Improving Classifier-Free Guidance of Flow Matching

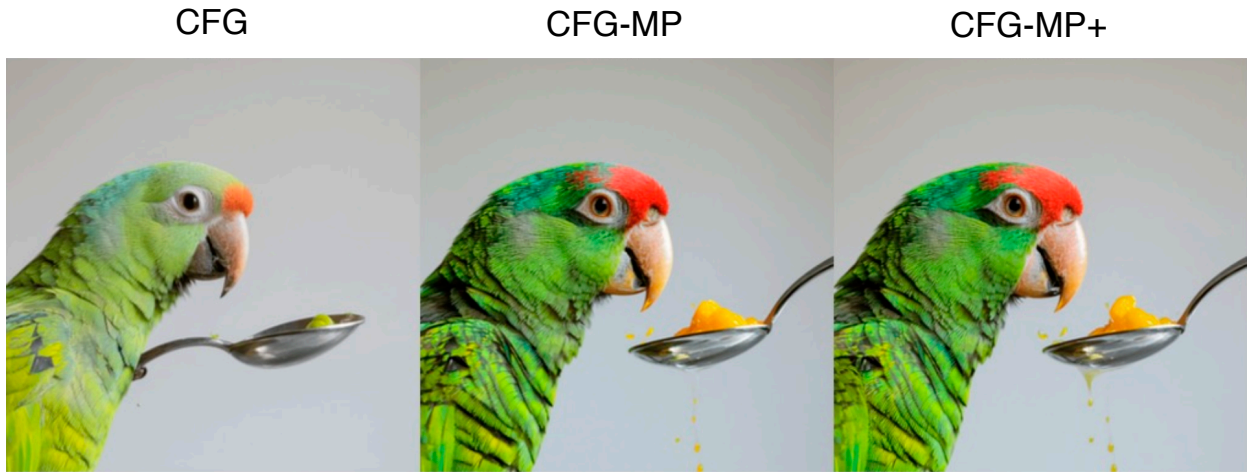
via Manifold Projection

CFG-MP · ICML 2026

Training-free projection for making CFG less sensitive to guidance scale.

Jian-Feng Cai · Haixia Liu · Zhengyi Su · Chao Wang

github.com/LeonSuZhengYi/CFG-MP



Prompt: Close-up picture of a parrot dropping a spoon.

Prompt: A war weary hamster soldier.

CFG works, but the guidance scale has no objective

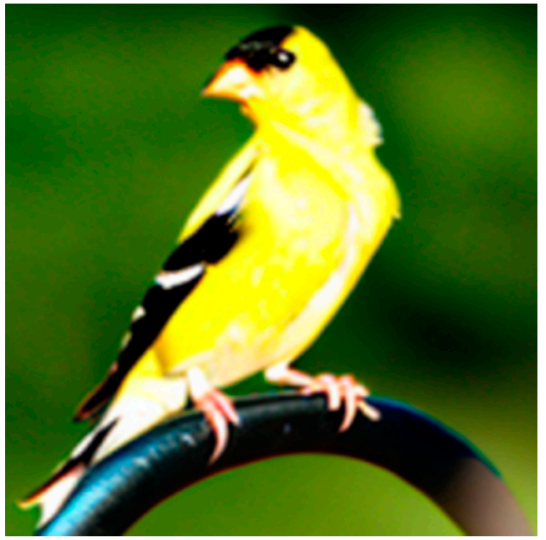
$$\begin{cases} v_\theta^{\text{cfg}}(t,x,y) = v_\theta(t,x,\varnothing) + w \cdot (v_\theta(t,x,y) - v_\theta(t,x,\varnothing)), \;\; w \geq 1. \\ x_{t+\Delta t} = x_t + v_\theta^{\text{cfg}}(t,x,y) \cdot \Delta t \,. \end{cases}$$

Practical weakness

- small w: weak condition alignment
- large w: oversaturation / artifacts

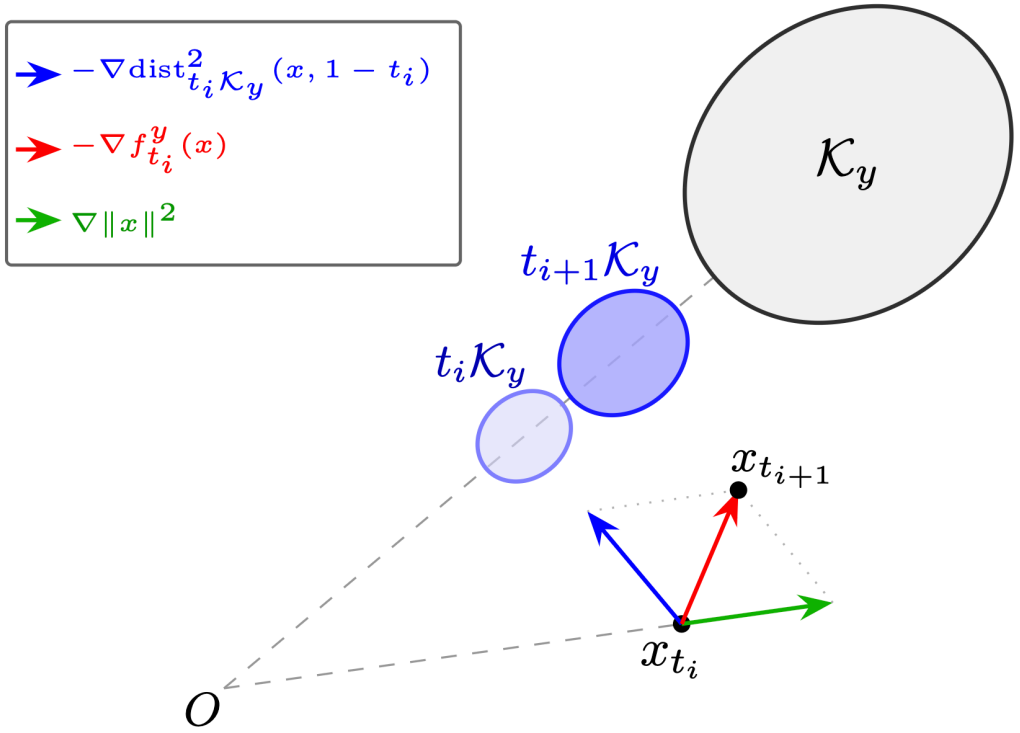
Question: can we explain CFG as an approximation, and thus derive an optimal w?

$w = 1.0$



$w = 10$

With ideal velocity, flow matching is homotopy optimization



$$\begin{cases} v^*_{t,y}(x) = -\dfrac{1}{2t(1-t)} \nabla_x \left(\mathrm{dist}^2_{t\mathcal{K}_y}(x, 1-t) - (1-t)\|x\|^2 \right), \\ x_{t_{i+1}} = x_{t_i} + v^*_{t_i,y}(x_{t_i})\, \Delta t\,. \end{cases}$$

The target set moves from small $t\mathcal{K}$ toward the data set $\mathcal{K}$, x_t is attracted by $t\mathcal{K}$.

This gives CFG an objective-level interpretation, but only through the ideal velocity field.

The optimal guidance scale is an approximation knob

$$w^* = \frac{\left\langle v_\theta(t,x,y) - v_\theta(t,x,\varnothing),\, v^*_{t,y}(x) - v_\theta(t,x,\varnothing) \right\rangle}{\left\| v_\theta(t,x,y) - v_\theta(t,x,\varnothing) \right\|^2}$$

Exists in theory

It minimizes approximation error to the ideal conditional velocity.

Hard in practice

It depends on the unknown ideal velocity and varies with class / prompt.

Better target

Instead of finding w*, we seek for another sensitivity term.

Error decomposition turns w-tuning into gap reduction

$$\left\| v_\theta^{\mathrm{cfg}}(t,x,y) - v_{t,y}^*(x) \right\|^2 = \left\| v_\theta^{\mathrm{cfg}^*}(t,x,y) - v_{t,y}^*(x) \right\|^2 + (w^* - w)^2 \left\| v_\theta(t,x,y) - v_\theta(t,x,\varnothing) \right\|^2 .$$

approximation error = intrinsic model error + **(scale error)^2 × prediction gap**

We do not need to solve for w* directly.

CFG-MP targets the prediction gap during sampling.

CFG-MP: sample, then project toward zero prediction gap

1 CFG step

move with extrapolated velocity

2 Projection

Implemented by fixed-point iteration

3 Next state

closer to the ideal trajectory

Training-free

no retraining inserted at sampling time

A standard CFG step is followed by a projection step on the zero manifold $\mathscr{M}_t$.

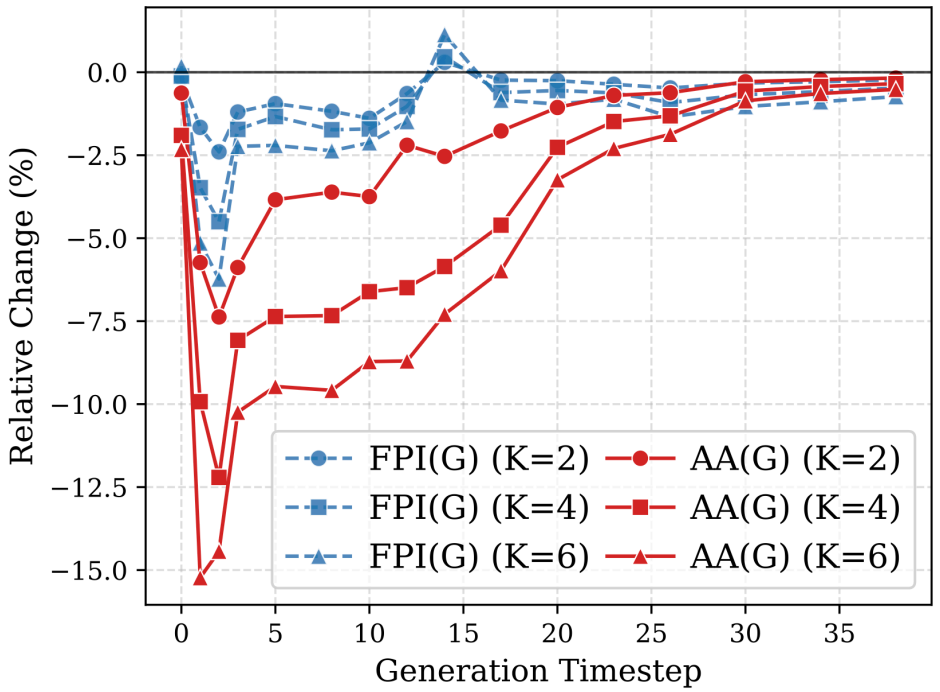
$$x_{i+1} = \mathrm{Proj}_{\mathscr{M}_{t_{i+1}}}\Big(x_i + \Delta t\, v_\theta^{\mathrm{cfg}}(t_i, x_i, y)\Big), \quad \mathscr{M}_t := \Big\{ z \;\Big|\; v_\theta(t, z, y) = v_\theta(t, z, \varnothing) \Big\},$$

$$G(x,t) := x - \frac{1}{2}\Delta t\, v_\theta(t, x, \varnothing) + \frac{1}{2}\Delta t\, v_\theta\Big(t, x - \frac{1}{2}\Delta t\, v_\theta(t, x, \varnothing), y\Big),$$

$\mathrm{Proj}_{\mathscr{M}_{t_{i+1}}}(\cdot) \approx$ Fix point iteration of $G(\cdot, t_{i+1})$.

CFG-MP+: Anderson acceleration makes FPI fast and stable

Lower relative change indicates smaller prediction gap.



- **CFG-MP:** $z_{k+1} = G(z_k, t)\,.$

- **CFG-MP+:**

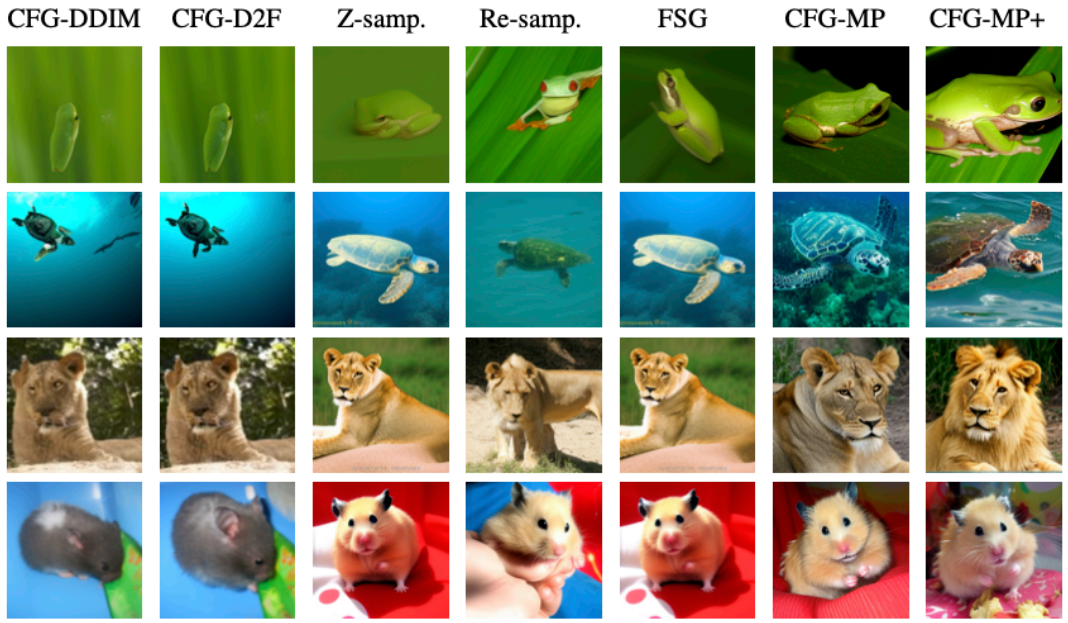
$$\begin{cases} f_k = G(z_k, t) - z_k, \\ \alpha^{k+1} = \underset{\alpha}{\operatorname{argmin}} \left\{ \left\| \sum_{i=k-m_k}^{k} \alpha_i f_i \right\|_2 : \sum_{i=k-m_k}^{k} \alpha_i = 1 \right\}, \\ z_{k+1} = \sum_{i=k-m_k}^{k} \alpha_i^{k+1} G(z_i, t)\,. \end{cases}$$

Observed behavior: vanilla fixed-point iteration can stagnate in middle stages; AA makes relative change more negative.

No extra evaluations of G.

Class-to-image: stronger, more robust ImageNet samples

DiT-XL-2-256 on ImageNet256; compared across three guidance scales and two NFE settings.



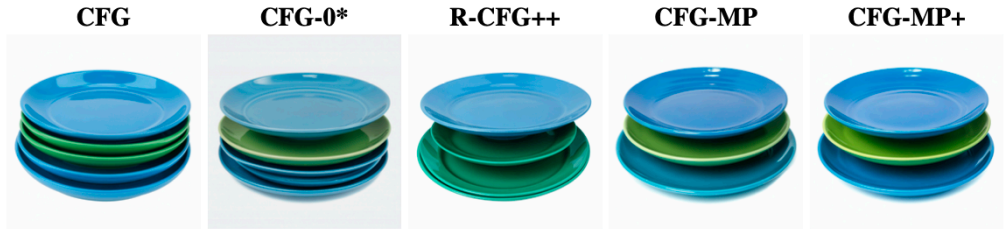
Guidance Scale	$\omega = 1.5$				$\omega = 2.0$				$\omega = 2.5$			
NFEs	60	120	60	120	60	120	60	120	60	120	60	120
Methods	FID (↓)		IS (↑)		FID (↓)		IS (↑)		FID (↓)		IS (↑)	
CFG(DDIM)	11.76	9.61	61.39	61.86	13.29	12.17	67.98	69.25	15.83	14.71	72.53	73.17
Z-sampling	11.53	9.21	63.56	63.91	12.85	11.68	70.47	72.52	15.21	14.07	74.82	75.86
Re-sampling	11.48	8.56	66.98	67.87	12.92	11.63	71.88	73.07	15.17	14.14	75.25	76.01
FSG	11.42	9.14	67.20	65.73	12.76	11.50	72.31	74.75	15.04	14.01	75.98	77.14
CFG(D2F)	8.95	**7.45**	65.90	66.27	10.26	**9.18**	70.57	71.48	12.85	11.49	76.26	77.83
CFG-MP	**8.85**	7.77	72.85	73.11	**10.03**	9.26	78.74	79.25	12.51	**11.48**	82.45	83.74
CFG-MP+	9.21	7.91	**76.78**	**78.26**	10.28	9.34	**84.67**	**86.26**	**12.30**	11.49	**88.72**	**90.45**

Main empirical message: the projection view is not just interpretive; it improves sample quality across guidance regimes.

Qualitatively, CFG-MP/MP+ sharpen textures and reduce over-smoothed surfaces.

Text-to-image: better binding and preference metrics

SD3.5 and Flux-dev experiments show gains where CFG often fails: attributes, objects, and composition.



Prompt: *A stack of 3 plates. The one on the top and bottom is blue, and the middle one is green.*

Prompt: *A vehicle composed of two wheels held in a frame one behind the other, propelled by pedals and steered with handlebars attached to the front wheel.*

Prompt: *Cute anime neko cat girl.*

Prompt: *An astronaut riding a horse.*

CFG-MP/MP+ turns guidance from heuristic extrapolation into a training-free projection step.

Thank you!

Paper (arXiv)

https://arxiv.org/abs/2601.21892

Code (GitHub)

https://github.com/LeonSuZhengYi/CFG-MP