CFG-Ctrl: Control-Based Classifier-Free Diffusion Guidance

Supplementary Material

1. More Theoretical Analysis

1.1. Notation Table

To facilitate the understanding of the theoretical derivations of CFG-Ctrl and SMC-CFG, we summarize the main symbols, their corresponding technical meanings, and relevant mathematical expressions or value constraints in Table 1. These notations cover core components such as velocity fields, semantic error signals, and stability analysis metrics, providing a clear reference for readers to follow the logical flow of the proposed framework.

1.2. Additional CFG Variants

CFG-Zero⋆ [2] introduces an optimizable scalar $s^\star \in \mathbb{R}_{>0}$ into the standard CFG framework, with its guided velocity field formulated as:

$$\begin{aligned}\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c}) &= (1-w)\cdot s^\star \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) + w\,\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) \\ &= s^\star \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) \\ &\quad + w \cdot \big(\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - s^\star \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)\big).\end{aligned} \tag{1}$$

As summarized in Table 1 of the main text, under the CFG-Ctrl paradigm, the guidance schedule K_t and direction operator Π_t of CFG-Zero⋆ are modeled as:

$$\begin{aligned}K_t &= \begin{bmatrix} wI & \frac{s^\star}{1-s^\star} I \end{bmatrix}, \quad s^\star = \frac{\mathbf{v}_\theta(\mathbf{c})^\top \mathbf{v}_\theta(\varnothing)}{|\mathbf{v}_\theta(\varnothing)|^2}), \\ \Pi_t &= \begin{bmatrix} I - P_t \\ P_t \end{bmatrix}, \quad P_t = \frac{\mathbf{v}_\theta(\varnothing)\mathbf{v}_\theta(\varnothing)^\top}{|\mathbf{v}_\theta(\varnothing)|^2}.\end{aligned} \tag{2}$$

Substituting these components into the closed-loop dynamics of CFG-Ctrl yields:

$$\begin{aligned}\frac{d\mathbf{x}_t}{dt} &= \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) + K_t\,\Pi_t(e_t) \\ &= \frac{\mathbf{v}_\theta^\top(\mathbf{c})\mathbf{v}_\theta(\varnothing)}{|\mathbf{v}_\theta(\varnothing)|^2}\mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) \\ &\quad + w\left(\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \frac{\mathbf{v}_\theta^\top(\mathbf{c})\mathbf{v}_\theta(\varnothing)}{|\mathbf{v}_\theta(\varnothing)|^2}\mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)\right) \\ &= s_t^\star \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) + w\left(\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - s_t^\star \cdot \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)\right),\end{aligned} \tag{3}$$

where $s_t^\star$ corresponds to $\frac{\mathbf{v}_\theta^\top(\mathbf{c})\mathbf{v}_\theta(\varnothing)}{|\mathbf{v}_\theta(\varnothing)|^2}$. Notably, CFG-Zero⋆ shares a similar design motivation with APG [10]: both adopt orthogonal projection transformations as their direction operators. The key distinction lies in the projection target—APG projects onto the conditional velocity field $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$, while CFG-Zero⋆ projects onto the unconditional velocity field $\mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)$. From a control-theoretic perspective, both methods fall into the category of projection-based structured feedback controllers.

Rectified-CFG++ [11] differs from standard CFG by incorporating not only the error signal derived from the current latent state $\mathbf{x}_t$ (defined as $\Delta\mathbf{v}_\theta(t) = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)$) but also predictive information from a future state $\mathbf{x}_{t-\frac{\Delta t}{2}}$. The error signal for this predicted future state is formulated as:

$$\Delta\mathbf{v}_\theta(t-\frac{\Delta t}{2}) = \mathbf{v}_\theta(\mathbf{x}_{t-\frac{\Delta t}{2}}, t-\frac{\Delta t}{2}, \mathbf{c}) - \mathbf{v}_\theta(\mathbf{x}_{t-\frac{\Delta t}{2}}, t-\frac{\Delta t}{2}, \varnothing), \tag{4}$$

and the guided velocity field of Rectified-CFG++ is given by:

$$\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c}) = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) + \alpha(t)\,\Delta\mathbf{v}_\theta(t - \frac{\Delta t}{2}). \tag{5}$$

As outlined in Table 1 of the main text, within the CFG-Ctrl framework, the guidance schedule K_t and direction operator Π_t of Rectified-CFG++ are structured as:

$$\begin{aligned}K_t &= \begin{bmatrix} I & \alpha(t) I \end{bmatrix}, \quad \alpha(t) = \lambda_{max}(1-t)^\gamma, \\ \Pi_t &= \begin{bmatrix} \Delta\mathbf{v}_\theta(t) \\ \Delta\mathbf{v}_\theta(t - \frac{\Delta t}{2}) \end{bmatrix}.\end{aligned} \tag{6}$$

Substituting these components into the closed-loop dynamics of CFG-Ctrl leads to the following derivation:

$$\begin{aligned}\frac{d\mathbf{x}_t}{dt} &= \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) + K_t\,\Pi_t(e_t) \\ &= \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing) + \Delta\mathbf{v}_\theta(t) + \alpha(t)\Delta\mathbf{v}_\theta(t - \frac{\Delta t}{2}) \\ &= \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) + \alpha(t)\Delta\mathbf{v}_\theta(t - \frac{\Delta t}{2}).\end{aligned} \tag{7}$$

Notably, Rectified-CFG++ adopts a time-varying gain scheduling strategy via $\alpha(t)$, which dynamically adjusts guidance strength throughout the sampling process. Beyond this, the method embodies the core principle of Model Predictive Control, which is a robust control paradigm that leverages a system model to predict future behavior over a finite horizon and optimize control actions accordingly. By integrating error information from the predicted future state $\mathbf{x}_{t-\frac{\Delta t}{2}}$, Rectified-CFG++ effectively anticipates potential deviations in the generative flow and pre-emptively adjusts guidance, thereby enhancing the stability of semantic alignment and the efficiency of the sampling process.

1.3. Theoretical Motivation: Robustness Analysis

In this section, we provide a theoretical motivation for the proposed SMC-CFG framework from a robust control per-

Table 1. **Notation table.**

Notation	Meaning	Value
$\mathbf{x}_t$	Latent state at time t during generative flow sampling.	$\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$ (initial state)
$\mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)$	Unconditional velocity field, obtained by dropping the condition $\mathbf{c}$.	/
$\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$	Conditional velocity field, incorporating the input condition $\mathbf{c}$.	/
$\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c})$	Guided velocity field, combined via guidance.	/
w	CFG guidance scale.	$w \geq 1$
$\mathbf{e}(t)$	Semantic error signal.	$\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)$
$\dot{\mathbf{e}}(t)$	Temporal derivative of the semantic error signal.	/
$\mathbf{u}_t$	General guidance control input.	$\mathbf{u}_t = K_t \Pi_t(\mathbf{e}(t))$
K_t	Guidance schedule matrix/scalar in CFG-Ctrl framework.	/
Π_t	Direction operator in CFG-Ctrl framework.	/
$\mathcal{S}$	Semantic sliding manifold.	$\mathcal{S} = \{(\mathbf{x}, t) \mid \mathbf{s}(t) = \mathbf{0}\}$
$\mathbf{s}(t)$	Sliding mode surface variable in SMC-CFG.	$\mathbf{s}(t) = \dot{\mathbf{e}}(t) + \lambda \mathbf{e}(t)$
λ	Shape parameter of the sliding mode surface.	Hyperparameter
k	Gain of the switching control term.	Hyperparameter
$\Delta\mathbf{e}(t)$	SMC correction term (Switching Control).	$-k \cdot \text{sign}(\mathbf{s}(t))$
$\mathbf{\Phi}(t, \mathbf{x}_t)$	Intrinsic drift dynamics (encapsulating model non-linearities).	$\|\mathbf{\Phi}\|_2 \leq \delta$
$\mathbf{\Gamma}(t)$	Effective control gain matrix (Jacobian of semantic difference).	$\mathbf{\Gamma} = w\mathbf{I} + \Delta\mathbf{\Gamma}(t)$
$\Delta\mathbf{\Gamma}(t)$	Anisotropic deviation from the nominal isotropic guidance.	/
δ	Upper bound of the intrinsic drift dynamics.	$\delta > 0$
ρ	Upper bound of the anisotropic deviation norm.	$\|\Delta\mathbf{\Gamma}\|_2 \leq \rho < w$
ϵ	Positive safety margin for the control gain.	$\epsilon > 0$
$V(\mathbf{s})$	Lyapunov function candidate for stability analysis.	$V(\mathbf{s}) = \frac{1}{2}\|\mathbf{s}\|_2^2$
Δt	Discrete time step size for sampling.	/

spective. Unlike standard CFG, which relies on linear extrapolation and assumes an ideal linear evolution of the semantic error, SMC-CFG explicitly introduces a nonlinear switching term to handle the unmodeled non-linearities and disturbances inherent in the diffusion flow. Our analysis demonstrates that under reasonable robustness assumptions, the proposed controller drives the generative trajectory toward the *semantic sliding manifold* $\mathcal{S} = \{(\mathbf{x}, t) \mid \mathbf{s}(t) = \mathbf{0}\}$.

1.3.1. Dynamics of the Sliding Variable

Let $\mathbf{e}(t) = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}_\theta(\mathbf{x}_t, t, \varnothing)$ denote the semantic error signal. Recall that the sliding variable is defined as $\mathbf{s}(t) = \dot{\mathbf{e}}(t) + \lambda \mathbf{e}(t)$. Substituting the closed-loop update law into the time derivative of the sliding variable, we obtain the governing equation:

$$\dot{\mathbf{s}}(t) = \mathbf{\Phi}(t, \mathbf{x}_t) + \mathbf{\Gamma}(t) \cdot \Delta\mathbf{e}(t), \tag{8}$$

where:

- $\mathbf{\Phi}(t, \mathbf{x}_t)$ represents the *intrinsic drift dynamics*, encapsulating the system's natural evolution and standard CFG terms.
- $\mathbf{\Gamma}(t)$ denotes the *effective control gain matrix*, which corresponds to the scaled Jacobian of the semantic difference: $\mathbf{\Gamma}(t) = w\nabla_{\mathbf{x}}(\mathbf{v}_\theta(\mathbf{c}) - \mathbf{v}_\theta(\varnothing))$.

A key challenge in diffusion models is that $\mathbf{\Gamma}(t)$ is highly non-linear and anisotropic. To address this, we adopt a robust control strategy by decomposing the gain into a nominal part and a deviation part.

1.3.2. Robustness Assumptions

Assumption 1 (Boundedness of Intrinsic Drift)**.** *While the gradients of diffusion models may diverge at time boundaries ($t \to 0$ or $t \to T$), we assume that within the effective sampling interval, the drift term $\mathbf{\Phi}(t, \mathbf{x}_t)$ is locally bounded:*

$$\sup_{t,\mathbf{x}\in\mathcal{D}} \|\mathbf{\Phi}(t,\mathbf{x})\|_2 \leq \delta. \tag{9}$$

Assumption 2 (Nominal Control Dominance)**.** *We decompose the effective gain matrix $\mathbf{\Gamma}(t)$ into a nominal isotropic gain $w\mathbf{I}$ and an anisotropic deviation $\Delta\mathbf{\Gamma}(t)$:*

$$\mathbf{\Gamma}(t) = w\mathbf{I} + \Delta\mathbf{\Gamma}(t). \tag{10}$$

We assume that the guidance scale w is sufficiently large such that the nominal control direction dominates the anisotropic deviation, in the sense that there exists a constant $\rho > 0$ with $w > \rho\sqrt{D}$ where the constant D is the dimension of $\mathbf{s}$. And the spectral norm of the deviation is bounded:

$$\|\Delta\mathbf{\Gamma}(t)\|_2 \leq \rho. \tag{11}$$

Remark: Assumption 2 is physically intuitive: it implies that the CFG guidance force w remains the dominant driver of the semantic correction, while the local curvature of the velocity field $\Delta\mathbf{\Gamma}$ acts as a subordinate disturbance.

1.3.3. Robust Stability Analysis

We now show that the proposed switching control law $\Delta\mathbf{e}(t) = -k \cdot \text{sign}(\mathbf{s}(t))$ ensures stability despite these uncertainties.

Theorem 1 (Robust Convergence)**.** *Consider the system in Eq. (8) under Assumptions 1 and 2. If the switching gain k satisfies:*

$$k > \frac{\delta}{w - \rho\sqrt{D}} + \epsilon, \tag{12}$$

where $\epsilon > 0$ is a safety margin.

Consider the Lyapunov function $V(\mathbf{s}) = \frac{1}{2}\|\mathbf{s}\|_2^2$. Its derivative is:

$$\dot{V} = \mathbf{s}^\top\dot{\mathbf{s}} = \mathbf{s}^\top\left(\mathbf{\Phi} + (w\mathbf{I} + \Delta\mathbf{\Gamma})\Delta\mathbf{e}\right). \tag{13}$$

Substituting the control law $\Delta\mathbf{e} = -k \cdot \text{sign}(\mathbf{s}(t))$ (for $\mathbf{s} \neq \mathbf{0}$):

$$\begin{aligned}\dot{V} &= \mathbf{s}^\top\mathbf{\Phi} - wk\|\mathbf{s}\|_1 - k\mathbf{s}^\top\Delta\mathbf{\Gamma} \cdot \text{sign}(\mathbf{s}(t)) \\ &\leq \|\mathbf{s}\|_2\|\mathbf{\Phi}\|_2 - wk\|\mathbf{s}\|_1 + k\|\mathbf{s}\|_2\|\text{sign}(\mathbf{s}(t))\|_2\|\Delta\mathbf{\Gamma}\|_2 \\ &\leq \delta\|\mathbf{s}\|_2 - wk\|\mathbf{s}\|_1 + k\rho\sqrt{D}\,\|\mathbf{s}\|_2,\end{aligned} \tag{14}$$

Let $\phi = \omega - \rho\sqrt{D}$ and apply the bounds δ and ρ from Assumptions 1 and 2:

$$\dot{V} \leq \|\mathbf{s}\|_2\left(\delta - wk + k\rho\sqrt{D}\right) = \|\mathbf{s}\|_2\left(\delta - k\phi\right). \tag{15}$$

From the condition in Eq. (12), we have $k\phi > \delta + \epsilon\phi$. Substituting this into the inequality:

$$\dot{V} \leq \|\mathbf{s}\|_2\left(\delta - (\delta + \epsilon\phi)\right) = -\epsilon\phi\|\mathbf{s}\|_2. \tag{16}$$

Let $\eta = \epsilon\phi > 0$. The differential inequality $\dot{V} \leq -\sqrt{2}\eta V^{1/2}$ guarantees finite-time convergence of $\mathbf{s}(t)$.

This analysis demonstrates that SMC-CFG is theoretically robust: as long as the gain k is chosen to cover the worst-case combination of intrinsic drift δ and the dimension-amplified Jacobian mismatch $\rho\sqrt{D}$ induced by the sign-based switching law, the system remains stable.

1.3.4. Discrete Implementation and Stability Corridor

The theoretical derivation above serves as a continuous-time design guide. In practice, diffusion models operate in discrete time steps Δt, where high-gain switching can lead to chattering. Based on the discrete evolution $\|\mathbf{s}_{t+1}\| \approx |\|\mathbf{s}_t\| - \Delta t(w_{eff}k - \delta)|$, we derive a heuristic **Stability Corridor** for hyperparameter tuning:

$$\underbrace{\frac{\delta_{est}}{w}}_{\text{Convergence}} < k < \underbrace{\frac{2\|\mathbf{s}_t\|_2}{w\Delta t}}_{\text{Stability}}. \tag{17}$$

This corridor highlights the trade-off: k must be large enough to overcome model drift (lower bound), but bounded by the inverse step size to prevent numerical oscillations (upper bound). This aligns with our experimental findings in Table 3, where a moderate fixed k achieves the optimal balance.

Remark on Hyperparameter Selection. The theoretical analysis in Eq. (17) establishes a stability corridor for the gain k, bounded by the intrinsic model drift δ (lower bound) and the discretization frequency $1/\Delta t$ (upper bound). In practice, while the exact value of δ varies across timesteps and samples, it is inherently bounded by the Lipschitz continuity of the pre-trained network. Furthermore, the upper bound is typically dominated by the inverse step size term, creating a wide margin for feasible k. Consequently, we treat k as a scalar hyperparameter. Our ablation studies (Table 3 in the main paper) empirically verify this theoretical corridor: excessively low k fails to overcome model drift (under-correction), while excessively high k induces numerical chattering (over-correction). A fixed intermediate value provides robust performance across diverse inputs without requiring real-time estimation of δ.

2. Additional Implementation Details

2.1. Datasets and Baselines.

To comprehensively evaluate the proposed SMC-CFG, we compare it with the standard CFG on the image-generation benchmark T2I-CompBench [5] using three different flow matching models. T2I-CompBench is a comprehensive

benchmark for open-world compositional text-to-image generation, comprising 6,000 compositional text prompts. In our experiments, we focus on four sub-categories that are most relevant to text-aligned image fidelity: color binding, shape binding, texture binding, and spatial relationships. For all flow matching models, we adopt publicly available checkpoints from HuggingFace. Specifically, Stable Diffusion 3.5 is based on the "stabilityai/stable-diffusion-3.5-large" public weights. Flux-dev uses "black-forest-labs/FLUX.1-dev". Given that Flux-dev is a guidance-distilled model, we set the embedded guidance to 1 in baseline experiments to ensure fairness when no CFG is applied. Qwen-Image uses the "Qwen/Qwen-Image" checkpoint. All models generate images at a resolution of 1024 $\times$ 1024 from textual prompts without any additional fine-tuning.

2.2. Metrics.

In the main text, we utilize a series of evaluation metrics. FID (Fréchet Inception Distance) [4] computes the Fréchet Distance between the multivariate Gaussian distribution estimated from the feature vectors of generated images and of real images, assessing the image quality and diversity of the generated results. CLIP Score [3], utilizing a pre-trained CLIP [9] model, quantifies the semantic alignment between the generated image and the text prompt by computing the cosine similarity between their respective L2-normalized feature vectors. Aesthetic Score [12] serves as an aesthetic regression model, evaluating the image's general aesthetic appeal, such as excellent composition and harmonious coloring. ImageReward [16] is a general-purpose reward model trained on a large dataset of expert human preference feedback, which quantifies the generated image's perceived quality and attractiveness to predict the probability of being preferred by humans. PickScore [6] is a CLIP-based scoring function trained on real users' preference data, specifically designed to predict the probability of a generated image being selected by humans in a competitive setting. HPSv2 and HPSv2.1 (Human Preference Score) [15] are multi-dimensional perceptual metrics that simultaneously assess the image adherence to the prompt, aesthetic quality, and visual fidelity. Finally, MPS (Multi-dimensional Preference Score) [17] is a unified model that utilizes a condition mask on top of the CLIP model to predict the quality of a text-to-image output across four distinct human preference dimensions: Overall, Aesthetics, Semantic Alignment, and Detail Quality.

2.3. Hyperparameters.

We determine the hyperparameters of SMC-CFG through grid search over the two parameters λ and k. Specifically, λ is searched within $\{2, 3, 4, 5, 6, 7, 8\}$, while k is explored over $\{0.01, 0.05, 0.1, 0.15, ..., 0.75, 0.8\}$. To avoid test-set leakage, the grid search is conducted on an auxiliary set of 200 cases sampled from the MS-COCO [8] dataset, which is entirely disjoint from the evaluation set used in the experiment. The optimal configurations selected for the three text-to-image models used in our experiments are as follows: for Stable Diffusion 3.5, λ = 6 and $k = 0.1$; for Flux, $\lambda = 6$ and $k = 0.7$; and for Qwen-Image, $\lambda = 6$ and $k = 0.1$. The main experiments adopt these hyperparameter settings without further modification.

Table 2. **Quantitative evaluation on T2I-CompBench.**

Model	Color ↑	Shape ↑	Texture ↑	Spatial ↑
SD3.5 [1]	0.6790	0.5915	0.7243	0.1625
w/ SMC-CFG	**0.7461**	**0.6009**	**0.7406**	**0.2563**
Flux-dev [7]	0.8172	0.5751	0.7432	0.2708
w/ SMC-CFG	**0.8216**	**0.6199**	**0.7901**	**0.2939**
Qwen-Image [14]	0.7747	0.5621	0.6747	0.2968
w/ SMC-CFG	**0.8191**	**0.5934**	**0.7421**	**0.4085**

3. More Experiments

3.1. Text-to-Image Benchmark Evaluation

We evaluate SMC-CFG on three flow matching text-to-image models using T2I-CompBench [5], and compare it with representative CFG-based baselines on VQAScore (GenAI-Bench). Table 2 shows that SMC-CFG improves the compositional generation performance of SD3.5, Flux-dev, and Qwen-Image on Color, Shape, Texture, and Spatial. The gains are generally larger on spatial and attribute-related dimensions. Table 3 reports the VQAScore results on SD3.5. SMC-CFG achieves the best Base, Advance, and Overall scores among the compared methods, outperforming standard CFG as well as recent variants such as CFG-Zero and Rect-CFG++. Visual comparisons on the three T2I models are shown in Figure 3, 4, and 5.

Table 3. **Compositional alignment evaluation on SD3.5.**

Method	VQAScore (GenAI-Bench)		
	Base ↑	Advance ↑	Overall ↑
Base (w/o CFG)	0.79	0.64	0.70
w/ CFG	0.83	0.64	0.72
w/ CFG-Zero*	0.88	0.66	0.75
w/ Rect-CFG++	0.87	0.64	0.73
w/ SMC-CFG	**0.89**	**0.68**	**0.77**

3.2. Text-to-Video Generation

We further extend our evaluation to the text-to-video generation task to assess the generalization capability of SMC-CFG. Using the Wan2.2-TI2V-5B [13] model, we conduct

Table 4. **Video comparison on Wan2.2-TI2V-5B.**

Method	Total Score	Quality Score	Semantic Score	Color	Human Action	Subject Consistency
CFG	0.5594	0.6581	0.4607	0.9087	0.5313	0.9450
SMC-CFG	**0.5839**	**0.6747**	**0.4931**	**0.9818**	**0.6000**	**0.9609**

a qualitative comparison against the standard CFG baseline. As visualized in Figure 6, our method demonstrates superior stability in the spatiotemporal domain. Specifically, SMC-CFG enhances temporal consistency, producing smoother motion trajectories with fewer visual artifacts or flickering compared to the baseline. Furthermore, it exhibits robust semantic adherence in complex compositional scenarios, effectively maintaining the spatial structure and identity of generated objects throughout the video sequence. We also show quantitative evaluation in Table 4. SMC-CFG improves the total VBench score and gives higher Quality and Semantic scores than CFG. It also performs better on Color, Human Action, and Subject Consistency. These results suggest that the behavior of SMC-CFG is not limited to text-to-image generation and can transfer to text-to-video generation as well.

3.3. Computational Efficiency

We further assess the computational overhead and inference latency of our method at different output resolutions to demonstrate its practicality in real-world deployment scenarios. As presented in Table 5, SMC-CFG exhibits memory consumption and FLOPs that are comparable to those of standard CFG in a single inference pass, and the average inference time remains nearly identical. These results indicate that SMC-CFG preserves the computational efficiency of standard CFG and does not introduce additional computational cost or latency during inference.

Table 5. Computational cost and inference time comparison of standard CFG and SMC-CFG.

Resolution	Guidance	Memory (GB)	FLOPs (G)	Runtime (s)
512×512	CFG	31.99	1203370.06	23.84
	SMC-CFG	31.99	1203370.07	23.97
1024×1024	CFG	33.59	3590870.89	44.78
	SMC-CFG	33.59	3590870.93	45.09

3.4. Ablation Study on Hyperparameter Effects

We conduct visual comparison with fixed initial noise to show impact of hyperparameters. As shown in Figure 1, λ governs the stability of structural details by shaping the sliding mode manifold, while k regulates the overall semantic alignment and its trade-off with aesthetic realism.

Figure 1. Qualitative results under various hyperparameters.

4. More Discussion

4.1. CFG Scale

We analyze the effect of the CFG scale by visualizing the performance curves of the main evaluation metrics under varying guidance strengths on the Flux-dev model, as shown in Figure 2. When the CFG scale reaches the model's default optimal value of 2, both standard CFG and other baseline methods achieve their best performance. However, their performance rapidly degrades as the scale increases further, revealing the strong nonlinear distortions introduced by high guidance. In contrast, SMC-CFG continues to improve as the CFG scale increases, demonstrating that it can better exploit the potential of guidance without suffering from the instability observed in conventional methods. Even under extremely large scales, SMC-CFG shows only a slight performance drop, indicating strong robustness against over-guidance effects.

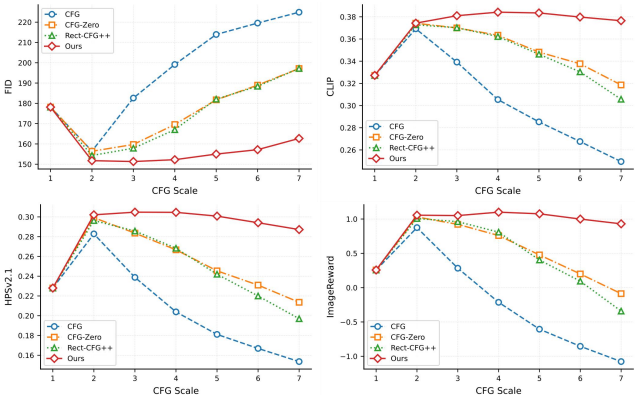
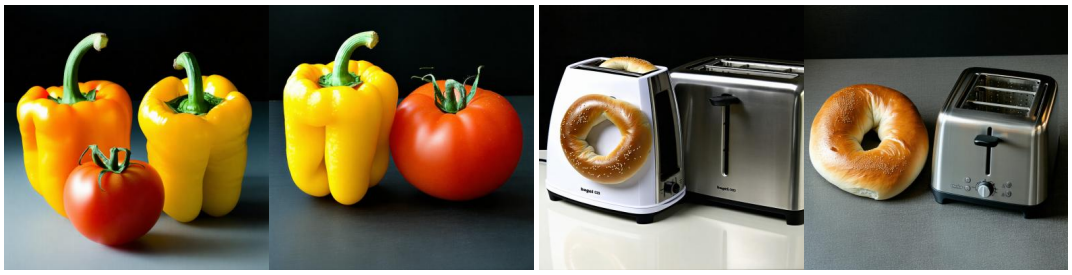
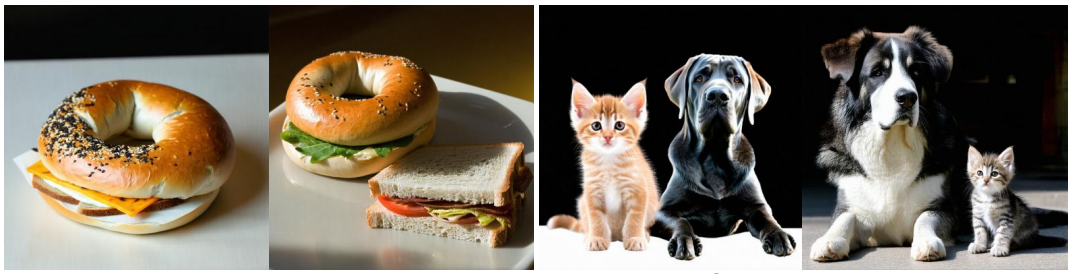


Figure 2. Performance curves of different methods under varying CFG scales.

4.2. Limitations and Future Work

Despite its ability to alleviate the nonlinear effects associated with high CFG scales and to substantially improve compositional image generation, SMC-CFG introduces two additional hyperparameters, which increase the complexity of deployment and may require manual tuning for different models. In the future, we plan to explore adaptive guidance control mechanisms capable of dynamically adjusting control parameters according to the evolving state of the generative process. In particular, one promising way is to incorporate error-differential feedback, where changes in text–image alignment across successive steps are used to automatically increase or decrease the effective guidance strength. The adaptive strategy offers the potential to eliminate manual tuning while improving stability and performance under varying guidance scales.





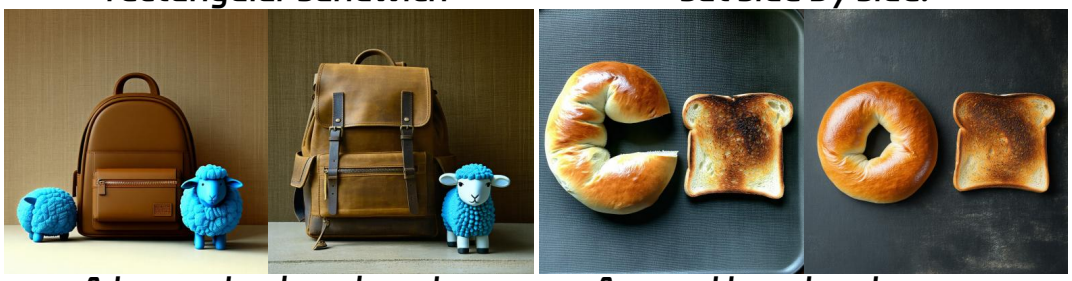


Figure 3. Addtional visual comparison between CFG (left) and SMC-CFG (right) in SD3.5.

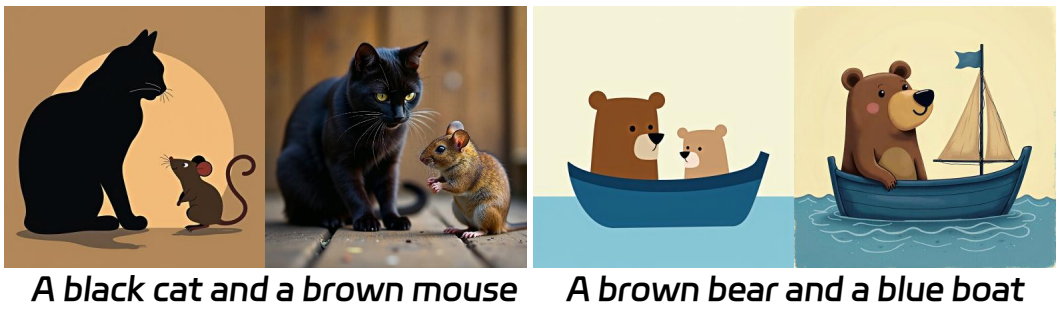


Figure 4. Addtional visual comparison between CFG (left) and SMC-CFG (right) in Flux-dev.

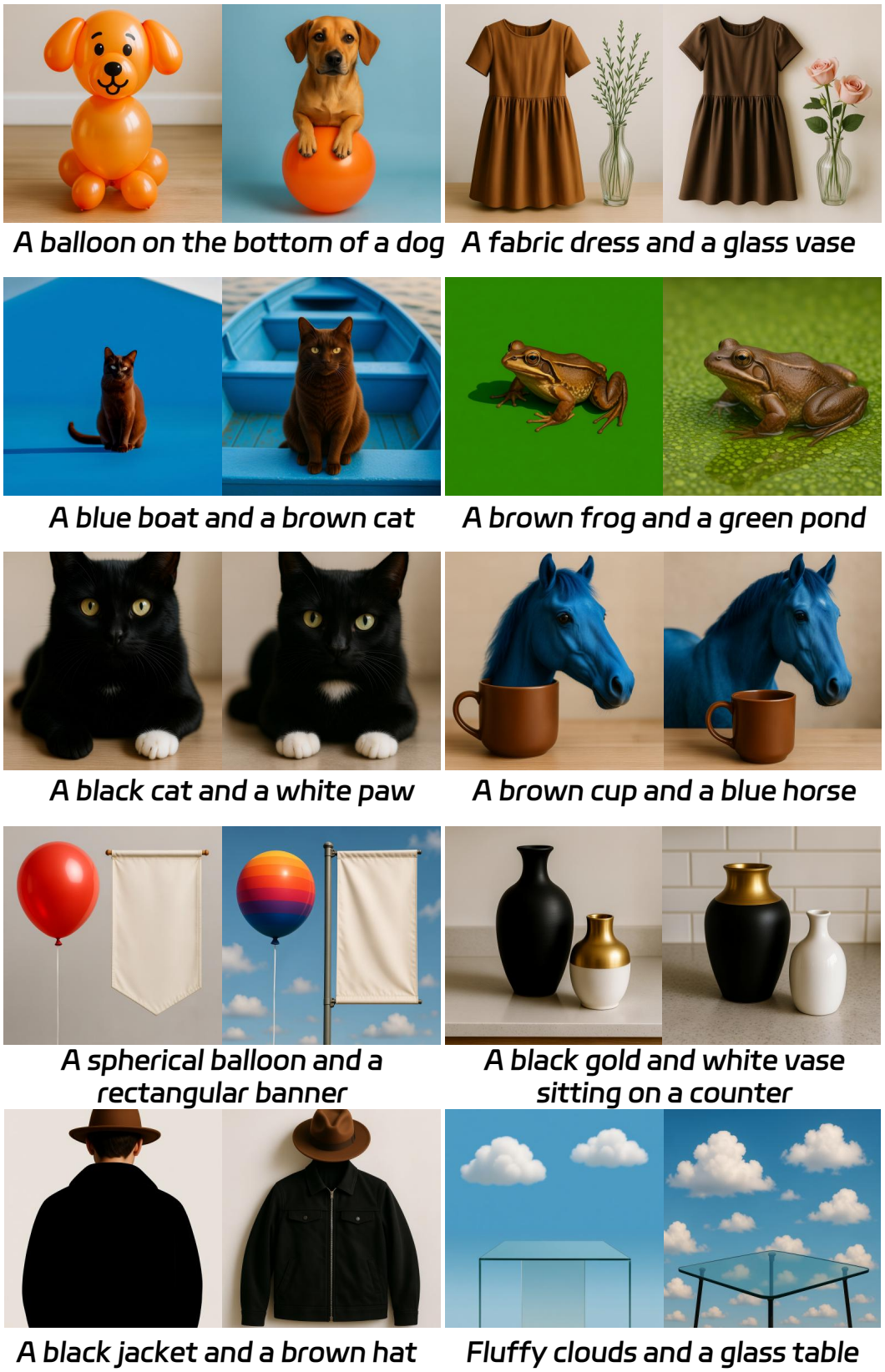


Figure 5. Addtional visual comparison between CFG (left) and SMC-CFG (right) in Qwen-Image.

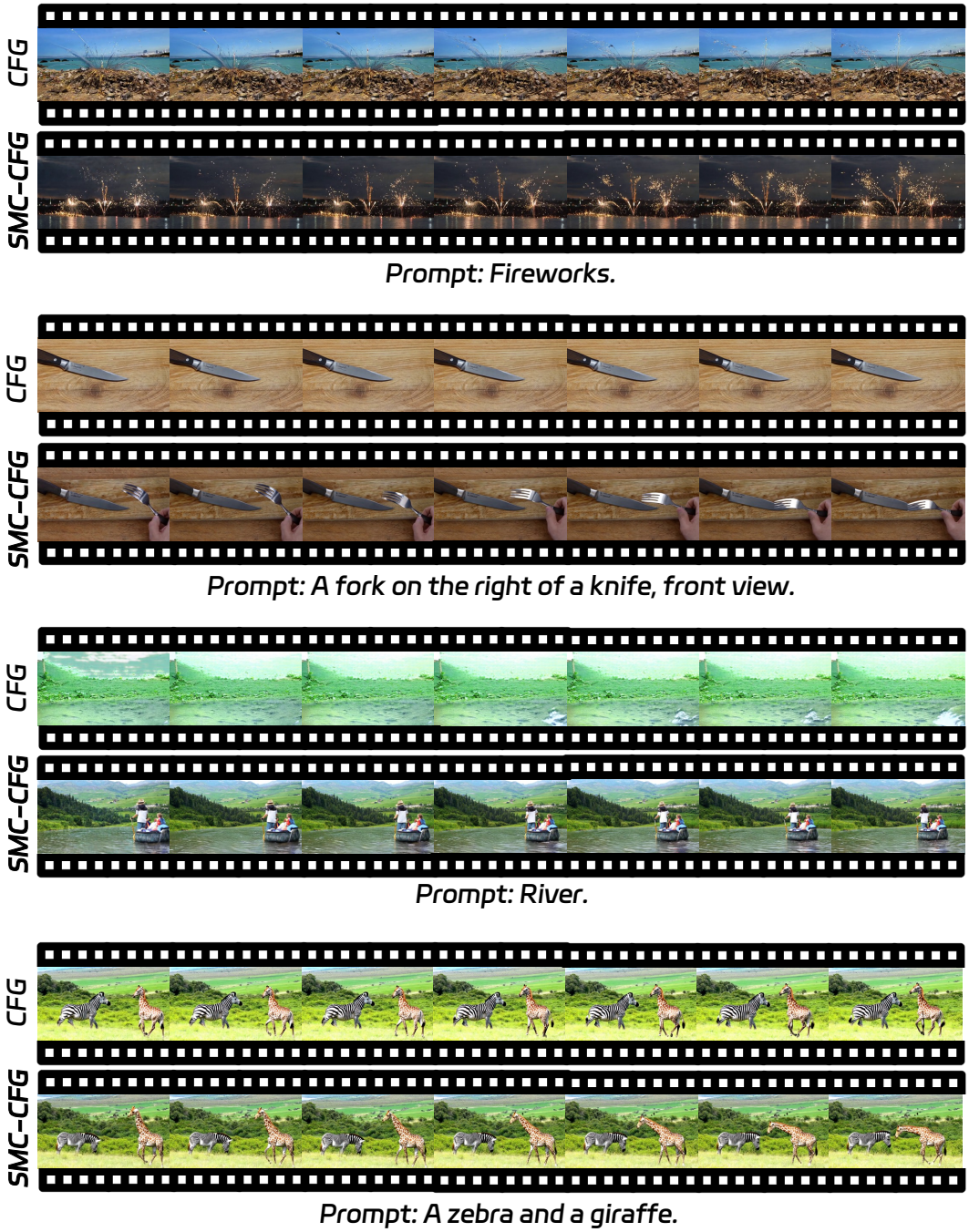


Figure 6. Additional video comparisons between CFG (above) and SMC-CFG (below) in Wan2.2-TI2V-5B.

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4

[2] Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025. 1

[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021. 4

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[5] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 3, 4

[6] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 4

[7] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 4

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[10] 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. 1

[11] Shreshth Saini, Shashank Gupta, and Alan C Bovik. Rectified-cfg++ for flow based models. *arXiv preprint arXiv:2510.07631*, 2025. 1

[12] Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2023-11-10. 4

[13] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

[14] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 4

[15] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 4

[16] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 4

[17] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024. 4