c) confirms the same interaction: successive updates are most useful under stronger, orthogonal-dominant guidance, while their effect nearly disappears at $\lambda = 1$. These results show that the effect of successive updates depends jointly on guidance strength and composition. Appendix B.3 analyzes this dependence through the local Fisher geometry and provides additional analysis results.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks and Backbones. We evaluate our findings across class-conditional and text-conditioned image generators. For class-conditional generation, we evaluate LlamaGen-L (Sun et al., 2024), RAR-XL/XXL (Yu et al., 2025), and AliTok-L/XL (Wu et al., 2026) on ImageNet (Deng et al., 2009) at 256×256 resolution. Despite differences in their factorization and tokenization designs, all these models follow the same next-token autoregressive formulation over discrete visual tokens. We further evaluate SimpleAR-0.5B-RL (Wang et al., 2025) on GenEval (Ghosh et al., 2023) for text-conditioned generation.

Evaluation Settings. For class-conditional generation, we follow the standard ImageNet-256 evaluation protocol and report FID, Inception Score (IS), precision, and recall, using FID as the primary metric for mechanism analysis. For text-conditioned generation, we follow the standard GenEval protocol and report the overall score together with per-category results. In the mechanism experiments, we vary only the guidance strength, composition ratio λ, and number of successive local updates K, while keeping all other settings fixed. For the final generation results, we compare standard CFG, SoftCFG (Xu et al., 2026), and the guidance configurations derived from our analysis. Broader parameter sweeps are provided in Appendix D.1.

5.2 GENERALITY OF THE GUIDANCE MECHANISMS

The two guidance patterns identified on AliTok in Sec. 4, concerning guidance composition and successive local updates, are further evaluated. We test their generality on LlamaGen, which provides a different class-conditional autoregressive backbone, and on SimpleAR, which further extends the evaluation to text-conditioned generation.

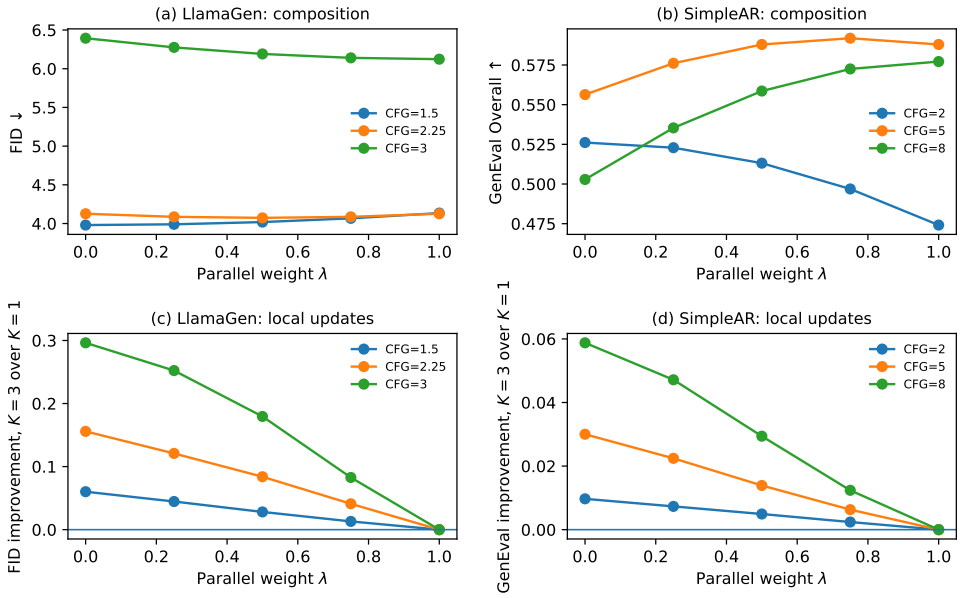


Figure 6: **Generality of the guidance mechanisms on LlamaGen and SimpleAR.** (a,b) Generation performance across composition ratios λ at representative guidance strengths, with $K = 1$. (c,d) Improvement from $K = 1$ to $K = 3$ under the same guidance strength and composition.

Guidance Composition across Strengths. We further evaluate how guidance composition varies with strength on LlamaGen and SimpleAR with $K = 1$, using λ to vary the parallel component relative to the orthogonal component across representative guidance strengths. Fig. 6(a,b) shows the same strength-dependent shift on both models: smaller λ is preferred under weaker guidance, while larger λ becomes preferable as guidance strengthens.

Successive Local Updates across Strengths and Compositions. We compare successive local updates on LlamaGen and SimpleAR by measuring the gain of $K = 3$ over $K = 1$ at the same guidance strength and composition. Fig. 6(c,d) shows the same interaction observed on AliTok: the gain increases with the guidance strength and decreases as λ grows on both models. More analysis results are provided in Appendix D.2.

5.3 Improving CFG on Class-Conditional Generation

The above mechanisms allow us to narrow the configurations that need to be evaluated, rather than independently enumerating all combinations of guidance strength, composition, and successive local updates. As guidance strength changes, we first identify the corresponding preferred composition, and then use the dependence of successive local updates on this composition to further reduce the remaining evaluations. We use this procedure to obtain the results reported in Tab. 1. Appendix C provides the complete procedure and analysis of the number of evaluations, and compares the method to a greedy search starting from standard CFG.

Table 1: **Class conditional generation on ImageNet** 256×256**.** The upper block reports representative published results, while the lower block compares standard CFG, SoftCFG, and ours using the same released checkpoints and evaluation protocol.

Type	Model	Params	Method	FID ↓	IS ↑	Pre. ↑	Rec. ↑
GAN	StyleGAN-XL (Sauer et al., 2022)	166M	–	2.30	265.1	0.78	0.53
Diff.	DiT-XL/2 (Peebles & Xie, 2023)	675M	Standard CFG	2.27	278.2	0.83	0.57
Mask	MaskGIT (Chang et al., 2022)	227M	–	6.18	182.1	0.80	0.51
MAR	MAR-H (Li et al., 2024)	943M	Standard CFG	1.55	303.7	0.81	0.62
VAR	VAR-d30 (Tian et al., 2024)	2.0B	Standard CFG	1.92	323.1	0.82	0.59
AR	LlamaGen-L	343M	Standard CFG	3.72	**249.1**	**0.83**	0.51
			SoftCFG	3.53	245.2	0.82	0.52
			Ours	**3.25**	238.0	0.82	**0.53**
	RAR-XL	955M	Standard CFG	1.51	308.4	**0.81**	0.62
			SoftCFG	1.49	304.7	0.80	0.62
			Ours	**1.44**	**310.2**	0.80	**0.63**
	RAR-XXL	1.5B	Standard CFG	1.48	325.6	**0.81**	0.62
			SoftCFG	1.47	319.8	0.81	**0.63**
			Ours	**1.42**	**326.3**	0.80	0.62
	AliTok-L	177M	Standard CFG	1.39	**325.6**	**0.79**	0.64
			SoftCFG	1.37	278.4	0.77	**0.65**
			Ours	**1.27**	308.4	0.78	0.64
	AliTok-XL	662M	CFG	1.32	**312.6**	0.80	0.64
			SoftCFG	1.27	298.4	0.79	**0.65**
			Ours	**1.24**	306.2	**0.80**	0.64

Tab. 1 shows that our method achieves better FID than both standard CFG and SoftCFG across all five evaluated autoregressive backbones, while precision and recall remain broadly comparable. These consistent gains show that the guidance relationships identified above also improve generation quality in practice. Appendix D.3 further examines the search effort, the trade-off between FID and Inception Score, and the stability of the results across random seeds.

5.4 Extension to Text-Conditional Generation

Sec. 5.2 shows that the guidance behaviors identified above also persist on SimpleAR. We further evaluate their practical effect on text-conditioned generation using the released SimpleAR-0.5B-RL checkpoint on GenEval. Tab. 2 shows a clear improvement in the overall GenEval score over both standard CFG and SoftCFG, indicating that the same findings can also translate into better generation performance under text-conditioning.

Table 2: **Text conditioned generation on GenEval.** All methods use the same released checkpoint and evaluation protocol. We report the overall score, image accuracy, prompt accuracy, and the six category scores.

Method	Overall Evaluation			Prompt Categories					
	Overall ↑	Image Acc. ↑	Prompt Acc. ↑	Single Obj. ↑	Two Obj. ↑	Counting ↑	Colors ↑	Position ↑	Color Attr. ↑
Standard CFG	0.585	0.579	0.787	0.975	0.811	0.291	0.814	0.253	0.368
SoftCFG	0.588	0.581	0.796	0.984	0.813	0.297	0.814	**0.258**	0.360
Ours	**0.611**	**0.604**	**0.826**	**0.991**	**0.869**	**0.334**	**0.827**	0.240	**0.405**

5.5 Visualizing the Guidance Mechanisms

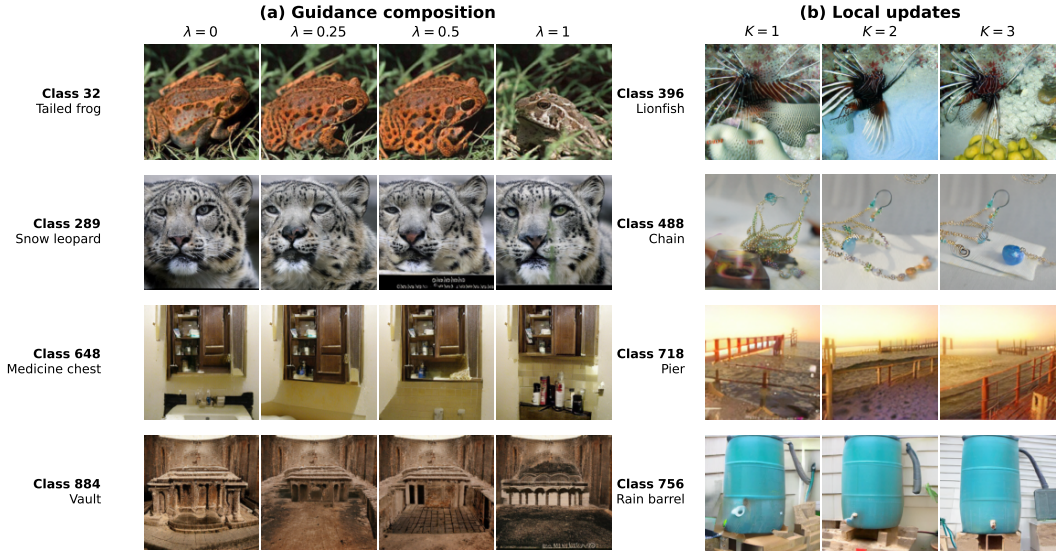


Figure 7: **Qualitative visualization of the guidance mechanisms on AliTok-L.** (a) Guidance composition at guidance scale 3 and $K = 1$, varying λ. (b) Successive local updates at guidance scale 15 and $\lambda = 0.25$, varying K.

Fig. 7 provides qualitative examples of the two guidance effects on AliTok-L. Fig. 7(a) shows how generation quality changes with guidance composition, while Fig. 7(b) shows the effect of successive local updates when both guidance strength and composition are fixed. Additional examples in more classes, along with comparisons to standard CFG are provided in Appendix D.4.

6 Conclusion

In this work, we rethink classifier-free guidance in autoregressive image generation beyond a single guidance strength. Using the local Fisher geometry of next-token categorical distributions, we show that the composition of the guide changes with strength, while successive local updates improve the effectiveness of the guide. These findings generalize across different autoregressive backbones and lead to consistent performance improvements. Overall, our results show that CFG depends not only on guidance strength, but also on how the guidance direction is composed and how the update is applied.

Limitations and Future Work Our study is limited to autoregressive image generation. Since language models share the same next-token formulation, it would be interesting to examine whether similar guidance behaviors also arise in language generation. Another direction is to dynamically adapt the strength, composition, and local updates of the guide to the evolving generation state.

AI USE STATEMENT

In this work, we used generative AI tools only to organize the writing structure and polish the language. We did not use them to develop analyses, run experiments, derive mathematics, generate figures, tables, or datasets, or produce citations or scientific claims. All AI suggestions were reviewed and revised by the authors. The mathematics, experiments, results, and conclusions were completed independently by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This work studies inference-time guidance for autoregressive image generation using existing models and public benchmarks. We do not conduct human-subject experiments, collect personal data, or introduce a new dataset. The authors take responsibility for the reported experiments and conclusions.

REPRODUCIBILITY STATEMENT

We provide the detailed search procedures, and experimental settings needed to reproduce our results using publicly released checkpoints. Code will be released upon acceptance.

REFERENCES

Shun Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zrWNtzSZsf.

N. N. Čencov. *Statistical decision rules and optimal inference*, volume 53. American Mathematical Society, 1982.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11315–11325, June 2022.

Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=gM6aboVgTO.

Huayu Chen, Hang Su, Peize Sun, and Jun Zhu. Toward guidance-free AR visual generation via condition contrastive alignment. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=kGvXIlIVLM.

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5fybcQZ0g4.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail Ilkan Ceylan, Michael M. Bronstein, and Joey Bose. Fisher flow matching for generative modeling over discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6jOScqwdHU.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 248–255, 2009.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A: Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 01 1922. doi: 10.1098/rsta.1922.0009.

Alexandre Galashov, Ashwini Pokle, Arnaud Doucet, Arthur Gretton, Mauricio Delbracio, and Valentin De Bortoli. Learn to guide your diffusion model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=l8XOk4ylBH.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Wbr51vK331.

Damien MARTINS GOMES, Yanlei Zhang, Eugene Belilovsky, Guy Wolf, and Mahdi S. Hosseini. Adafisher: Adaptive second order optimization via fisher information. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=puTxuiK2qO.

Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Shu-Tao Xia, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Ndnwg9oOQO.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744, 2025.

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

Shun ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, and Samuli Laine Timo Aila. Guiding a diffusion model with a bad version of itself. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=bg6fVPVs3s.

Mingi Kwon, Shin seong Kim, Jaeseok Jeong, Yi Ting Hsiao, and Youngjung Uh. Tcfg: Tangential damping classifier-free guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2620–2629, June 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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nAIhvNy15T.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=VNBIF0gmkb.

Xinyao Liao, Qiyuan He, Yicong Li, Jiayin Zhu, Xiaoye Qu, Wei Wei, and Angela Yao. Vpg: Visual prefix guidance for autoregressive image and video generation. *arXiv preprint arXiv:2605.30317*, 2026.

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 (WACV)*, pp. 5404–5411, 2024.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

Ky Dan Nguyen, Hoang Lam Tran, Anh-Dung Dinh, Daochang Liu, Weidong Cai, Xiuying Wang, and Chang Xu. Rethinking visual autoregressive sampling with information-grounding guidance. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4PL5wouooK.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 2020. doi: 10.3390/e22101100. URL https://www.mdpi.com/1099-4300/22/10/1100.

Yann Ollivier. Riemannian metrics for neural networks i: feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 06 2015.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajic, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=z9YC9bvfUL.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations*, 2014.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1607–1612, 2010.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8821–8831. PMLR, 2021.

C.R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 1945.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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*, 2025. URL https://openreview.net/forum?id=e2ONKX6qzJ.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*. Association for Computing Machinery, 2022. doi: 10.1145/3528233.3530738.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37. PMLR, 2015.

Youngwoo Shin, Jiwan Hur, and Junmo Kim. SSG: Scaled spatial guidance for multi-scale visual autoregressive generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=S6oLw7VixT.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, BINGYUE PENG, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gojL67CfS8.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48. PMLR, 2016.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025.

Pingyu Wu, Kai Zhu, Yu Liu, Longxiang Tang, Jian Yang, Yansong Peng, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Towards sequence modeling alignment between tokenizer and autoregressive model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GT3obnZ5Fk.

Xiaodong Wu, Wenyi Yu, Chao Zhang, and Phil Woodland. An improved empirical fisher approximation for natural gradient descent. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LmjLRHVCMG.

Dongli Xu, Aleksei Tiulpin, and Matthew B. Blaschko. SoftCFG: Uncertainty-guided stable guidance for visual autoregressive model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=G7tqQ5Upcs.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=AFDcYJKhND. Featured Certification.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=tOXoQPRzPL.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

A Additional Details for the Preliminaries

Fig. 8 summarizes the two views used in our analysis. The Evidence–KL view interprets standard CFG as trading conditional evidence against deviation from the original conditional prediction, while the local Fisher geometry characterizes how this trade-off depends on guidance direction and strength. We provide the corresponding derivations below.

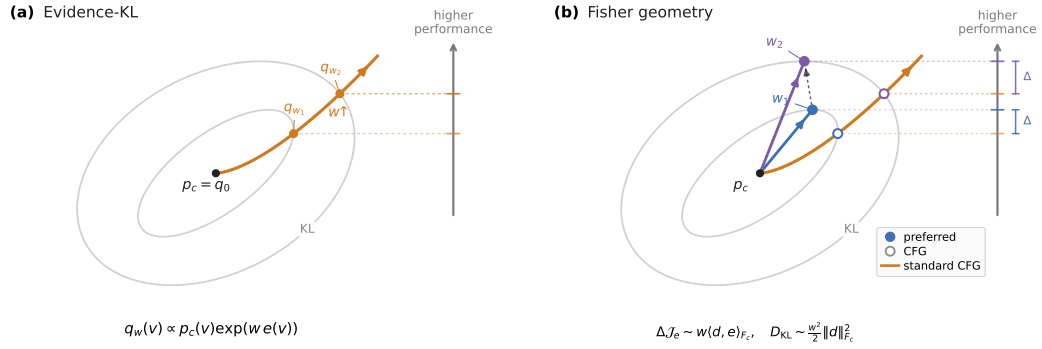


Figure 8: **From the Evidence–KL view to the Fisher geometry of CFG. (a)** Increasing the guidance strength moves the sampling distribution toward higher conditional evidence while increasing its KL deviation from the original conditional distribution p_c. **(b)** Under the Fisher geometry, the evidence gain depends jointly on the applied guidance direction d and its strength w. Standard CFG follows the fixed direction $d = e$.

A.1 Derivation of the Evidence–KL View

Consider a single step of autoregressive decoding. Let

$$p_b(v) = \mathrm{Softmax}(z_b)(v), \qquad v \in \mathcal{V} := \{1, \dots, V\}, \tag{15}$$

be a base categorical distribution, where $p_b(v) > 0$ for every $v \in \mathcal{V}$.

Given a token-level evidence function $E : \mathcal{V} \to \mathbb{R}$, we seek a new sampling distribution q that assigns greater probability to tokens with larger evidence, while remaining close to the base distribution p_b. We formulate this objective as

$$q^* = \arg\max_{q \in \Delta^{V-1}} \left[\sum_{v=1}^{V} q(v)E(v) - \lambda D_{\mathrm{KL}}(q \| p_b) \right], \tag{16}$$

where $\lambda > 0$ controls the penalty for deviating from the base distribution. Expanding the KL divergence gives

$$q^* = \arg\max_{q \in \Delta^{V-1}} \left[\sum_{v=1}^{V} q(v)E(v) - \lambda \sum_{v=1}^{V} q(v) \log \frac{q(v)}{p_b(v)} \right]. \tag{17}$$

To enforce the normalization constraint $\sum_v q(v) = 1$, we introduce a Lagrange multiplier ν and construct

$$\mathcal{L}(q, \nu) = \sum_{v=1}^{V} q(v)E(v) - \lambda \sum_{v=1}^{V} q(v) \log \frac{q(v)}{p_b(v)} + \nu \left(1 - \sum_{v=1}^{V} q(v) \right). \tag{18}$$

Taking the partial derivative with respect to $q(v)$ yields

$$\frac{\partial \mathcal{L}}{\partial q(v)} = E(v) - \lambda \left(\log \frac{q(v)}{p_b(v)} + 1 \right) - \nu. \tag{19}$$

At the optimum, the first-order condition gives

$$E(v) - \lambda \left(\log \frac{q^*(v)}{p_b(v)} + 1 \right) - \nu = 0. \tag{20}$$

Rearranging the terms, we obtain

$$q^*(v) = p_b(v) \exp\left(\frac{E(v)}{\lambda}\right) \exp\left(-1 - \frac{\nu}{\lambda}\right). \tag{21}$$

Normalizing the preceding expression gives

$$q^*(v) = \frac{p_b(v) \exp\left(E(v)/\lambda\right)}{Z_E}, \qquad Z_E := \sum_{j=1}^{V} p_b(j) \exp\left(\frac{E(j)}{\lambda}\right). \tag{22}$$

Thus, the optimal distribution q^* is obtained by exponentially tilting the base distribution according to the token-level evidence E.

The solution in Eq. 22 can be written as

$$q^*(v) \propto \exp\left(z_b(v) + \frac{1}{\lambda} E(v)\right), \qquad p_b(v) = \frac{\exp(z_b(v))}{\sum_j \exp(z_b(j))}. \tag{23}$$

Consequently, the corresponding logit update is

$$z^*(v) = z_b(v) + \frac{1}{\lambda} E(v). \tag{24}$$

Thus, KL-regularized evidence maximization corresponds to exponentially tilting the base distribution by adding a scaled evidence function to its logits.

Let z_c and z_u denote the conditional and unconditional logits respectively. Write the standard CFG relative to the conditional prediction as

$$z_{\text{cfg}} = z_c + w(z_c - z_u), \qquad w \geq 0. \tag{25}$$

This is equivalent to the commonly used form $z_u + s(z_c - z_u)$ under the relation $s = 1 + w$.

In the general formulation, take the conditional prediction as the base and set

$$(p_b, z_b) = (p_c, z_c), \qquad e = z_c - z_u, \qquad \lambda^{-1} = w. \tag{26}$$

and substituting these choices into Eq. 24 gives

$$z^*(v) = z_c(v) + w\big(z_c(v) - z_u(v)\big) = z_{\text{cfg}}(v). \tag{27}$$

Here, we refer to $z_c - z_u$ as the **conditional evidence** signal e, since it captures how the condition changes the relative preference among tokens.

Therefore, standard CFG is an exact instance of KL-regularized evidence maximization with the conditional distribution as the base. Under this view, the guidance strength controls the trade-off between increasing conditional evidence and limiting deviation from the original conditional prediction.

A.2 Fisher Geometry Induced by KL Divergence

The Evidence–KL view measures the distributional deviation induced by guidance through KL divergence. We next characterize its local geometry for categorical next-token predictions. In particular, the second-order expansion of KL divergence gives rise to the Fisher–Rao metric (ichi Amari, 2016; Nielsen, 2020).

Let $z \in \mathbb{R}^V$ be a vector of logits and p_z the corresponding categorical distribution. Using the log-partition function $A(z)$, we write

$$A(z) := \log \sum_{j=1}^{V} \exp(z(j)), \qquad p_z(v) = \exp\big(z(v) - A(z)\big). \tag{28}$$

Consider a local perturbation $\delta \in \mathbb{R}^V$ and let $p = p_z$ and $q = p_{z+\delta}$. For each token v,

$$\log \frac{q(v)}{p(v)} = \delta(v) - A(z + \delta) + A(z). \tag{29}$$

The KL divergence from the perturbed distribution to the original distribution can therefore be written exactly as

$$D_{\mathrm{KL}}(q\|p) = \sum_{v=1}^{V} q(v) \log \frac{q(v)}{p(v)} = q^\top \delta - A(z+\delta) + A(z). \tag{30}$$

The first derivative of the log-partition function is

$$\frac{\partial A(z)}{\partial z(i)} = \frac{\exp(z(i))}{\sum_k \exp(z(k))} = p(i). \tag{31}$$

Differentiating once more gives

$$\begin{aligned} \frac{\partial^2 A(z)}{\partial z(i) \partial z(j)} &= \frac{\partial p(i)}{\partial z(j)} \\ &= \frac{\mathbf{1}\{i=j\} \exp(z(i)) \sum_k \exp(z(k)) - \exp(z(i)) \exp(z(j))}{\left(\sum_k \exp(z(k))\right)^2} \\ &= p(i)\big(\mathbf{1}\{i=j\} - p(j)\big). \end{aligned} \tag{32}$$

Thus, in matrix form,

$$\nabla A(z) = p, \qquad \nabla^2 A(z) = \mathrm{Diag}(p) - pp^\top. \tag{33}$$

We now relate this Hessian to the Fisher information matrix of the same categorical distribution. For an observed token v, the score function is

$$s_z(v) := \nabla_z \log p_z(v) = \nabla_z \big(z(v) - A(z)\big) = e_v - p, \tag{34}$$

where e_v denotes the one-hot vector corresponding to token v. The Fisher information matrix is

$$\begin{aligned} F(p) &:= \mathbb{E}_{v\sim p}\left[s_z(v) s_z(v)^\top\right] \\ &= \mathbb{E}_{v\sim p}\left[(e_v - p)(e_v - p)^\top\right] \\ &= \mathrm{Diag}(p) - pp^\top, \end{aligned} \tag{35}$$

where $\mathbb{E}_{v\sim p}[e_v] = p$ and $\mathbb{E}_{v\sim p}[e_v e_v^\top] = \mathrm{Diag}(p)$. Comparing Eq. 33 and Eq. 35 gives

$$\nabla^2 A(z) = F(p). \tag{36}$$

Thus, for a categorical distribution parameterized by logits, the Hessian of the log-partition function coincides with its Fisher information matrix.

A second-order Taylor expansion around z gives

$$A(z+\delta) = A(z) + p^\top \delta + \frac{1}{2} \delta^\top F(p) \delta + \mathcal{O}\big(\|\delta\|_2^3\big). \tag{37}$$

Since $p_z = \nabla A(z)$ and $\nabla^2 A(z) = F(p)$, a first-order expansion of the probability map gives

$$q = p_{z+\delta} = p + F(p)\delta + \mathcal{O}\big(\|\delta\|_2^2\big). \tag{38}$$

and hence

$$q^\top \delta = p^\top \delta + \delta^\top F(p) \delta + \mathcal{O}\big(\|\delta\|_2^3\big). \tag{39}$$

Substituting Eq. 37 and Eq. 39 into Eq. 30, we obtain

$$\begin{aligned} D_{\mathrm{KL}}(p_{z+\delta}\|p_z) &= p^\top \delta + \delta^\top F(p) \delta - p^\top \delta - \frac{1}{2} \delta^\top F(p) \delta + \mathcal{O}\big(\|\delta\|_2^3\big) \\ &= \frac{1}{2} \delta^\top F(p) \delta + \mathcal{O}\big(\|\delta\|_2^3\big). \end{aligned} \tag{40}$$

Eq. 40 shows that the local KL deviation induced by a small logit perturbation is characterized by the quadratic form associated with $F(p)$. For two directions $a, b \in \mathbb{R}^V$, we define

$$\langle a, b \rangle_{F(p)} = a^\top F(p) b, \qquad \|a\|_{F(p)}^2 = a^\top F(p) a. \tag{41}$$

Under this local geometry, $\frac{1}{2}\|a\|^2_{F(p)}$ approximates the KL deviation induced by a small perturbation a, while $\langle a, b\rangle_{F(p)}$ measures the alignment between two perturbation directions.

We next characterize the corresponding change in conditional evidence. Recall that $e = z_c - z_u$ denotes the conditional evidence vector, and define

$$\mathcal{J}_e(q) = \mathbb{E}_{v\sim q}[e(v)] = q^\top e. \tag{42}$$

Using Eq. 38, we obtain

$$\begin{aligned}\mathcal{J}_e(p_{z+\delta}) - \mathcal{J}_e(p_z) &= \big(p_{z+\delta} - p\big)^\top e \\ &= \delta^\top F(p) e + \mathcal{O}\big(\|\delta\|_2^2\big) \\ &= \langle \delta, e\rangle_{F(p)} + \mathcal{O}\big(\|\delta\|_2^2\big)\,.\end{aligned} \tag{43}$$

Together with Eq. 40, this shows that, locally, the evidence gain depends on the Fisher alignment between the perturbation and the evidence signal, while the KL deviation depends on the Fisher length of the perturbation.

Finally, write the logit perturbation as $\delta = wd$, where $d \in \mathbb{R}^V$ denotes the guidance direction and $w \geq 0$ its strength. Standard CFG corresponds to $d = e = z_c - z_u$. Substituting $\delta = wd$ gives

$$\mathcal{J}_e(p_{z+wd}) - \mathcal{J}_e(p_z) = w\langle d, e\rangle_{F(p)} + \mathcal{O}(w^2), \qquad D_{\mathrm{KL}}(p_{z+wd}\|p_z) = \frac{w^2}{2}\|d\|^2_{F(p)} + \mathcal{O}(w^3). \tag{44}$$

These relations show that the Evidence–KL behavior depends jointly on the guidance direction and its strength. Since $F(p)$ is determined by the current predictive distribution, the corresponding local geometry also changes as the distribution moves during guidance. These two properties provide the basis for our subsequent analyses of how guidance should be composed and applied.

B Additional Analysis of Guidance Composition and Application

This section provides additional analysis results that support the functional decomposition, strength-dependent guidance composition, and successive local updates studied in Sec. 4.

B.1 Additional Analysis of the Functional Decomposition

Why select the midpoint reference. Our decomposition uses the midpoint logits:

$$m = \frac{z_c + z_u}{2}. \tag{45}$$

Let

$$p_c(v) = \frac{\exp z_c(v)}{Z_c}, \qquad p_u(v) = \frac{\exp z_u(v)}{Z_u}. \tag{46}$$

The distribution corresponding to the midpoint satisfies

$$\begin{aligned}p_m(v) &= \frac{\exp\left[\frac{1}{2}(z_c(v) + z_u(v))\right]}{\sum_j \exp\left[\frac{1}{2}(z_c(j) + z_u(j))\right]} \\ &= \frac{\sqrt{p_c(v)p_u(v)}}{\sum_j \sqrt{p_c(j)p_u(j)}}.\end{aligned} \tag{47}$$

Thus, arithmetic averaging in logit space corresponds to a normalized geometric mean in probability space. Unlike a probability space arithmetic average, the geometric mean assigns substantial mass only when a token receives support from both branches. This makes the midpoint a symmetric reference for the probability structure jointly supported by the conditional and unconditional predictions.

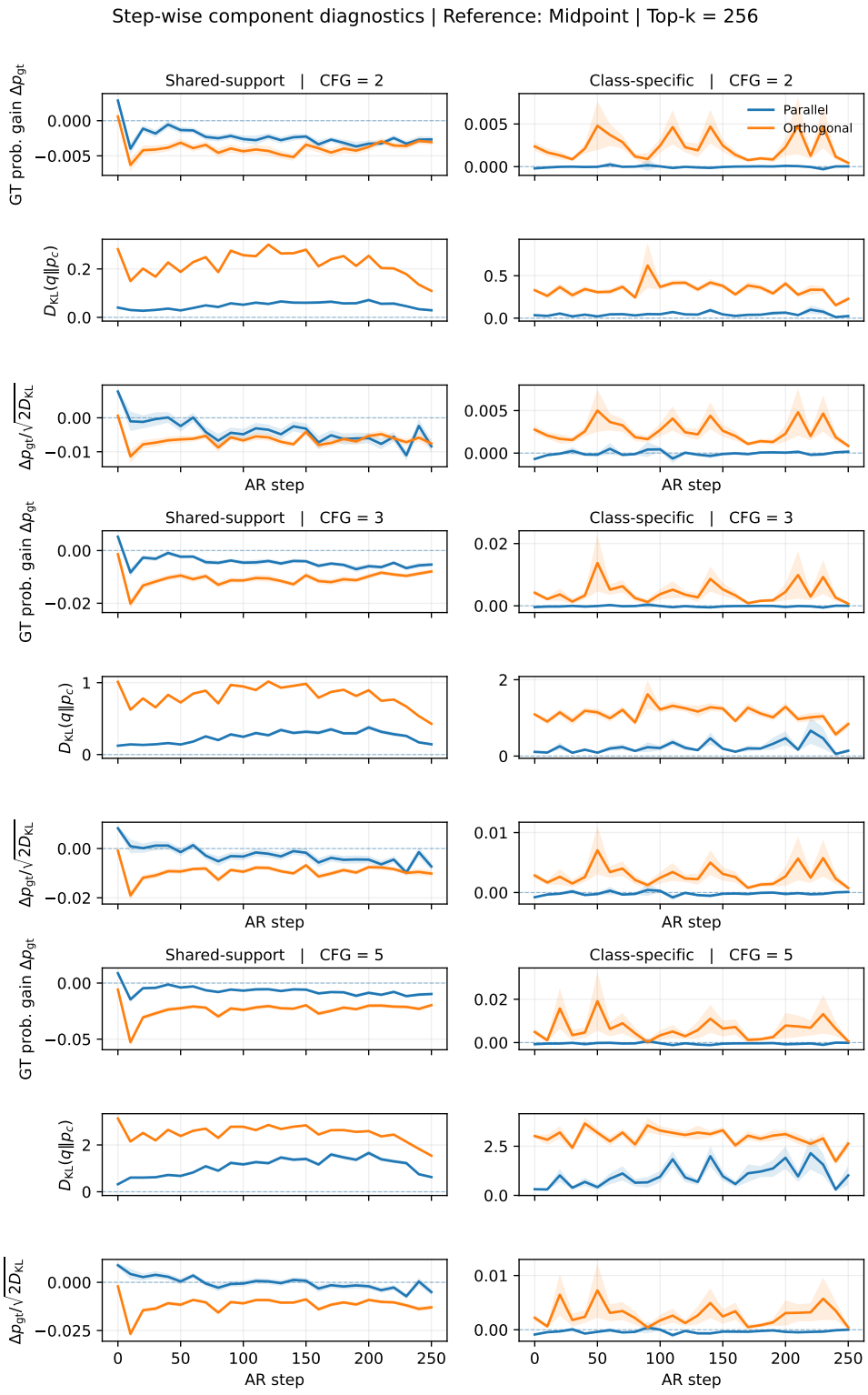


Figure 9: **Step-wise diagnostics with the midpoint reference.** The parallel and orthogonal components consistently favor shared-support preservation and class-specific amplification, respectively.

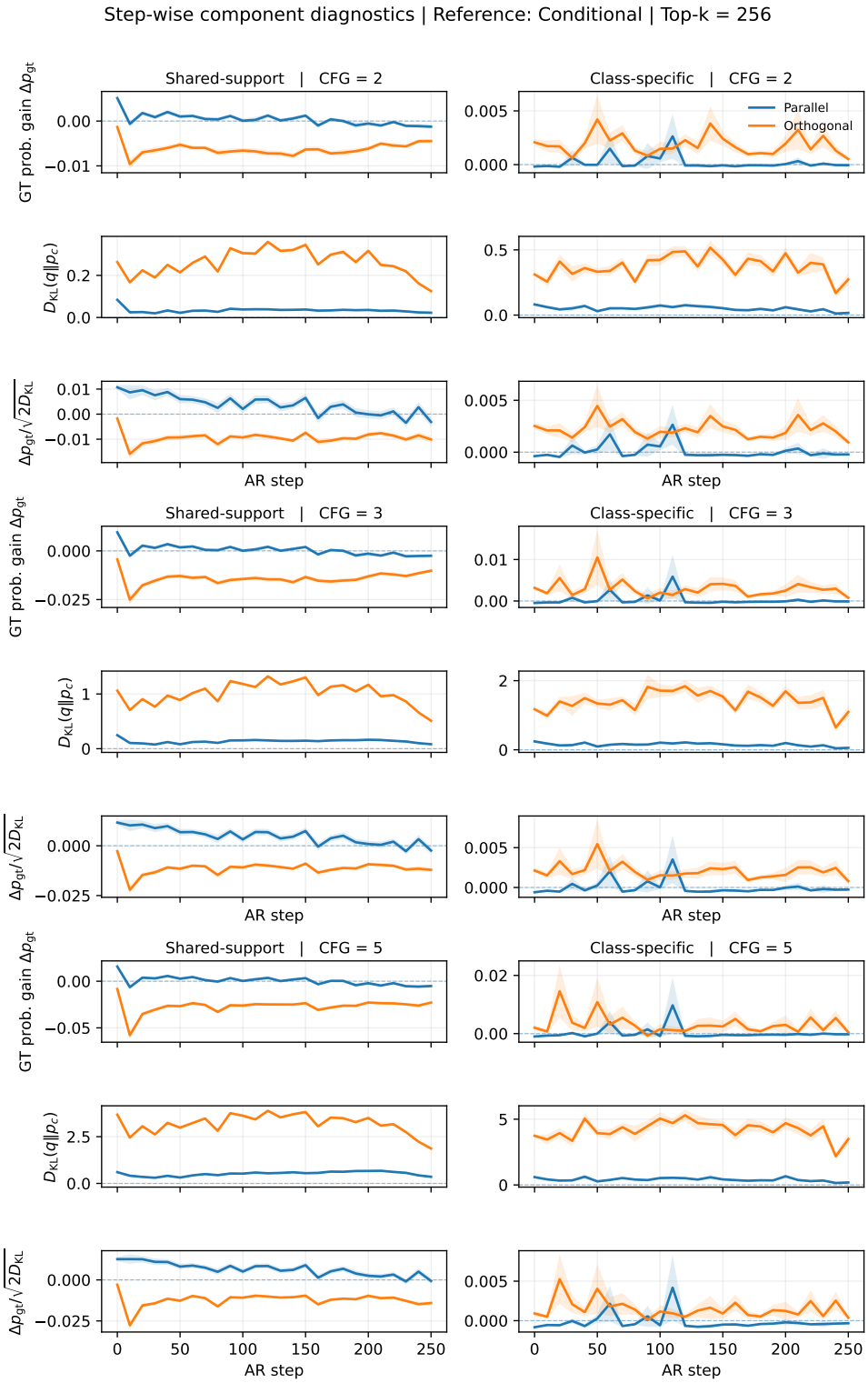


Figure 10: **Step-wise diagnostics with the conditional reference.** The shared-support separation remains strong, while the class-specific separation becomes weaker and less consistent.

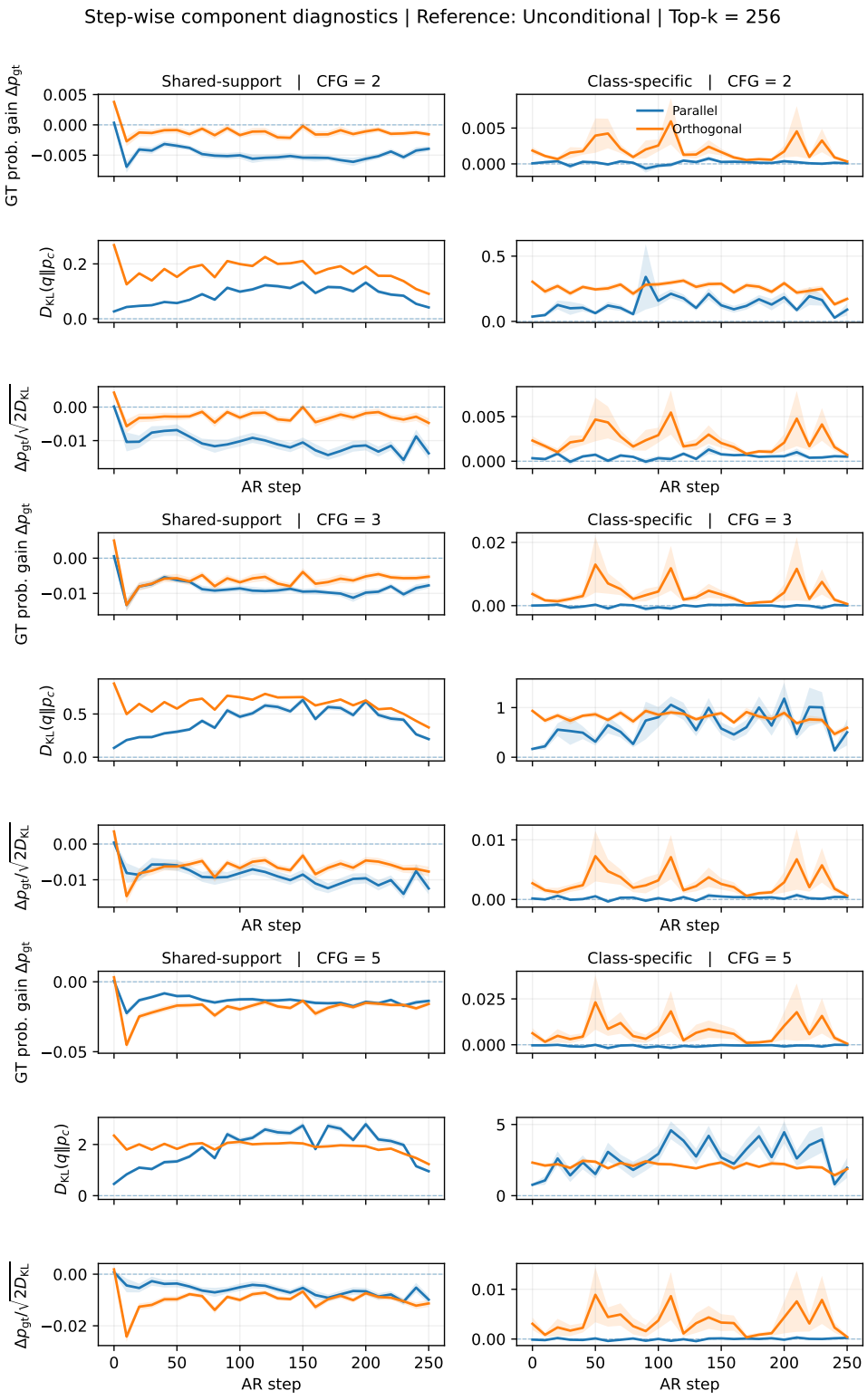


Figure 11: **Step-wise diagnostics with the unconditional reference.** The class-specific separation remains strong, while the expected shared-support separation is not consistently observed.

Reference ablation. To examine whether the observed functional separation depends on the choice of reference, we consider a general reference r and define:

$$e_{\parallel}^{(r)} = \frac{\langle e, r \rangle_{F_c}}{\|r\|_{F_c}^2} r, \qquad e_{\perp}^{(r)} = e - e_{\parallel}^{(r)}. \tag{48}$$

We compare three choices:

$$r \in \left\{ \frac{z_c + z_u}{2}, z_c, z_u \right\}, \tag{49}$$

corresponding to the midpoint, conditional, and unconditional references. We use AliTok-L (Wu et al., 2026) on a balanced subset of 2,048 ImageNet validation images and condition the model on the GT prefix at each evaluated autoregressive position. We probe 26 positions from $t = 0$ to 250 at intervals of 10, and use guidance strengths $w \in \{2, 3, 5\}$. At each position, the same conditional and unconditional logits are used for all three references. The parallel and orthogonal components are then applied separately with the same guidance strength. Following Sec. 4.1, ground-truth tokens are divided into shared-support and class-specific groups using Top-k with $k = 256$.

To summarize the separation between the two functional roles, we define

$$\begin{aligned} G_{\text{shared}}^{(r)} &= \mathbb{E}\left[\Delta p_{\text{gt}}\left(e_{\parallel}^{(r)}\right) - \Delta p_{\text{gt}}\left(e_{\perp}^{(r)}\right) \mid \text{shared-support}\right], \\ G_{\text{class}}^{(r)} &= \mathbb{E}\left[\Delta p_{\text{gt}}\left(e_{\perp}^{(r)}\right) - \Delta p_{\text{gt}}\left(e_{\parallel}^{(r)}\right) \mid \text{class-specific}\right]. \end{aligned} \tag{50}$$

Positive values indicate the expected functional ordering. We further report the fraction of evaluated CFG-strength/AR-position pairs for which this ordering holds as the corresponding consistency.

Table 3: **Reference ablation for the functional decomposition.** G_{shared} measures the advantage of the parallel component on shared-support tokens, while G_{class} measures the advantage of the orthogonal component on class-specific tokens. Consistency is the fraction of 78 evaluated position settings that follow the corresponding functional ordering.

Reference	$G_{\text{shared}} \uparrow$	$G_{\text{class}} \uparrow$	Shared Cons. $\uparrow$	Class Cons. $\uparrow$
Midpoint	0.0082	0.0044	98.7%	98.7%
Conditional	0.0162	0.0025	100.0%	89.7%
Unconditional	-0.0004	0.0046	30.8%	100.0%

Tab. 3 shows that the two endpoint references favor different sides of the decomposition. Using z_c as the reference produces the largest shared-support gap, but weakens the class-specific separation and reduces its consistency to 89.7%. Using z_u shows the opposite behavior: the class-specific role remains strong, while the shared-support gap becomes slightly negative and the expected ordering holds in only 30.8% of the evaluated settings. In contrast, the midpoint maintains positive role gaps and nearly perfect consistency for both token groups. The midpoint therefore provides a balanced reference that exposes both shared-support preservation and class-specific amplification consistently. Together with its symmetric geometric mean interpretation, this balanced behavior motivates our use of the midpoint as the default reference throughout the main analysis.

Step-wise Functional Diagnostics across References. Figs. 9–11 show the corresponding diagnostics over autoregressive positions. For each reference, we report CFG strengths 2, 3, and 5. Within each CFG block, the columns correspond to shared-support and class-specific tokens, while the rows report ground-truth probability change, KL deviation from p_c, and probability change normalized by $\sqrt{2D_{\text{KL}}}$. The midpoint preserves the expected ordering across nearly all positions and guidance strengths. The conditional reference shows a stronger bias toward the shared-support role, whereas the unconditional reference shows the complementary bias toward the class-specific role, consistent with Tab. 3.

Robustness to the Top-k threshold. Using the midpoint reference, our default analysis sets $k = 256$ to distinguish shared-support and class-specific ground-truth tokens. To examine whether the observed functional separation depends on this threshold, we repeat the same component-wise diagnostic with

$$k \in \{128, 256, 512\}. \tag{51}$$

As shown in Fig. 12, changing k alters the exact composition of the two token groups, but the qualitative roles of the two components remain stable across thresholds. The orthogonal component consistently provides stronger amplification for class-specific tokens, while the parallel component better preserves shared-support tokens.

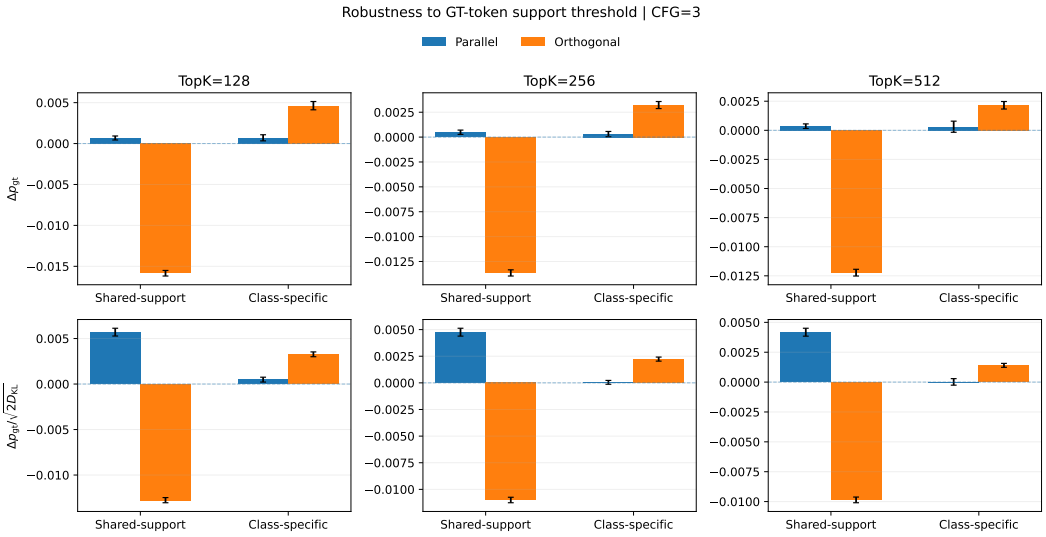


Figure 12: **Robustness to the Top-k threshold used for ground-truth token categorization.** We repeat the component-wise analysis for $k \in \{128, 256, 512\}$. Top: ground-truth token probability change. Bottom: corresponding probability change normalized by KL displacement.

B.2 Additional Analysis of Strength-Dependent Composition

Sec. 4.2 shows that the preferred balance between the orthogonal and parallel components changes with guidance strength. We provide additional analysis of this interaction and report the complete sweeps underlying the observations in main text.

Interpreting guidance composition. Using the parameterization in Eq. 12, we write

$$z_{a,\lambda} = z_c + a\left(e_\perp + \lambda e_\parallel\right), \qquad \lambda = \frac{b}{a}. \tag{52}$$

The guidance strength a controls the magnitude of the orthogonal contribution, while λ controls the balance between the two components. Since the decomposition in Sec. 4.1 satisfies

$$\left\langle e_\perp, e_\parallel \right\rangle_{F_c} = 0, \tag{53}$$

the local KL deviation around p_c takes the form

$$D_{\mathrm{KL}}\left(q_{a,\lambda} \,\|\, p_c\right) = \frac{a^2}{2}\left(\|e_\perp\|_{F_c}^2 + \lambda^2 \|e_\parallel\|_{F_c}^2\right) + \mathcal{O}(a^3), \tag{54}$$

The corresponding guided distribution can be written exactly as

$$q_{a,\lambda}(v) \propto p_c(v) \exp\left\{a\left[e_\perp(v) + \lambda e_\parallel(v)\right]\right\}. \tag{55}$$

For a class-specific token x_c and a shared-support token x_s, their relative odds therefore satisfy

$$\log \frac{q_{a,\lambda}(x_c)/q_{a,\lambda}(x_s)}{p_c(x_c)/p_c(x_s)} = a\left[e_\perp(x_c) - e_\perp(x_s) + \lambda\left(e_\parallel(x_c) - e_\parallel(x_s)\right)\right]. \tag{56}$$

Eq. 54 shows that, for a fixed composition, the local deviation from p_c grows approximately with a^2, while Eq. 56 shows that the relative preference induced by the same composition is amplified proportionally with a. As observed in Sec. 4.1, this preference reflects a trade-off between the two

components: stronger orthogonal guidance tends to favor class-specific predictions while weakening shared-support predictions, whereas the parallel component shows the complementary tendency.

When a is small, both the overall deviation and this relative preference remain limited, so greater emphasis on the orthogonal component can be beneficial. As a increases, however, the deviation grows rapidly and the same preference between the two token groups is amplified more strongly, making the cost in shared-support predictions increasingly consequential. Increasing λ gives greater relative weight to the parallel component, helping compensate for this loss and rebalance the two functional effects under stronger guidance.

This trade-off is particularly relevant to autoregressive generation, where each sampled token becomes part of the prefix for subsequent predictions. As the deviation of this induced guidance grows, an overly skewed token preference can therefore affect not only the current prediction but also later decoding steps. This provides an intuitive explanation for why the preferred composition can shift toward a larger parallel contribution as guidance strength becomes stronger.

Broader composition sweep. We report the broader sweep underlying the results in Sec. 4.2. For each guidance strength a, we vary the composition ratio λ over the restricted range $\lambda \in [0, 1]$. Fig. 13 shows the same strength-dependent shift across the full range of evaluated settings. At stronger guidance strengths, the optimum reaches $\lambda = 1$ and remains at the boundary of the restricted search range.

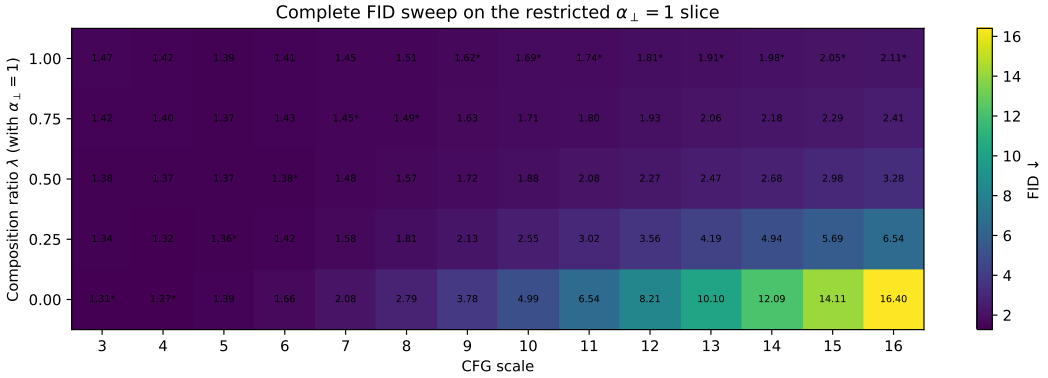


Figure 13: **Broader sweep over guidance strength and composition.** Columns correspond to guidance strengths a and rows correspond to composition ratios λ. Each cell reports FID on AliTok-L, and the minimum within each guidance strength is marked by a star.

Composition beyond $\lambda = 1$. We further evaluate compositions with $\lambda > 1$. For each fixed orthogonal strength a, we vary the parallel strength b and report the corresponding ratio $\lambda = b/a$. This covers compositions from stronger orthogonal emphasis to stronger parallel emphasis, with $\lambda = 1$ corresponding to standard CFG. The fixed a sweeps extend the same trend beyond $\lambda = 1$. As Fig. 14 shows, the preferred composition shifts from $\lambda^\star = 0$ at $a = 4$, to $\lambda^\star = 1$ at $a = 7.5$, and further to $\lambda^\star = 1.33$ at $a = 12$. Thus, as guidance becomes stronger, the preferred composition continues to shift toward a larger parallel contribution and can move beyond the standard CFG composition.

B.3 Local Geometry Along Successive Guidance Updates

Evolution of the local guidance direction. The Fisher decomposition in Sec. 4.1 is defined locally at the current predictive distribution. During successive updates, although the underlying evidence e and reference m are unchanged, the predictive distribution evolves, causing the Fisher geometry and the resulting decomposition to change accordingly. Let $e_\perp^{(j)}$ and $e_\parallel^{(j)}$ denote the decomposition of the same evidence e at the j-th intermediate distribution. Since

$$e = e_\perp^{(j)} + e_\parallel^{(j)} \tag{57}$$

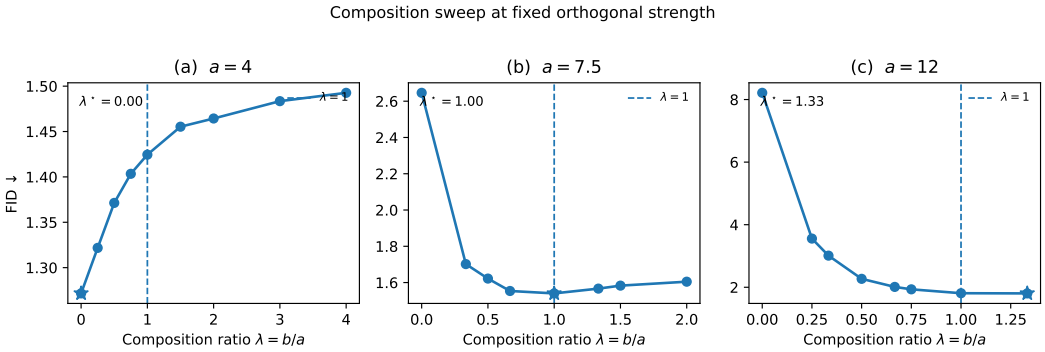


Figure 14: **Composition sweeps at fixed orthogonal strength.** Each panel fixes a and varies the composition ratio $\lambda = b/a$ on AliTok-L. The dashed vertical line denotes the standard CFG composition $\lambda = 1$, and the star marks the minimum FID. **(a)** $a = 4$. **(b)** $a = 7.5$. **(c)** $a = 12$.

holds at every intermediate state, comparison with the initial decomposition gives

$$e_{\perp}^{(j)} = e_{\perp}^{(0)} + \left(e_{\parallel}^{(0)} - e_{\parallel}^{(j)}\right). \tag{58}$$

Eq. 58 shows that the two components adjust together while preserving the same evidence e: a change in the parallel component is balanced by a corresponding change in the orthogonal component. For a fixed composition λ, define the local guidance direction as

$$d^{(j)} = e_{\perp}^{(j)} + \lambda e_{\parallel}^{(j)}. \tag{59}$$

Using Eq. 58, we obtain

$$d^{(j)} = d^{(0)} + (1-\lambda)\left(e_{\parallel}^{(0)} - e_{\parallel}^{(j)}\right). \tag{60}$$

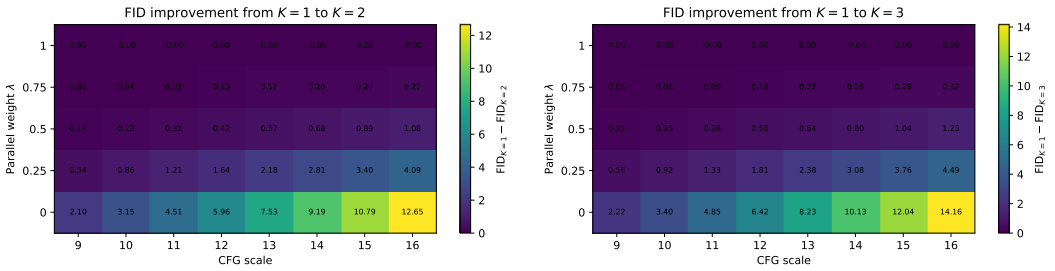
Eq. 60 shows that the change in the guidance direction introduced by refreshing the decomposition is scaled by $1-\lambda$. As λ approaches 1, this change becomes smaller and vanishes at $\lambda = 1$. Beyond this point, although the correction reverses sign for $\lambda > 1$, our additional checks show only small FID differences across different values of K. Overall, the strong effect of successive updates is mainly observed when the orthogonal component receives greater emphasis.

Interaction with guidance strength. The role of guidance strength can be understood from how much the prediction changes along the update. When the displacement is small, the intermediate distributions remain close to the starting prediction, so the local Fisher geometry and the resulting decomposition change little:

$$e_{\parallel}^{(j)} \approx e_{\parallel}^{(0)}. \tag{61}$$

As the guidance strength increases, the intermediate predictions can move farther from the starting state, allowing the local decomposition to vary more along the update. This provides a geometric interpretation of the larger effect of successive updates observed at stronger guidance strengths in Sec. 4.3.

Interaction across guidance strength and composition. We further examine successive updates across the evaluated combinations of guidance strength and composition. We fix the orthogonal weight to 1 and vary $a \in \{9, \ldots, 16\}$ and $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. For each setting, we compare $K = 2$ and $K = 3$ with the one-shot update $K = 1$. Fig. 15 shows the same trend across the evaluated grid: when the orthogonal component receives more weight, the benefit of successive updates grows clearly with guidance strength. As λ approaches 1, this benefit becomes progressively smaller across all strengths. At $\lambda = 1$, the differences are nearly zero for both $K = 2$ and $K = 3$, consistent with Eq. 60.



(a) $K = 2$ compared with $K = 1$.

(b) $K = 3$ compared with $K = 1$.

Figure 15: **Interaction between guidance strength, composition, and successive local updates on AliTok-L.** Each cell reports the FID improvement over $K = 1$ for a given guidance strength a and composition ratio λ. Positive values indicate better FID with successive local updates.

C Searching over Guidance Strength, Composition, and Local Updates

The experiments in Sec. 4 consider three factors of classifier-free guidance: the guidance strength a, the composition ratio λ, and the number of successive local updates K. Given discrete candidate values for these factors, the most straightforward approach is to evaluate all possible combinations. This remains a meaningful baseline even if the three factors are independent, since it identifies the configuration within the chosen ranges and provides a direct better reference.

Our experiments further show that these factors exhibit dependencies rather than varying independently. This suggests that the same configuration space can be searched more efficiently once these relationships are taken into account. We first formalize the discrete search space over (a, λ, K), and then show how the two observations can be used to eliminate unnecessary evaluations.

C.1 Systematic Search Using the Guidance Relationships

Let the candidate values for guidance strength, composition ratio, and successive local updates be

$$\begin{aligned} \mathcal{A} &= \{a_1, \dots, a_i\}, \qquad a_1 < a_2 < \cdots < a_i, \\ \Lambda &= \{\lambda_1, \dots, \lambda_j\}, \qquad \lambda_1 < \lambda_2 < \cdots < \lambda_j, \\ \mathcal{K} &= \{K_1, \dots, K_m\}, \quad K_1 = 1 < K_2 < \cdots < K_m. \end{aligned} \tag{62}$$

The direct Cartesian search therefore contains $i \times j \times m$ configurations. For a configuration (a_s, λ_t, K_u), where $1 \le s \le i$, $1 \le t \le j$, and $1 \le u \le m$, let

$$F_{stu} := F(a_s, \lambda_t, K_u) \tag{63}$$

denote its evaluation objective, where lower values are preferred. Since $K_1 = 1$, F_{st1} corresponds to the one-shot setting. For each guidance strength a_s, define

$$r_s := \arg \min_{1 \le t \le j} F_{st1}, \tag{64}$$

where r_s is the index of the preferred composition ratio under the one-shot setting. The observed shift toward larger λ as guidance becomes stronger corresponds to

$$r_1 \le r_2 \le \cdots \le r_i. \tag{65}$$

Thus, once r_{s-1} has been identified, the preferred composition at a_s should satisfy $r_s \ge r_{s-1}$. When locating r_s, compositions with $t < r_{s-1}$ can therefore be skipped, and only $t = r_{s-1}, \dots, j$ need to be evaluated.

For the remaining local-update settings, define the improvement over the corresponding one-shot configuration as

$$G_{stu} := F_{st1} - F_{stu}, \qquad u = 2, \dots, m. \tag{66}$$

At a fixed guidance strength, this improvement decreases as the composition ratio increases. Accordingly,

$$t < q \quad \Longrightarrow \quad G_{stu} \geq G_{squ}, \qquad u = 2, \ldots, m. \tag{67}$$

Once r_s has been identified, consider any composition index $t > r_s$. By the definition of r_s and the ordering above,

$$F_{sr_s1} \leq F_{st1}, \qquad G_{sr_su} \geq G_{stu}. \tag{68}$$

It follows that

$$\begin{aligned} F_{sr_su} &= F_{sr_s1} - G_{sr_su} \\ &\leq F_{st1} - G_{stu} = F_{stu}, \qquad t > r_s, \;\; u = 2, \ldots, m. \end{aligned} \tag{69}$$

Therefore, for each $u = 2, \ldots, m$, configurations with $t > r_s$ are dominated by $(a_s, \lambda_{r_s}, K_u)$ and do not need to be evaluated. The search at a_s can thus be restricted to $t = 1, \ldots, r_s$ for all remaining local-update settings.

Together, the two orderings define the systematic search procedure. At each guidance strength a_s, the one-shot setting evaluates only $t = r_{s-1}, \ldots, j$ to identify the new optimum r_s. Once r_s is found, each remaining local-update setting evaluates only $t = 1, \ldots, r_s$. Repeating this procedure over $a_1, \ldots, a_i$ gives Algorithm 1.

Algorithm 1 Systematic Search over Guidance Configurations

Require: $\mathcal{A} = \{a_1, \ldots, a_i\}, \Lambda = \{\lambda_1, \ldots, \lambda_j\}, \mathcal{K} = \{K_1, \ldots, K_m\}$, with $K_1 = 1$

▷ Initialize at the smallest guidance strength

for $t = 1, \ldots, j$ **do**
  Evaluate F_{1t1}
end for
$r_1 \leftarrow \arg\min_{t \in \{1, \ldots, j\}} F_{1t1}$
for $u = 2, \ldots, m$ **do**
  for $t = 1, \ldots, r_1$ **do**
    Evaluate F_{1tu}
  end for
end for

▷ Search the remaining guidance strengths

for $s = 2, \ldots, i$ **do**
  for $t = r_{s-1}, \ldots, j$ **do**
    Evaluate F_{st1}
  end for
  $r_s \leftarrow \arg\min_{t \in \{r_{s-1}, \ldots, j\}} F_{st1}$
  for $u = 2, \ldots, m$ **do**
    for $t = 1, \ldots, r_s$ **do**
      Evaluate F_{stu}
    end for
  end for
end for
return the evaluated configuration with the lowest objective

Search cost. We quantify the number of configurations evaluated by Algorithm 1. For the one-shot setting, the first guidance strength evaluates all j composition ratios, while each subsequent a_s evaluates $j - r_{s-1} + 1$ candidates. Therefore,

$$\begin{aligned} N_{\text{one}} &= j + \sum_{s=2}^{i} (j - r_{s-1} + 1) \\ &= ij + i - 1 - \sum_{s=1}^{i-1} r_s. \end{aligned} \tag{70}$$

For the remaining $m-1$ local-update settings, r_s compositions are evaluated at each a_s, giving

$$N_{\text{local}} = (m-1)\sum_{s=1}^{i} r_s. \tag{71}$$

Hence, the total search cost is

$$N_{\text{search}} = ij + i - 1 + (m-2)\sum_{s=1}^{i-1} r_s + (m-1)r_i. \tag{72}$$

Since $1 \le r_1 \le \cdots \le r_i \le j$, for $m \ge 2$ we obtain

$$i(j+m-1) \le N_{\text{search}} \le ijm - (i-1)(j-1). \tag{73}$$

Compared with the ijm evaluations required by exhaustive Cartesian search, this search strategy skips between $(i-1)(j-1)$ and $i(m-1)(j-1)$ configurations.

Example on AliTok-L. Consider the $15 \times 7 \times 2$ search space used for AliTok-L, which contains 210 configurations under exhaustive Cartesian search. Based on the bounds derived above, the systematic search requires only 120–126 evaluations. In practice, applying the procedure to AliTok-L evaluates 125 configurations, which falls within this range and means that the number of evaluated configurations is reduced by 40.5% compared with exhaustive search.

C.2 Greedy Search from Standard CFG

The systematic search above reduces the configuration space while retaining the optimum under the two ordering relations. When a smaller evaluation budget is preferred, we further consider a greedy alternative that starts from a tuned standard CFG configuration and follows a short search path over the same candidate sets.

We start from a tuned standard CFG configuration $(a_{s_0}, \lambda_{t_0}, K_1)$ with $\lambda_{t_0} = 1$ and $K_1 = 1$. We first greedily refine guidance strength and composition, following the tendency for weaker guidance to favor smaller λ. After selecting $(a^\star, \lambda^\star)$, we evaluate the remaining local-update settings $K_2, \ldots, K_m$ and choose the best configuration. The complete procedure is summarized in Algorithm 2. Compared with the systematic search above, the greedy procedure follows a substantially shorter search path, allowing strong configurations to be identified with far fewer evaluations.

Example on AliTok. Tab. 4 illustrates the greedy path on AliTok-L. Starting from the tuned standard CFG configuration $(a, \lambda, K) = (5, 1, 1)$, the search progressively decreases the composition ratio and then the guidance strength. It reaches $(4, 0, 1)$ with an FID of 1.27, after which $K = 2$ does not provide a further improvement. Overall, the procedure reaches a strong configuration after only seven evaluations, improving FID from 1.39 to 1.27.

Table 4: Greedy search path on AliTok-L.

Step	a	λ	K	FID ↓
1	5	1.00	1	1.39
2	5	0.75	1	1.37
3	5	0.50	1	1.37
4	4	0.50	1	1.37
5	4	0.25	1	1.32
6	4	0.00	1	**1.27**
7	4	0.00	2	1.29

As shown in Tab. 5, the preferred number of local updates can differ across backbones. While AliTok-L favors $K = 1$, AliTok-XL further improves from 1.28 to 1.26 FID when increasing K from 1 to 2 at the selected $(w, \lambda) = (7.0, 0.5)$, whereas a larger $K = 3$ brings no further improvement.

Algorithm 2 Greedy Search from Standard CFG

Require: $\mathcal{A} = \{a_1, \ldots, a_i\}$, $\Lambda = \{\lambda_1, \ldots, \lambda_j\}$, $\mathcal{K} = \{K_1, \ldots, K_m\}$, starting indices (s_0, t_0) with $\lambda_{t_0} = 1$ and $K_1 = 1$

$(s^\star, t^\star) \leftarrow (s_0, t_0)$, $F^\star \leftarrow F_{s_0 t_0 1}$, $t_{\text{start}} \leftarrow t_0$
found $\leftarrow$ False
▷ Stage 1: refine guidance strength and composition
for $s = s_0, s_0 - 1, \ldots, 1$ **do**
    $t \leftarrow t_{\text{start}}$, $F \leftarrow F_{st1}$ ▷ Evaluate if not cached
    while $t > 1$ **do**
        $q \leftarrow t - 1$; evaluate F_{sq1}
        $t_{\text{start}} \leftarrow q$
        if $F_{sq1} < F$ **then**
            $(t, F) \leftarrow (q, F_{sq1})$
        else
            break
        end if
    end while
    if $F < F^\star$ **then**
        $(s^\star, t^\star, F^\star) \leftarrow (s, t, F)$
        found $\leftarrow$ True
    else if found **then**
        break
    end if
    if $t = 1$ **then**
        break
    end if
end for
▷ Stage 2: refine local updates
for $u = 2, \ldots, m$ **do**
    Evaluate $F_{s^\star t^\star u}$
end for
$u^\star \leftarrow \arg\min_{u \in \{1, \ldots, m\}} F_{s^\star t^\star u}$
return $(a_{s^\star}, \lambda_{t^\star}, K_{u^\star})$

Table 5: Greedy search path on AliTok-XL. After identifying a favorable guidance strength and composition, increasing the number of local updates from $K = 1$ to $K = 2$ further improves FID.

Step	a	λ	K	FID ↓
1	8.0	1.00	1	1.32
2	8.0	0.75	1	1.30
3	8.0	0.50	1	1.31
4	7.0	0.50	1	1.28
5	7.0	0.25	1	1.29
6	6.0	0.25	1	1.30
7	6.0	0.00	1	1.31
8	7.0	0.50	2	**1.26**
9	7.0	0.50	3	1.27

D Additional Experimental Details and Results

D.1 Implementation and Evaluation Details

All experiments are built on the released checkpoints and sampling code of the corresponding backbones, without retraining or changes to the model architecture or tokenizer. We preserve the standard CFG implementation of each backbone and introduce our method only after obtaining the

conditional and reference logits. The detailed inference procedure at one autoregressive step is summarized in Algorithm 3.

Algorithm 3 Logit-space guidance at one autoregressive step

Require: Conditional logits z_c, reference logits z_u, guidance strength a, composition ratio λ, local steps K

$e \leftarrow z_c - z_u$
$m \leftarrow (z_c + z_u)/2$
$z^{(0)} \leftarrow z_c$
for $j = 0, \ldots, K-1$ **do**
  $q^{(j)} \leftarrow \mathrm{Softmax}(z^{(j)})$
  $F^{(j)} \leftarrow \mathrm{Diag}(q^{(j)}) - q^{(j)} q^{(j)\top}$
  $e_{\parallel}^{(j)} \leftarrow \dfrac{\langle e, m \rangle_{F^{(j)}}}{\|m\|_{F^{(j)}}^2} m$
  $e_{\perp}^{(j)} \leftarrow e - e_{\parallel}^{(j)}$
  $z^{(j+1)} \leftarrow z^{(j)} + \dfrac{a}{K} \left(e_{\perp}^{(j)} + \lambda e_{\parallel}^{(j)} \right)$
end for
Sample the next token from $\mathrm{Softmax}(z^{(K)})$

Table 6 summarizes the sampling settings used for each backbone. LlamaGen and SimpleAR use constant CFG, whereas RAR and AliTok retain their original power cosine schedules. In our experiments, we vary only the guidance strength, composition ratio λ, and number of successive local updates K; all other sampling parameters remain unchanged, and the original CFG schedule is preserved for each backbone. For ImageNet evaluation, FID, IS, precision, and recall are computed from $50{,}000$ generated samples distributed evenly across the $1{,}000$ classes, and all experiments are run on eight NVIDIA A100 GPUs.

Table 6: **Sampling settings.** The listed parameters follow the released configuration of each backbone and remain fixed unless they correspond to the guidance quantities explicitly studied in our experiments.

Backbone	Task	Resolution	Temperature	CFG schedule	Schedule power
LlamaGen-L	C2I	256×256	1.00	Constant	–
RAR-XL	C2I	256×256	1.02	Power cosine	1.5
RAR-XXL	C2I	256×256	1.02	Power cosine	1.2
AliTok-L	C2I	256×256	1.00	Power cosine	0.6
AliTok-XL	C2I	256×256	1.00	Power cosine	1.4
SimpleAR-0.5B-RL	T2I	1024×1024	1.00	Constant	–

For the overall performance comparison, we include SoftCFG (Xu et al., 2026), a recent inference method developed specifically for CFG in next-token autoregressive image generation. Like our approach, SoftCFG operates at inference time without retraining the backbone, making it a closely related baseline for evaluating improvements over standard CFG. We reproduce SoftCFG from its released implementation, retaining its uncertainty-guided perturbation and Step Normalization. When applying it to additional backbones, we adapt the same inference procedure while keeping the model weights, tokenizer, and original sampling pipeline unchanged.

For each backbone, setting $\lambda = 1$ and $K = 1$ reduces our implementation to standard CFG at the corresponding guidance strength. We use this setting to reproduce the standard CFG results for each backbone. The reproduced FID values are close to the originally reported results, with differences within approximately performance. Importantly, all results in our comparisons, including standard CFG, SoftCFG, and our method, are obtained from our own runs **under the same evaluation pipeline rather than copied from previously reported numbers**.

There is also a minor difference in how guidance strength is parameterized across the evaluated backbones. Some released implementations write standard CFG as $z_u + s(z_c - z_u)$, whereas our

formulation measures the guidance displacement from the conditional prediction z_c. Under this parameterization, the corresponding guidance strength in our formulation is $a = s - 1$.

D.2 Additional Results for Guidance Mechanisms

To supplement the representative results in Sec. 5.2, we report broader parameter sweeps over the evaluated guidance settings. The composition analysis varies guidance strength a and composition ratio λ with $K = 1$, while the local-update analysis further varies K under different strengths and compositions. All remaining sampling settings follow Appendix D.1.

Guidance composition sweeps. Fig. 16 reports the composition sweeps on LlamaGen and SimpleAR across the evaluated guidance strengths. We fix the orthogonal weight to 1 and $K = 1$, and vary the composition ratio λ at each guidance strength a. Color indicates the gap to the best evaluated composition at the same guidance strength, using FID for LlamaGen and the GenEval overall score for SimpleAR. Across both models, the preferred composition shifts toward larger λ as guidance becomes stronger.

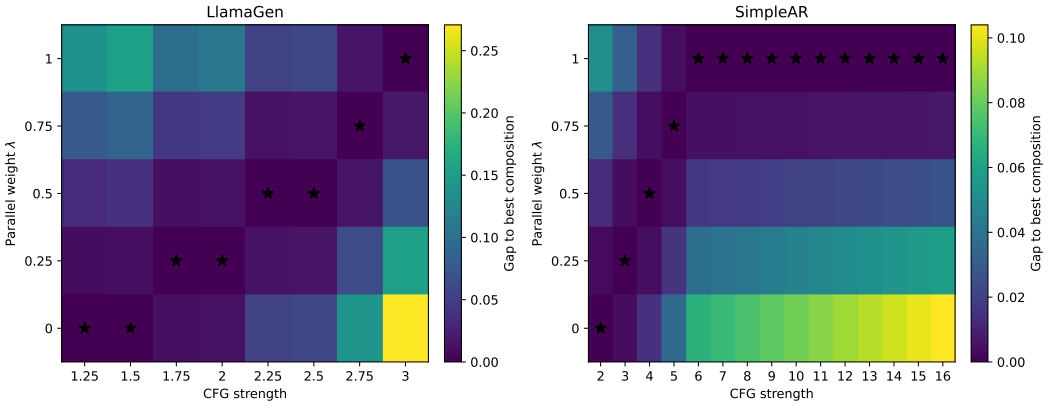


Figure 16: **Guidance composition sweeps on LlamaGen and SimpleAR.** The orthogonal weight is fixed to 1 and $K = 1$. Columns correspond to guidance strengths a, and rows correspond to composition ratios λ. Color indicates the gap to the best evaluated composition at each strength, defined by the lowest FID for LlamaGen and the highest GenEval overall score for SimpleAR. Stars mark the best evaluated composition in each column.

Successive local update sweeps. Fig. 17 reports the effect of $K = 2$ and $K = 3$ relative to $K = 1$ across the evaluated guidance strengths a and composition ratios λ. On both LlamaGen and SimpleAR, the gain increases with guidance strength when λ is small and decreases as λ grows. The same trend appears for both $K = 2$ and $K = 3$, with a stronger effect for $K = 3$.

D.3 Additional Analysis for Class-Conditional Generation

We further characterize the class-conditional results by considering three practical aspects: the number of evaluations needed to find strong guidance settings, the trade-off between FID and IS, and the stability of the gains across random seeds. The corresponding search procedures are detailed in Appendix C, while the results below compare their practical performance, examine the relation between FID and IS, and test whether the gains remain consistent across random seeds.

Search results and number of evaluations. We compare the search results on LlamaGen-L and AliTok-L, using the same candidate ranges considered in our experiments. For LlamaGen-L, the standard CFG sweep suggests an effective guidance range of $a \in \{1.0, 1.25, \ldots, 3.5\}$, giving 11 candidate strengths. We further use $\lambda \in \{0, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$ and $K \in \{1, 2\}$, resulting in $11 \times 7 \times 2 = 154$ possible combinations. Standard CFG and SoftCFG only tune the guidance strength and therefore require 11 evaluations. Following the systematic search in Appendix C, the

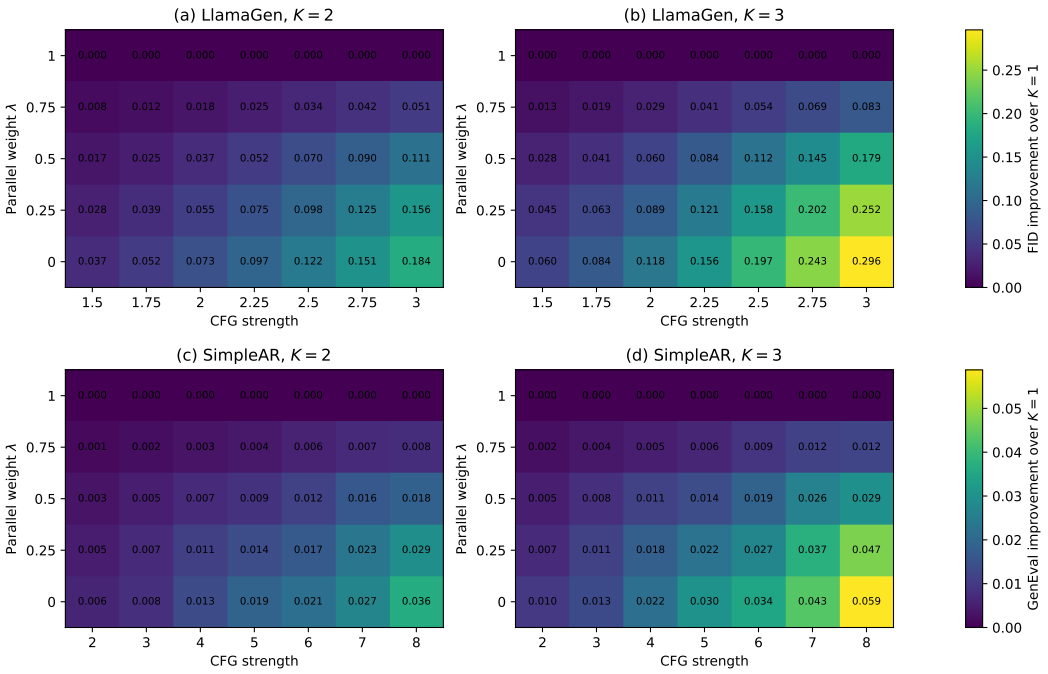


Figure 17: **Successive local updates across guidance settings on LlamaGen and SimpleAR.** Each heatmap shows the improvement over $K = 1$ across guidance strengths a and composition ratios λ, with the orthogonal weight fixed to 1. The first two panels show $K = 2$ and $K = 3$ on LlamaGen, and the last two show $K = 2$ and $K = 3$ on SimpleAR.

number of evaluations is bounded between 88 and 94, and 93 configurations are evaluated in practice. The greedy search starts from the best standard CFG setting and uses six additional evaluations, giving 17 evaluations in total.

For AliTok-L, we similarly use 15 guidance strengths $a \in \{2, 3, \ldots, 16\}$, together with the same seven values of λ and two values of K, giving $15 \times 7 \times 2 = 210$ possible combinations. Standard CFG and SoftCFG require 15 evaluations, the systematic search evaluates 125 configurations within its theoretical range of 120–126, and the greedy search requires only seven additional evaluations after the standard CFG sweep, for 21 evaluations in total.

Tab. 7 summarizes the corresponding FID and evaluation counts. Compared with the full Cartesian search, the systematic procedure substantially reduces the number of required evaluations. The greedy procedure further lowers the search effort and still retains clear gains over both standard CFG and SoftCFG.

Table 7: FID and number of evaluations for different guidance methods.

Method	LlamaGen-L FID ↓	LlamaGen-L Eval. ↓	AliTok-L FID ↓	AliTok-L Eval. ↓
Standard CFG	3.72	11	1.39	15
SoftCFG	3.53	11	1.37	15
Ours(systematic search)	**3.25**	93	**1.27**	125
Ours(greedy search)	3.36	17	**1.27**	21

Trade-off between FID and Inception Score. We compare standard CFG and our method across several representative settings on LlamaGen-L and AliTok-L. Fig. 18 shows that, at similar IS levels, our method generally achieves lower FID than standard CFG. This indicates that the FID gains reported in Tab. 1 are not simply obtained by sacrificing IS.

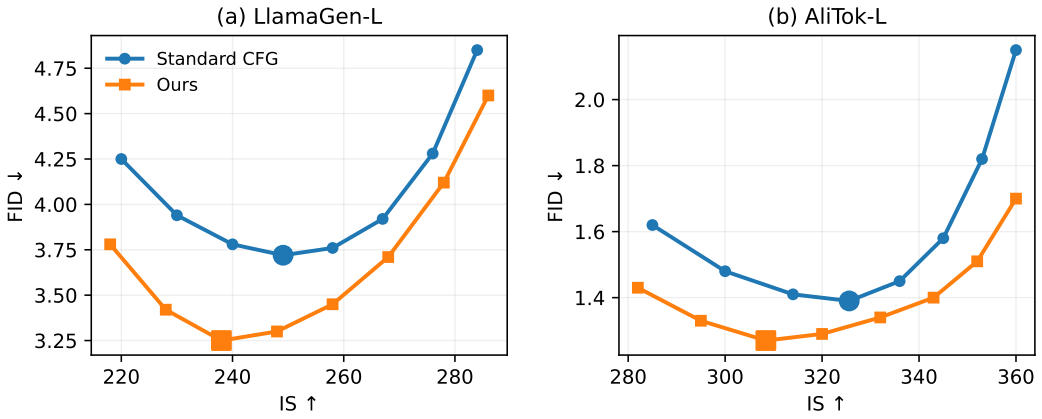


Figure 18: FID and IS across representative settings on (a) LlamaGen-L and (b) AliTok-L. Lower FID and higher IS are preferred.

Variation across random seeds. We further evaluate the best settings of LlamaGen-L and AliTok-L using two additional random seeds, while keeping all guidance settings fixed without further tuning. As shown in Table 8, the FID remains stable across the three seeds on both backbones. In particular, the variation across seeds is small relative to the improvements reported over standard CFG in Table 1, indicating that the observed gains are not specific to the random seed used in the main experiments.

Table 8: FID across different random seeds using the same guidance settings. Seed 42 corresponds to the results reported in Table 1.

Backbone	Seed 0	Seed 42	Seed 2027
LlamaGen-L	3.14	3.25	3.30
AliTok-L	1.28	1.27	1.26

D.4 Additional Qualitative Results

D.4.1 Additional Examples of Guidance Composition

We provide additional examples of the strength-dependent changes in guidance composition discussed in Sec. 4.2. Figs. 19 and 20 visualize the effect of varying λ at lower and higher guidance strengths, respectively.

D.4.2 Additional Examples of Successive Local Updates

We further visualize the effect of successive local updates discussed in Sec. 4.3. Fig. 21 compares different values of K under lower and higher parallel weights.

D.4.3 Comparison with Standard CFG

Fig. 22 compares standard CFG with the guidance setting derived from our analysis on AliTok-XL across a broader set of ImageNet classes.

D.4.4 Comparison under Text Conditioning

We also compare standard CFG with our guidance setting on SimpleAR. Fig. 23 shows representative generations across different GenEval prompts.

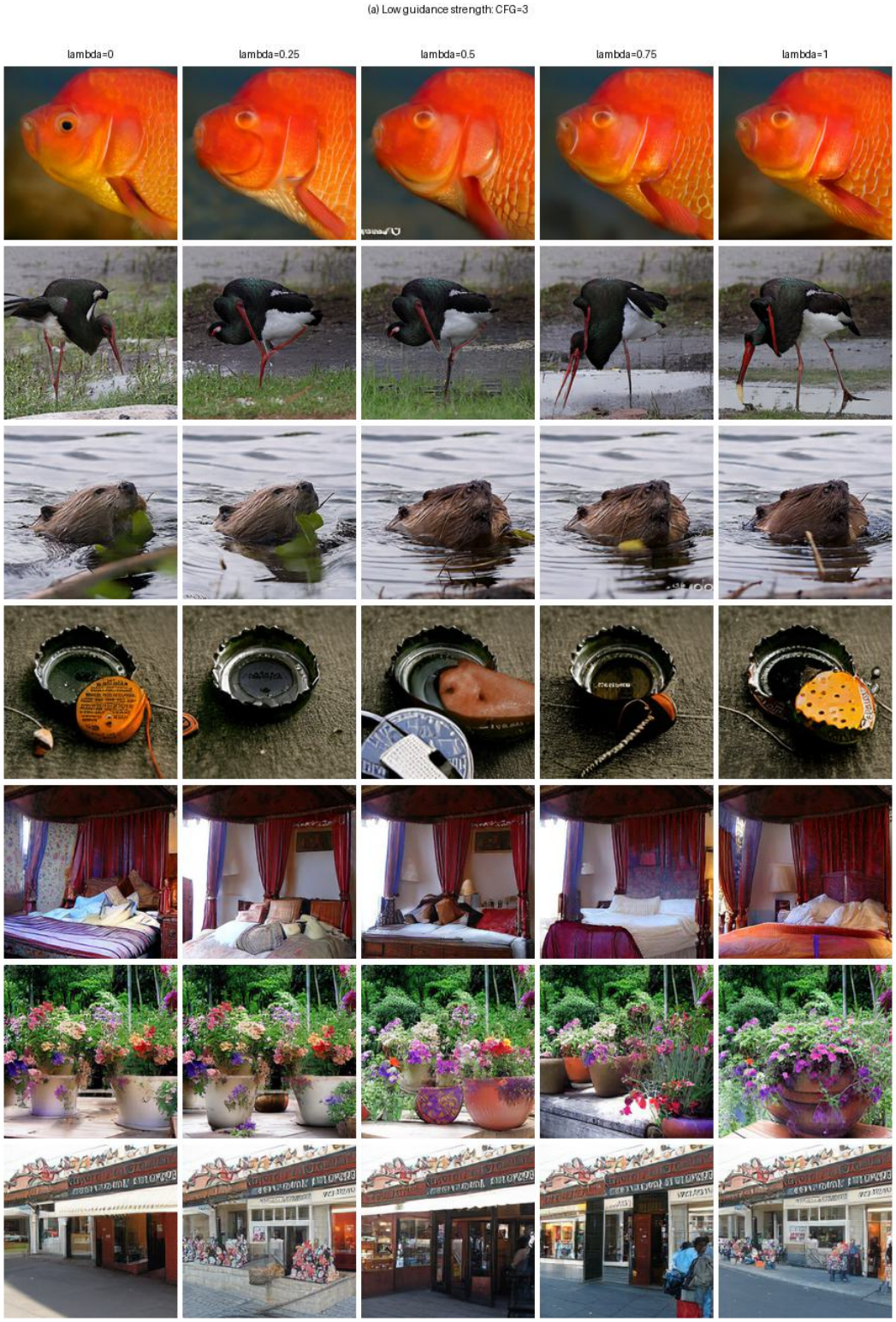


Figure 19: **Guidance composition under lower guidance strength.** Each row varies the composition ratio λ from 0 to 1 at a fixed guidance strength.

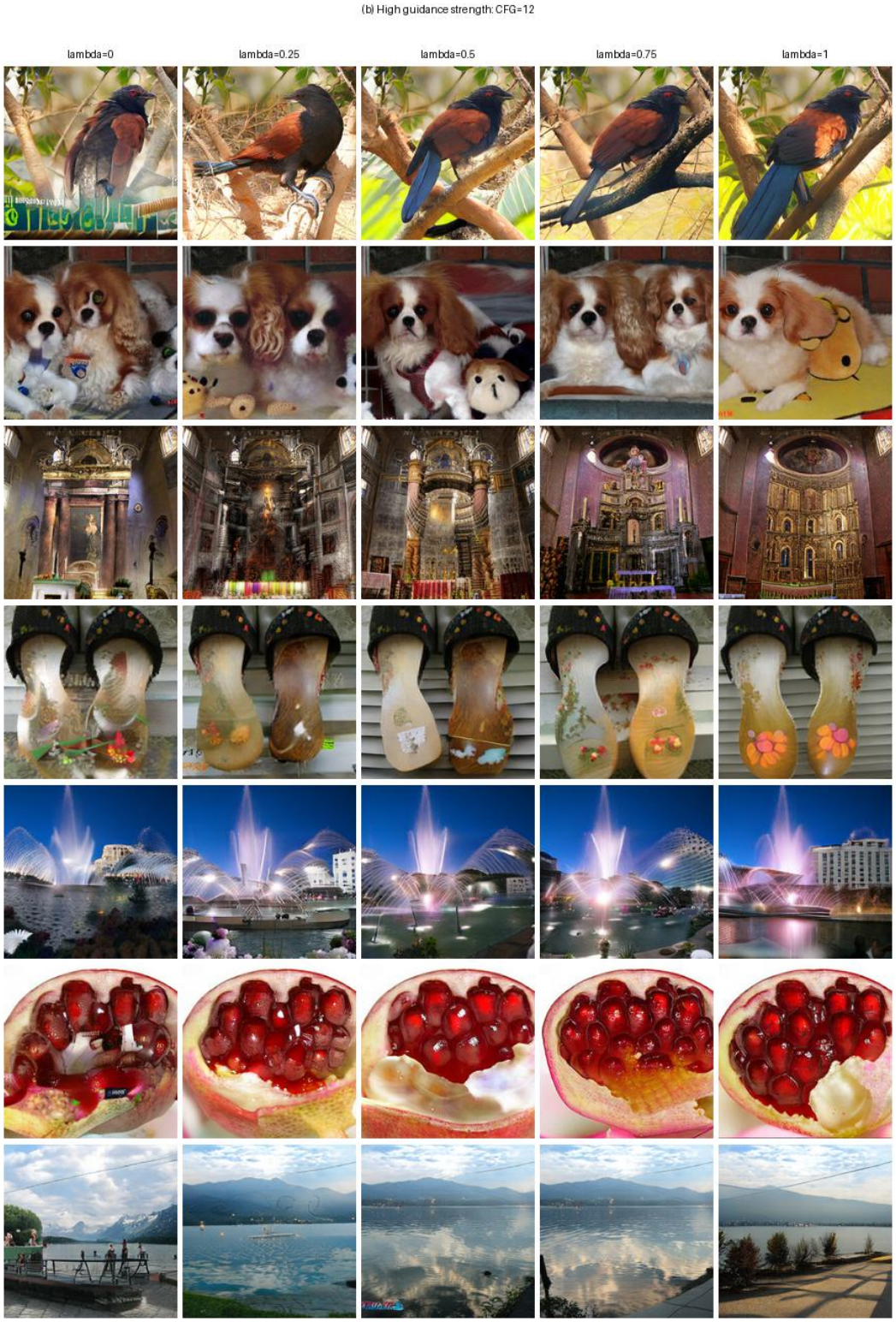


Figure 20: **Guidance composition under higher guidance strength.** Each row varies the composition ratio λ from 0 to 1 at a fixed guidance strength.

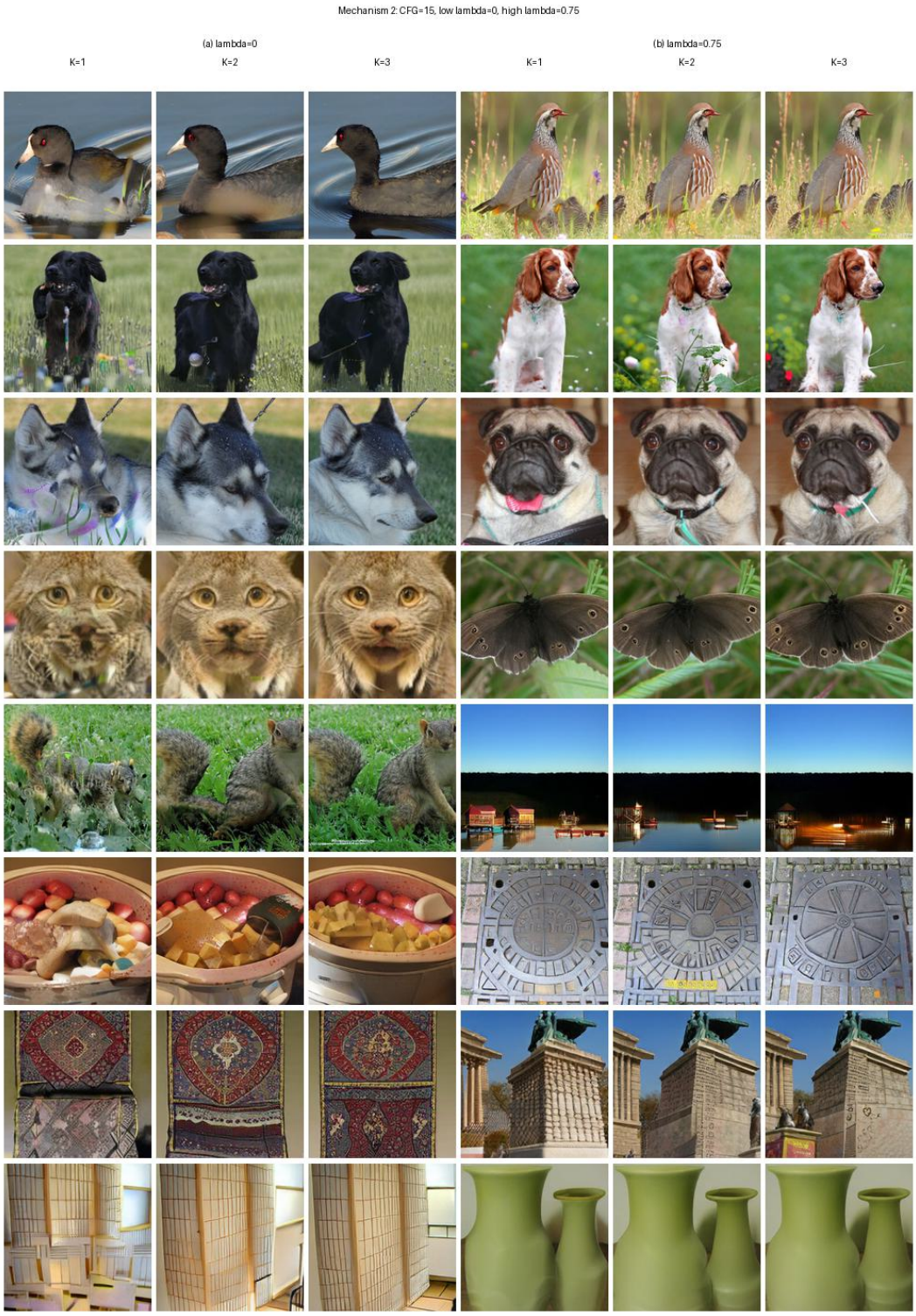


Figure 21: **Additional examples of successive local updates.** We vary the number of local updates K at fixed guidance strength and composition, under (a) a lower parallel weight and (b) a higher parallel weight.

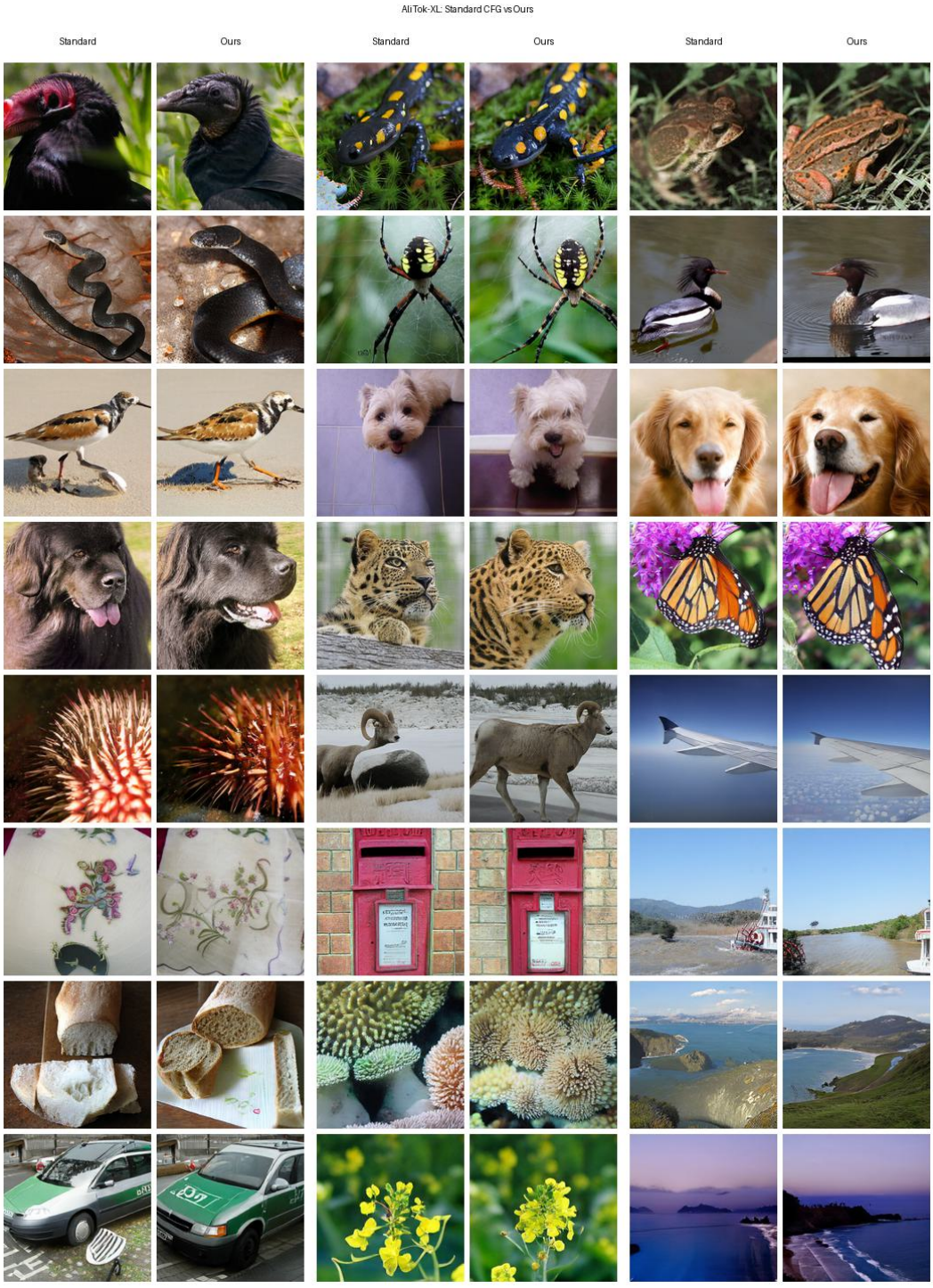


Figure 22: **Qualitative comparison with standard CFG on AliTok-XL.** Each pair compares standard CFG with the guidance setting derived from our analysis for the same ImageNet class.



Figure 23: **Qualitative comparison with standard CFG on SimpleAR.** Each pair shows generations from the same text prompt using standard CFG and our guidance setting.