Scaled Momentum Guidance for Flow Models

Wooyeol Baek* Yonsei University
Seongdo Kim* Yonsei University
Jinseong Kim* Yonsei University
Jongyoo Kim† Yonsei University

{wooyeol.baek, sdokim07, kjs007549, jy.kim}@yonsei.ac.kr
* Equal contribution † Corresponding author

Figure 1. Comparisons between CFG and **SMG(Ours)**

Abstract

*We propose **Scaled Momentum Guidance (SMG)**, a simple, training-free sampling method with no additional computation cost that improves quality and alignment in diffusion and flow-based generative models. While guidance methods like Classifier-Free Guidance (CFG)[9] improve conditional fidelity, they often introduce artifacts due to over-amplification and error accumulation. SMG mitigates this by incorporating a momentum term that stabilizes trajectories and encourages convergence toward coherent semantic modes. A scheduling function further controls guidance strength over time, balancing structure and diversity. Evaluated on the MS COCO dataset[17] with the text-conditional SD3 model [6], SMG consistently improves FID [8] and CLIPScore [21], and outperforms CFG, PAG [1], and CFG++ [4], demonstrating strong generalization across sampling regimes.*

1. Introduction

Generative models based on continuous-time dynamics—such as diffusion probabilistic models (DPMs)[12, 24, 25] and continuous normalizing flows (CNFs)[3, 19]—have demonstrated impressive capabilities in high-fidelity data synthesis. DPMs operate by reversing a stochastic noise process via learned score functions, while CNFs use neural ODEs to deterministically map samples from a base distribution to the data manifold. Recent theoretical work unifies both approaches under a common probabilistic flow (PF) framework [12, 25], offering a continuous-time interpretation of generative dynamics. Despite this theoretical clarity, effective sampling remains a challenge due to the multimodal nature of the data and the instability of trajectories in low-density regions [13].

Diffusion and flow-based models inherently predict averages over multiple data modes [2, 7], resulting in semantically ambiguous or blended outputs []. When combined with guidance mechanisms such as Classifier-Free Guidance (CFG)[9, 27], these models are more likely to reach high-probability regions. However, this process amplifies the influences of multiple conditional direction over the generative trajectory, often leading to undesirable artifacts such as oversaturation, loss of diversity, or semantic drift [13, 16, 23]. These artifacts are particularly problematic because generation is sequential: samples at each timestep are conditioned on the previous state, and early ar-

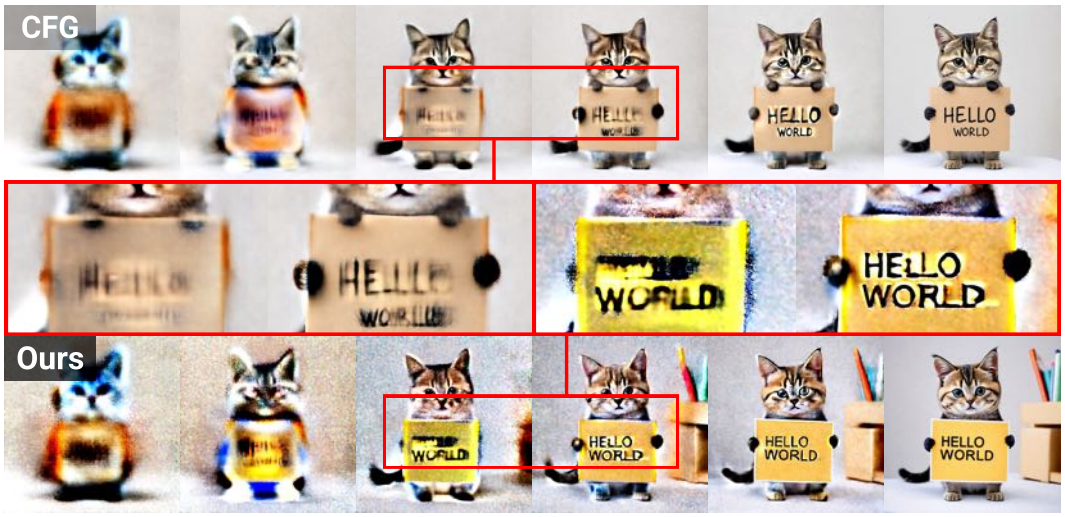

Figure 2. Visualization of $\mathbf{v}_{0|t}$ across timesteps under CFG and SMG, where $\mathbf{v}_{0|t}$ denotes the model's prediction of the final output image given the intermediate timestep t. CFG produces overlapped objects, resulting in unnatural artifacts that persist throughout sampling (four duplicated cat paws). In contrast, SMG resolves such ambiguity in the early stages, yielding coherent object structures that remain stable in subsequent steps (two properly positioned paws).

tifacts tend to persist throughout the process, making them difficult to eliminate in later stages [26]. To alleviate these issues, two main research directions have been proposed. One line constructs weak models to hueristically redirect generation toward more coherent outcomes—examples [1, 10, 11, 13]. Another line of work focuses on decomposing the CFG vector into content-relevant and quality-relevant components, allowing selective amplification of the latter [15, 22].

To overcome these limitations, we propose **Scaled Momentum Guidance (SMG)**—a simple, training-free method with no neural computational cost that improves sampling stability and output quality while maintaining semantic fidelity. SMG modifies the standard CFG formulation by introducing a momentum term that tracks the deviation between the current sampling direction and the accumulated trajectory. This encourages convergence toward a single coherent semantic mode, rather than a blended average of multiple plausible interpretations. In addition, SMG incorporates a scheduling function that applies stronger guidance in the early stages—where layout ambiguity is highest—and gradually reduces its effect over time. This allows the model to make confident semantic decisions early while preventing overcorrection in later steps. SMG does not require retraining, neural network modification, or additional model evaluations, and can be seamlessly integrated into existing diffusion or flow-based generative frameworks.

We evaluate SMG on the MS COCO dataset using the text-conditional SD3 model [6], focusing on two key aspects. First, we demonstrate the scalability of SMG in improving both sample quality and text-image alignment across a wide range of guidance and SMG scales. SMG consistently enhances generation performance as measured by FID [8] and CLIPScore [21], indicating its effectiveness in navigating the trade-off between fidelity and conditional consistency. Second, we compare SMG with other guidance methods including CFG [9], PAG [1], and CFG++ [4]. Experimental results show that SMG consistently outperforms these baselines in both quantitative metrics and qualitative visual quality, demonstrating strong generalization across diverse sampling regimes.

2. Method

2.1. Overlapped Multi-mode Layout

In flow-based models [18, 20], overlapping image layouts may implicitly form during early timesteps, Figure 2, due to the intrinsic properties of ordinary differential equations (ODEs). This occurs because the predicted vector at a given timestep represents an average of multiple possible trajectories, potentially resulting in blended layouts in the generated images, as illustrated in the first row of Figure. 2. The sampled outputs look plausible locally but fail to correspond to any individual semantic mode. Consequently, the generated layout may appear coherent on average, yet fail to reflect any specific, valid configuration. To overcome this limitation, we propose a strategy to decisively select one coherent mode and reinforce its corresponding denoising direction.

2.2. Scaled Momentum Guidance

To resolve directional ambiguity among multiple coexisting modes, we propose Scaled Momentum Guidance

(SMG)—a novel guidance formulation inspired by the momentum mechanism. [5, 14, 22] SMG leverages the discrepancy between the current prediction direction and the accumulated momentum to guide the sampling trajectory toward a distinct and coherent semantic mode.

Momentum-guided Direction Nudging In CFG [9] and Guided Flow [27], the predicted direction is defined as:

$$\mathbf{v}_t^{cfg} = (1+w)\cdot \mathbf{v}_t^{\text{cond}} - w\cdot \mathbf{v}_t^{\text{uncond}} \tag{1}$$

where w denotes the guidance scale. However, this linear interpolation tends to blend multiple plausible layout modes, making it difficult to steer denoising toward a decisive semantic direction when ambiguity is high.

To overcome this limitation, we enhance CFG's prediction using a momentum vector and define SMG as follows:

$$\mathbf{v}_t^{smg} = \mathbf{v}_t^{cfg} + \gamma(t)(\mathbf{v}_t^{cfg} - \mathbf{v}_t^{mom}) \tag{2}$$

where $\mathbf{v}_t^{cfg}$ is the current vector prediction and $\mathbf{v}_t^{mom}$ is the accumulated momentum up to the previous timestep. The guidance scale $\gamma(t)$ is defined as a function of the timestep t. It is designed to apply strong guidance solely during the early stages, when the object-overlapping artifact is most likely to occur. Additionally, a scaling factor η is introduced to control the maximum value of $\gamma(t)$, allowing flexible adjustment of the SMG strength during sampling. $\mathbf{v}_t^{cfg} - \mathbf{v}_t^{mom}$ captures the deviation from the momentum vector, thus offers a differentiated direction that facilitates convergence toward a single dominant mode rather than a blended average. Therefore, $\mathbf{v}_t^{smg}$ is our refined direction. The final guidance direction is then constructed by scaling this refined direction by $\gamma(t) = \eta \cdot f$.

Furthermore, to avoid instability during sampling, we modulate the guidance strength over time using a scheduling function $\gamma(t)$, which is empirically modeled by a logit-normal distribution. The scheduling function is designed to serve three distinct roles:

- **Early Stage:** $\gamma(t)$ rises sharply, softly nudging the sampling trajectory toward a coherent layout when multiple semantic modes coexist.
- **Middle Stage:** $\gamma(t)$ decays, reducing the correction strength to avoid over-biasing and allowing a natural separation of modes.
- **Late Stage:** $\gamma(t)$ approaches zero, recovering standard sampling behavior and enabling fine-grained refinement without altering the global structure.

This gradual modulation allows SMG to resolve early-stage ambiguities by softly guiding the sampling process toward a plausible configuration, while maintaining stability and semantic consistency throughout the trajectory. Moreover, by ensuring that the integral of $\gamma(t)$ over the full trajectory is close to zero, the scheduling design avoids cumulative drift and preserves the overall flow structure.

	FID(↓) [8] / CLIPScore (↑) [21]			
	$\eta = 0 (CFG)$	$\eta = 1.0$	$\eta = 1.5$	$\eta = 2.0$
$\omega = 1.0$	89.68 / **29.48**	**77.56** / 29.45	80.11 / 29.27	90.65 / 29.06
$\omega = 2.0$	44.79 / **31.20**	**41.45** / 31.06	45.72 / 30.92	42.05 / 30.99
$\omega = 3.0$	39.65 / **31.48**	37.76 / 31.37	37.71 / 31.33	**36.79** / 31.55
$\omega = 4.0$	38.75 / **31.55**	37.53 / 31.45	37.47 / 31.35	**36.98** / 31.40
$\omega = 5.0$	39.88 / **31.56**	38.41 / 30.71	37.54 / 30.70	**36.56** / 30.68
$\omega = 6.0$	39.97 / **31.57**	**36.81** / 31.43	37.33 / 31.54	37.54 / 30.68

Table 1. Comparison of FID (↓) [8] and CLIP score (↑) [21] on the MS COCO dataset [17] using SD 3 [6], evaluated under various CFG (ω) and SMG (η) scales. Our method achieves a more efficient trade-off between FID and CLIP scores compared to CFG.

SD3 [6]	CFG [9]	PAG [1]	CFG++ [4]	**+SMG (Ours)**
FID (↓)	39.65	37.61	89.71	**36.79**
CLIP score (↑)	31.48	**31.67**	29.48	31.35

Table 2. Comparison of guidance methods on MS COCO using the SD3 model. SMG achieves the best FID score, indicating superior visual fidelity, and ranks second in CLIPScore, reflecting strong text-image alignment. These results demonstrate that SMG effectively balances perceptual quality and conditional consistency without requiring retraining or additional computation.

3. Experiments

We evaluate proposed method on the MS COCO dataset [17] using Stable Diffusion 3 [6] with 30 sampling steps, focusing on both quantitative metrics and human evaluations. All baseline methods follow their official implementations and recommended best settings for fair comparison.

3.1. Quantitative Results

FID / CLIP score tradeoff In Table 1, our method is compared with standard CFG with varying guidance scales ($w = 1, 2, 3, 4, 5$ for CFG and $\eta = 1.0, 1.5, 2.0$ for SMG), which shows that ours achieves a more favorable trade-off between FID and CLIP score compared to CFG. Specifically, SMG significantly reduces FID while maintaining CLIP scores close to those of standard CFG, indicating that SMG improves sample quality without sacrificing semantic relevance between text and image.

Comparison with Existing Guidance Methods In Table 2, SMG is compared with other guidance methods on Stable Diffusion 3 (SD3), adhering to the recommended settings from the original authors. Our method substantially improves FID over existing methods while maintaining competitive CLIP scores. To complement quantitative results, we also conduct extensive qualitative evaluations, including user studies, which confirm that our method produces images with better text-image alignment and overall perceptual quality.

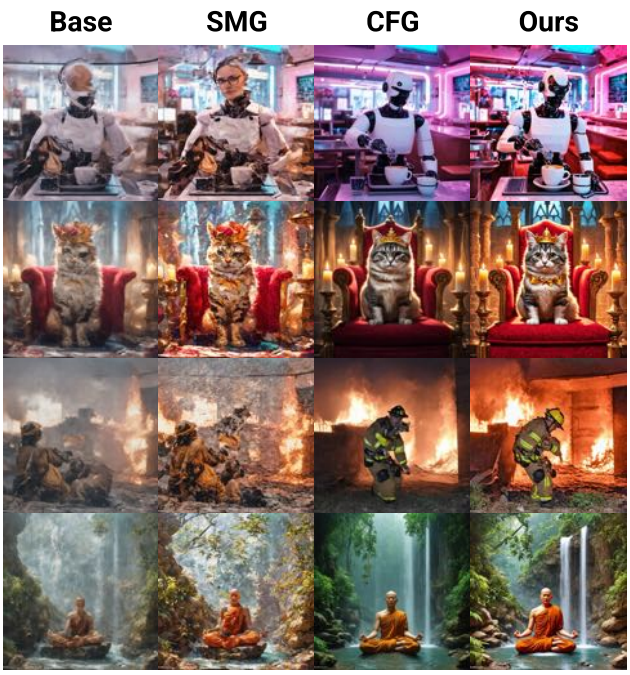

Figure 3. Ablation study. 'Base' is conditional generation, 'SMG' is Base with momentum nudging, 'CFG' is sampling with CFG, and 'Ours' is CFG combined with SMG.

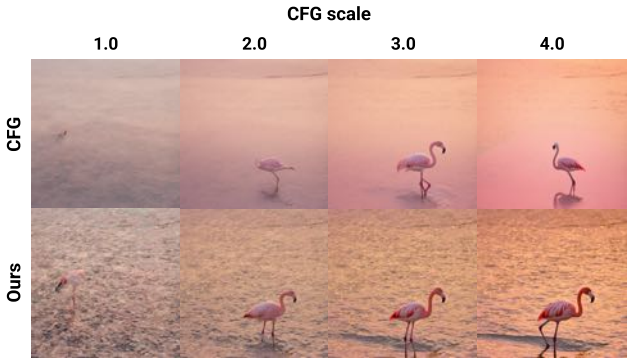

Figure 4. CFG scale comparison. "A flamingo standing in shallow pink-tinted salt lakes under a golden sunrise."

3.2. Qualitative Results

Ablation Study To validate the effectiveness of SMG, we conducted a series of ablation studies. Figure 3 compares the impact of SMG under both conditional-only and CFG settings. In the conditional setting (left), SMG effectively refines ambiguous layouts, leading to overall improvements in image quality. In the CFG setting (right), SMG not only corrects unnatural object configurations but also enhances visual clarity and better captures fine-grained details specified by the text conditions. Second, Figure 4 analyzes text-image alignment performance across different CFG scales. Across all tested scales, SMG consistently outperforms CFG in accurately reflecting text conditions, demonstrating stronger semantic alignment with key attributes (e.g., "flamingo", "golden", etc.). These results highlight SMG's robustness in both layout refinement and semantic consistency under various guidance configurations. Additional results and visualizations can be found in Appendix C.

Figure 5. Qaulitative comparison with the prompt: "A arctic fox curled up on a glowing ice crystal under the aurora corealis.", "A parade of forgotten toys marching through a rainy alley."

User study To complement the quantitative results with a more objective qualitative assessment, we conducted a user study with approximately 50 participants, comparing SMG and CFG in terms of image quality and text-image alignment. SMG was preferred in over 70% of cases for both criteria, confirming its effectiveness in improving perceptual quality and semantic consistency. Due to page limitation, details are provided in Appendix B.

Comparison with Existing Guidance Methods To further validate SMG, we conduct qualitative comparisons against other guidance methods (CFG++ and PAG) as shown in Figure 5 and Appendix A. Our method consistently achieves better image quality and text-image alignment, preserving fine-grained details (e.g., rainy scenes) and enhancing global layout coherence (e.g., curled shapes). Overall, SMG improves layout robustness and semantic consistency in flow-based generation while maintaining competitive or better performance than recent guidance techniques.

4. Conclusion

We proposed an empirical guidance method inspired by the linear trajectories of flow models and momentum dynamics, encouraging early convergence to a single semantic mode. Therefore, SMG outperforms existing methods in text-image alignment and image quality, achieving the best FID and strong text-image alignment validated by human preference.

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance. In *The 18th European Conference on Computer Vision ECCV 2024*. arXiv, 2024. 1, 2, 3

[2] Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian Mixture Flow Matching Models. In *Forty-Second International Conference on Machine Learning*. arXiv, 2025. 1

[3] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 1

[4] 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*, 2024. 1, 2, 3

[5] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In *International Conference on Learning Representations*, 2021. 3

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Forty-First International Conference on Machine Learning*, 2024. 1, 2, 3

[7] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One Step Diffusion via Shortcut Models. In *The Thirteenth International Conference on Learning Representations*, 2024. 1

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 2, 3

[9] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 2, 3

[10] Susung Hong. Smoothed Energy Guidance: Guiding Diffusion Models with Reduced Energy Curvature of Attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[11] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving Sample Quality of Diffusion Models Using Self-Attention Guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023. 2

[12] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, 2022. 1

[13] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a Diffusion Model with a Bad Version of Itself. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The Third International Conference on Learning Representations*. arXiv, 2014. 3

[15] Mingi Kwon, Shin seong Kim, Jaeseok Jeong Yi Ting Hsiao, and Youngjung Uh. TCFG: Tangential Damping Classifier-free Guidance. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025*. arXiv, 2025. 2

[16] 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. 1

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *The 13th European Conference on Computer Vision ECCV 2014*. arXiv, 2015. 1, 3

[18] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[19] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow Matching Guide and Code, 2024. 1

[20] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[22] Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models. In *The Thirteenth International Conference on Learning Representations*, 2024. 2, 3

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, 2022. 1

[24] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using

Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[25] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2020. 1

[26] Suttisak Wizadwongsa, Worameth Chinchuthakun, Pramook Khungurn, Amit Raj, and Supasorn Suwajanakorn. Diffusion Sampling with Momentum for Mitigating Divergence Artifacts. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[27] Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided Flows for Generative Modeling and Decision Making. *CoRR*, 2023. 1, 3

Scaled Momentum Guidance for Flow Models

Supplementary Material

A. Model Comparison

Figure 6. **Comparison with recent guidance methods.** ”A desk lamp on a wooden pier, its light reflecting softly on the lake at night.”, ”A cat-shaped spaceship landing on a candy planet surrounded by gumdrop tress.”, ”A toolbox placed carefully on a picnic blanket in a meadow.”, ”A snow globe containing a snowy cabin with warm lights and pine trees.”

B. User Study

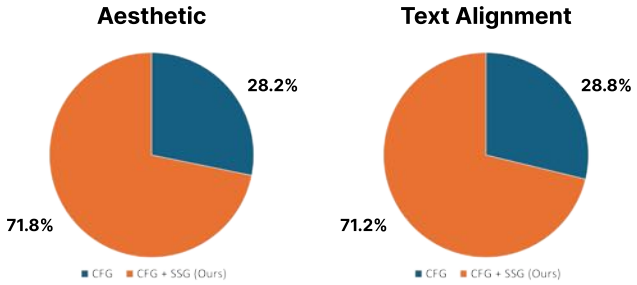

Figure 7. **The results of the user study.**

C. Extended results on MS COCO

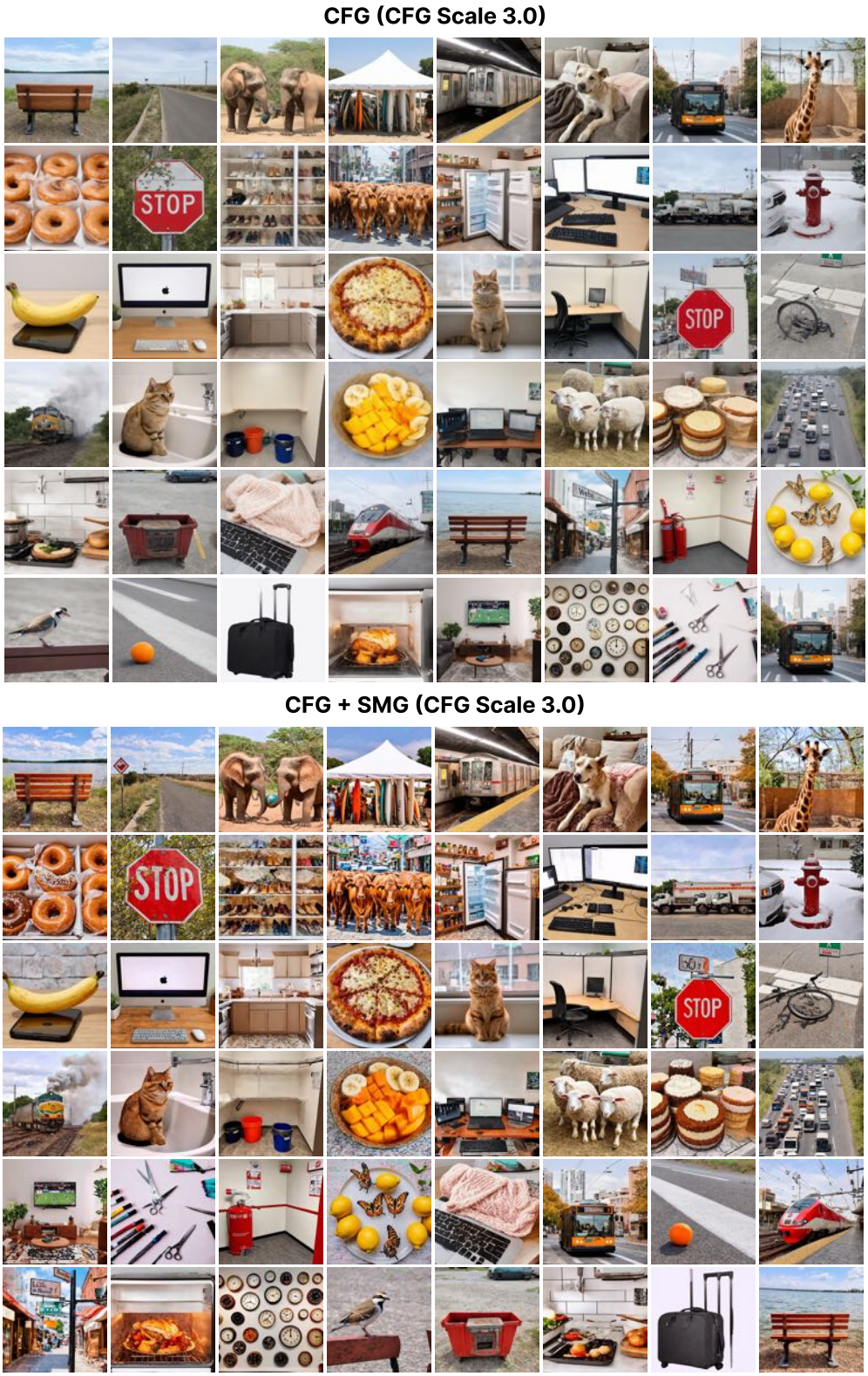

Figure 8. **MS-COCO Comparison.** CFG scale 3.0, SMG scale 2.0

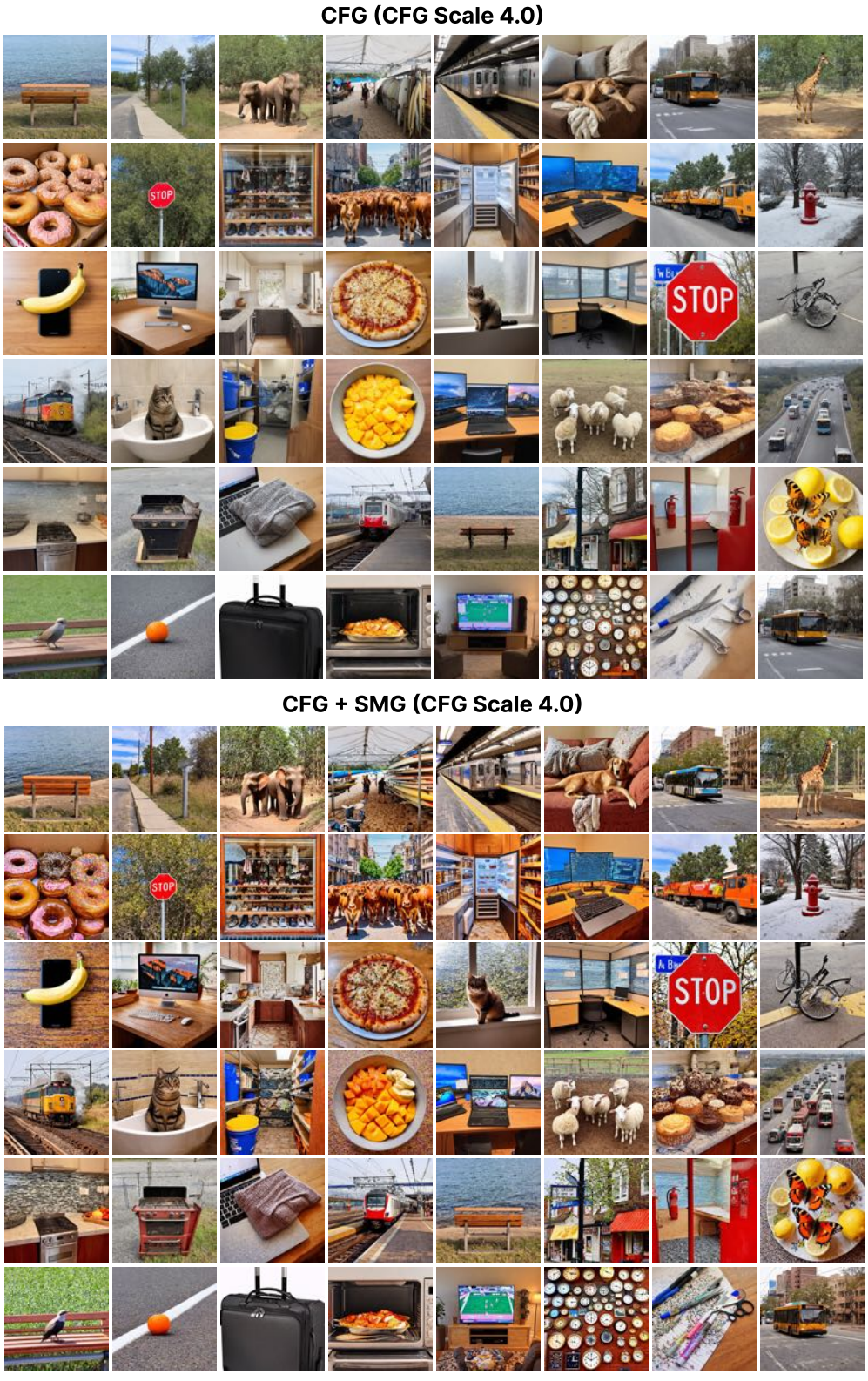

Figure 9. MS-COCO Comparison. CFG scale 4.0, SMG scale 2.0