SCFLOW: SPECTRUM-AWARE CORRECTION FOR STABILIZING TRAINING-FREE SPATIAL CONTROL IN RECTIFIED FLOW MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rectified flow models (RFMs) have emerged as an alternative to score-based diffusion models (SDMs). Although training-free spatial-condition generation was successful in SDMs, so far, there has been no systematic study on RFMs. In this paper, we pinpoint that prior methods do not perform as well on RFMs. We focus on energy-guided sampling as a representative paradigm, and show that RFMs have a highly anisotropic latent-feature Jacobian compared to SDMs at the early diffusion stage, which can unevenly emphasize or neglect different components of the control signal. We then develop Spectral Correction Flow (ScFlow), a training-free spatial control framework targeting RFMs. ScFlow consists of two correction components, channel-wise whitening and prompt reweighting. Compared to the uncorrected baseline, ScFlow improves DINO self-similarity by 14.1% and CLIP score by 11.8% on average across evaluated RFMs. Extensive experiments demonstrate that ScFlow provides better controllability on RFMs and outperforms state-of-the-art training-free methods.

1 INTRODUCTION

Recent text-to-image (T2I) generative models (Rombach et al., 2022; Podell et al., 2024) can synthesize high-quality images from natural language prompts alone. However, text is often insufficient for specifying precise spatial arrangements that faithfully reflect user intent. This limitation has motivated a large body of work (Li et al., 2024; Ye et al., 2023; Li et al., 2023; Huang et al., 2023; Zhao et al., 2023) on controllable generation, where auxiliary conditions such as sketches or depth maps are introduced to regulate spatial structure. Existing approaches, exemplified by ControlNet (Zhang et al., 2023), have achieved impressive controllability, but typically require training on large-scale paired data. Training-free approaches have been proposed to avoid this training overhead. Existing training-free methods can be broadly divided into two groups. Injection-based approaches (Hertz et al., 2022; Tumanyan et al., 2023; Meng et al., 2022) first use an inversion (Song et al., 2021; Mokady et al., 2023; Lin et al., 2024; Kim et al., 2025b) or forward diffusion process (i.e., noising) to obtain features from the condition image, such as noisy latents or feature maps, and then reuse them during generation. Another line of work (Epstein et al., 2023; Yu et al., 2023; Sun et al., 2024; Bansal et al., 2024), which is referred to as energy-guided sampling in this paper, instead progressively adjusts the diffusion latents by optimizing a designed energy function during the sampling process. Although these two paradigms intervene differently, they both operate on the intermediate states of diffusion sampling, making their effect closely coupled to the underlying generative process. However, prior training-free spatial-control methods have been evaluated exclusively on score-based diffusion models (SDMs), such as SD1.5 and SDXL (Rombach et al., 2022; Podell et al., 2024). It remains unclear how reliably established mechanisms transfer to recent rectified flow models (RFMs) (Liu et al., 2023).

To study this transition while disentangling the generative dynamics of RFMs from architectural confounds, we first evaluate state-of-the-art spatial control methods on InstaFlow (Liu et al., 2024), a UNet-based RFM. We find that this transfer is non-trivial. Injection-based methods exhibit noticeable performance variations (Table 1), while energy-guided sampling also becomes unstable. Figure 1 illustrates this phenomenon: FreeControl operates reliably on SD1.5, while producing substantial degradation when applied to InstaFlow. This motivates us to investigate what underlying

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.

differences between SDMs and RFMs cause this performance difference. We focus on energy-guided sampling since its explicit optimization objective makes the control pathway directly analyzable. Energy-guided sampling introduces spatial control through the gradient of an energy function, which gradually reduces the gap between the features of the condition input and the generated output. Taking a closer look at the formulation, we notice that the guidance gradient is mediated by the latent-feature Jacobian matrices of the generative model. Our experiment reveals that the spectra of these matrices in RFMs at the early sampling stage are highly anisotropic, while the spectra of those in SDMs are closer to isotropic. Such anisotropy can over-amplify some control signals while neglecting others, suggesting a contributing factor to the observed task instability. Additionally, our analysis finds that text conditioning also influences this anisotropy of RFMs.

Motivated by these findings, we propose Spectral Correction Flow (ScFlow), a training-free spatial control framework of stabilized energy-guided sampling on RFMs. ScFlow stabilizes energy-guided sampling on RFMs with two lightweight inference-time corrections: channel-wise whitening and prompt reweighting. Extensive experiments show that ScFlow restores spatial controllability on RFMs and outperforms state-of-the-art training-free spatial control methods. Our contributions are:

(1) We perform a series of analyses on RFMs and SDMs to identify their performance differences in spatial control generation. We find that the spectral anisotropy of RFMs is a contributing factor to the observed task failure. We also show that such spectral anisotropy is influenced by text conditions. (2) We propose ScFlow, the training-free spatial control framework of stabilized energy-guided sampling with two novel inference-time spectral correction methods: channel-wise whitening and prompt reweighting. (3) We validate ScFlow with extensive experiments over different condition types and model backbones, demonstrating consistent improvements over the uncorrected baseline and superior performance over other training-free methods on RFMs.

2 Related Work

Rectified Flow Models. Rectified flow models (Liu et al., 2023; 2024; Esser et al., 2024; Labs et al., 2025) learn a velocity field that transports noise to data along near-straight trajectories, enabling high-quality generation with fewer sampling steps. Recent studies (Wang et al., 2025) suggest that rectification, i.e., the reflow method (Liu et al., 2023), is central to this paradigm and can be applied to non-flow-matching models. In this work, we use the term RFM in this broader rectified-model (Wang et al., 2025) sense. Recent work has begun to extend training-free methods from SDMs to RFMs for image-to-image translation and editing (Wei et al., 2025; Shen et al., 2025; Kim et al., 2025a; Shin et al., 2025; Rout et al., 2024), where the condition inputs are RGB images and provide rich visual details. In contrast, conditional generation from sparse spatial conditions remains less explored. RF-Inversion (Rout et al., 2024) has reported results on a stroke-to-generation task, but has not explored general spatial control. Moreover, many of these work primarily focuses on the implications of transformer architectures. Several studies (Sun et al., 2024; Patel et al., 2025) begin to discuss the implications of rectified generative dynamics on tasks such as personalized generation or inverse problems, but their task focus differs from ours.

Diffusion Spatial Control. Spatial control methods can be broadly divided into training-based and training-free ones. Training-based methods (Zhang et al., 2023; Li et al., 2024; Zhao et al., 2023) learn additional conditioning modules and achieve strong performance, but require substantial training data for each novel condition type. Training-free methods (Tumanyan et al., 2023; Hertz et al., 2022; Meng et al., 2022; Xie et al., 2023; Couairon et al., 2023; Chen et al., 2024; Mo et al., 2024; Lin et al., 2024; Kim et al., 2025b) avoid this cost by manipulating the generative process at inference time. Early methods mainly target specific input conditions, such as bounding boxes (Xie et al., 2023) or segmentation maps (Couairon et al., 2023), while later work aims to support general spatial conditions in a unified framework (Tumanyan et al., 2023; Mo et al., 2024; Lin et al., 2024; Kim et al., 2025b). A common strategy is to obtain intermediate features from the condition image, either through forward noising or inversion (Song et al., 2021; Mokady et al., 2023), and reuse these latents or feature maps during generation (Tumanyan et al., 2023; Hertz et al., 2022; Meng et al., 2022; Lin et al., 2024; Kim et al., 2025b; Mo et al., 2024). Injection-based methods directly insert these representations into the denoising process (Tumanyan et al., 2023; Hertz et al., 2022; Meng et al., 2022; Lin et al., 2024; Kim et al., 2025b). Another branch of work (Epstein et al., 2023; Yu et al., 2023; Mo et al., 2024; Sun et al., 2024) takes a different route by explicitly formulating control

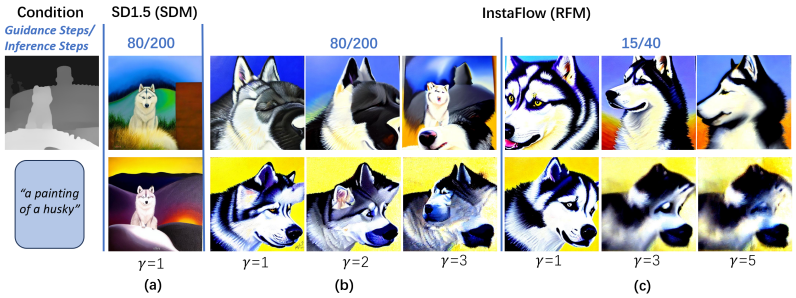

Figure 1: (a) With guidance weight $\gamma = 1$, energy-guided sampling is stable on SD1.5 (SDM) but fails on InstaFlow (RFM) under the same setting. (b) On InstaFlow, increasing γ degrades texture quality, e.g., causing artifacts, without realizing the target structure. We refer to this behavior as **instability**. (c) Instability is more severe with fewer inference steps.

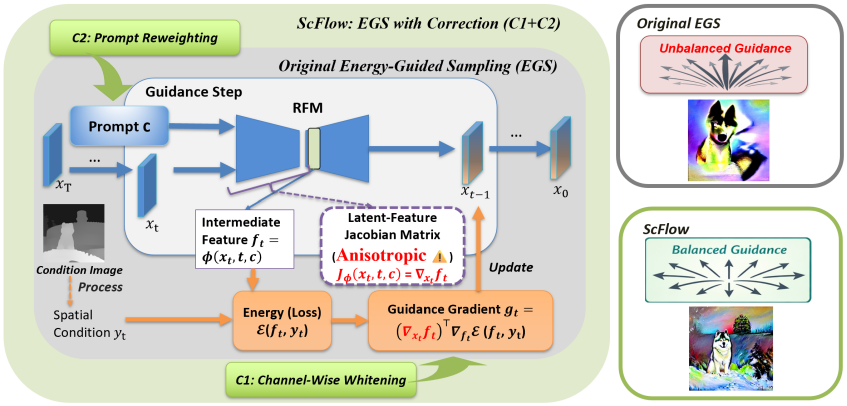

Figure 2: **Method overview.** Energy-guided sampling (EGS) updates each step of diffusion sampling with a gradient. In RFMs, the gradient is governed by the transpose of an anisotropic latent-feature Jacobian, which can unevenly transform the control signals. **ScFlow** adds two correction components, **channel-wise whitening** and **prompt reweighting**, to EGS and achieves more stable guidance.

as inference-time optimization on diffusion latents, allowing control objectives to be optimized as energy functions. However, both lines of existing methods are inherently mediated by the underlying generative process, while evaluated exclusively on SDMs. Their behavior on RFMs remains underexplored, motivating our study.

3 Preliminary

Score-based diffusion models (Ho et al., 2020; Song et al., 2021) and flow models (Lipman et al., 2023) generate images by transporting an initial Gaussian noise sample x_T along a sampling path conditioned on a text prompt c. Let $x_t \in \mathbb{R}^d$ denote the latent at timestep t. The sampling process can be described by

$$x_{t-1} = x_t + \Delta t \, u_\theta(x_t, t, c), \tag{1}$$

where $t \in \{0, 1, \cdots, T\}$, Δt is the sampling step size, and u_θ is the backbone of the T2I model. x_0 is the generated latent, which is eventually decoded into an image.

Energy-guided sampling (Mo et al., 2024) is developed to control the spatial structure of generated images. Figure 2 illustrates the process of energy-guided sampling. Let I_s be an image, e.g., a depth map, used as a spatial condition. Energy-guided sampling first uses an inversion scheme, e.g., DDIM inversion (Song et al., 2021), to determine the corresponding noise of I_s, denoted as s_T, and then applies the sampling process, i.e., Equation 1, to regenerate the condition. The intermediate features

y_t at timestep t, which are commonly extracted from the self-attention layer of the backbone model, are then used for guidance.

Energy-guided sampling measures the distance between the spatial condition y_t and the intermediate feature of u_θ through an energy function $\mathcal{E}$, whose gradient is used to control the sampling. It should be highlighted that the intermediate feature of u_θ is obtained from x_t and the prompt c, not from the sampling path of s_T, the noise inverted from I_s. To apply the guidance, the sampling process in Equation 1 is modified as

$$x_{t-1} = x_t + \Delta t\, u_\theta(x_t, t, c) - \gamma g_t, \tag{2}$$

where $g_t = \nabla_{x_t}\mathcal{E}(\phi(x_t, t, c), y_t)$ and $\phi(x_t, t, c)$ denotes the intermediate features extracted from u_θ. Previous works have demonstrated that energy-guided sampling is effective on SDMs, such as SD1.5. However, we find that it is unstable on RFMs.

4 A Spectral Analysis of the Latent-Feature Jacobian

Figure 1a shows that energy-guided sampling applied to SD1.5 can provide stable and high-quality outputs, but is unstable on RFMs. Even when we attempt to stabilize the generation by adjusting the weight γ, the generation quality is still far from that of SD1.5 (Figure 1b). Since the condition images, the sampling process (Equation 1), and energy-guided sampling (Equation 2) are the same, the instability should come from some inherent properties of RFMs.

4.1 The Influence of Spectral Anisotropy and its Measurement

Let us take a closer look at g_t, which can be rewritten as $g_t = J_\phi(x_t, t, c)^T w$, where $w = \nabla_\phi \mathcal{E}(\phi(x_t, t, c), y_t)$ and $J_\phi(x_t, t, c) = \nabla_{x_t}\phi(x_t, t, c)$. Since w is derived from the same energy function, the latent-feature Jacobian $J_\phi(x_t, t, c)$ better reflects the model-level difference between SDMs and RFMs. To study the latent-feature Jacobian matrix, we use SVD to decompose it as $J_\phi(x_t, t, c) = U\Sigma V^T$, where $\Sigma = \mathrm{diag}(\sigma_i)$ is the singular value matrix, and g_t can be written as

$$g_t = \sum_{j=1}^{r} \sigma_j (u_j^T w) v_j, \tag{3}$$

where $r = \min(m, n)$, m and n are the dimensions of the Jacobian matrix, and $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r \geq 0$. Since U and V are unitary matrices, the norms of their column vectors, u_j and v_j, are one. However, the values of σ_j can vary greatly. In an ideal isotropic case, all the singular values are the same, and therefore all v_j and $(u_j^T w)$ can contribute to g_t equally. In an extreme anisotropic case, a few singular values are significantly larger than the others, and g_t can be approximated by the dominative singular values and the corresponding components, i.e.,

$$g_t \approx \sum_{j=1}^{q} \sigma_j (u_j^T w) v_j, \tag{4}$$

where $\sigma_1, \cdots, \sigma_q$ are the dominative singular values and $q \ll r$. In this extreme case, the spatial control signals in the directions of $v_{q+1}, \cdots, v_r$ will be neglected. In the most extreme case, σ_1 entirely dominates the spectrum, resulting in $g_t \approx \sigma_1 (u_1^T w) v_1$. All spatial control signals, except those in the v_1 direction, will be neglected.

To compare the spectra of the latent-feature Jacobians of SDMs and RFMs, a straightforward approach is to use SVD to compute their singular values. However, the time complexity of SVD is $O(mn^2)$ (Xu et al., 2023), and the dimensions of the Jacobian matrices of state-of-the-art generative models are extremely large. In our study, the dimensions of the Jacobian matrices of SD1.5 and InstaFlow are both 16,384 by 327,680, which prohibit applying this straightforward approach. Nevertheless, there is an effective algorithm to compute the square of the largest singular value, σ_1^2, and $\sum_{j=1}^{r} \sigma_j^2$. We use their ratio

$$S(J_\phi) = \frac{\sigma_1^2}{\sum_{j=1}^{r} \sigma_j^2}, \tag{5}$$

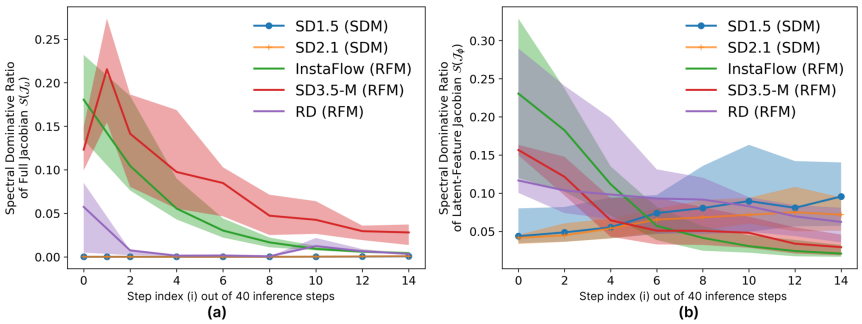

Figure 3: We compare the spectral dominative ratio of (a) the full model Jacobian, $S(J_u)$, and (b) the latent-feature Jacobian, $S(J_\phi)$. The shaded region indicates the 25–75 percentile range, and the line represents the medians. Both plots consistently show that RFMs exhibit higher anisotropy than SDMs at the early diffusion stage.

as a measure of spectral anisotropy. σ_1^2 can be computed through the power iteration algorithm (Trefethen & Bau, 1997), whose time complexity is $O(kmn)$, where k is the number of iterations, and $\sum_{j=1}^{r} \sigma_j^2 = \mathrm{tr}(J_\phi^T J_\phi)$. When $S(J_\phi)$ is close to one, the Jacobian matrix is highly anisotropic. When $S(J_\phi)$ is close to $1/r$, it is close to isotropic. We refer to our anisotropy index $S(J_\phi)$ as the spectral dominative ratio hereafter.

4.2 The Spectrum Difference

To study the spectral differences among different models, we compute the spectral dominative ratios of SD1.5, SD2.1, InstaFlow, the medium variant of SD3.5, and rectified diffusion (RD). The former two are SDMs, and the latter three are RFMs. Since $g_t = J_\phi(x_t, t, c)^T w$, where J_ϕ is determined by the generative model $u_\theta(x_t, t, c)$ and w carries the spatial control signal, we investigate the spectra of J_ϕ without control signal w. To prevent influence from prompts, we use an empty prompt, i.e., "", to generate 100 images and compute their $S(J_\phi)$ at different timesteps. In addition to J_ϕ, which is derived from the feature $\phi(x_t, t, c)$, we investigate the spectrum of the full generative model u_θ in Equation 1, whose Jacobian is given by $J_u(x_t, t, c) = \nabla_{x_t} u_\theta(x_t, t, c)$. J_u is the Jacobian matrix of the full generative model, and J_ϕ is the input-feature Jacobian matrix.

Figure 3a shows the $S(J_u)$ of the five models at different timesteps. Clearly, $S(J_u)$ of SD1.5 and SD2.1 are very small. Their means at timestep 2 are 6.1×10^{-5} and 1.26×10^{-4}, indicating very isotropic spectra. However, $S(J_u)$ of all three RFMs are very large, especially in the first few timesteps. At timestep 2, the $S(J_u)$ of InstaFlow is 2000 and 1000 times higher than those of SD1.5 and SD2.1, the $S(J_u)$ of SD3.5 is 2500 and 1200 times higher, and the $S(J_u)$ of RD is 100 and 50 times higher. Figure 3a illustrates clearly that the full Jacobian matrices of InstaFlow, SD3.5 and RD are highly anisotropic. Although energy-guided sampling uses the latent-feature Jacobian matrices, J_ϕ, to achieve spatial control, Figure 3a demonstrates that the spectra of SDMs and RFMs are fundamentally different. Figure 3b shows $S(J_\phi)$ of the five models. As with Figure 3a, in the first few timesteps, InstaFlow, SD3.5, and RD have significantly higher $S(J_\phi)$ than SD1.5 and SD2.1. This indicates that the latent-feature Jacobian matrices of the three RFMs are highly anisotropic in the early timesteps. Although in the later timesteps, $S(J_\phi)$ of the five models becomes similar, and some $S(J_\phi)$ of SD1.5 and SD2.1 are even higher than those of RFMs, they do not affect spatial control too much, as the coarse spatial structure of an image is developed primarily in the early diffusion stage (Hertz et al., 2022).

Figure 3 shows that anisotropy is a consistent spectral property of RFMs and is present in both the full model Jacobian and the latent-feature Jacobian directly mediating energy-guided sampling. Together with Eqs. 3 and 4, these observations implicate the strong spectral anisotropy of RFMs as a plausible mechanism underlying the instability shown in Figure 1.

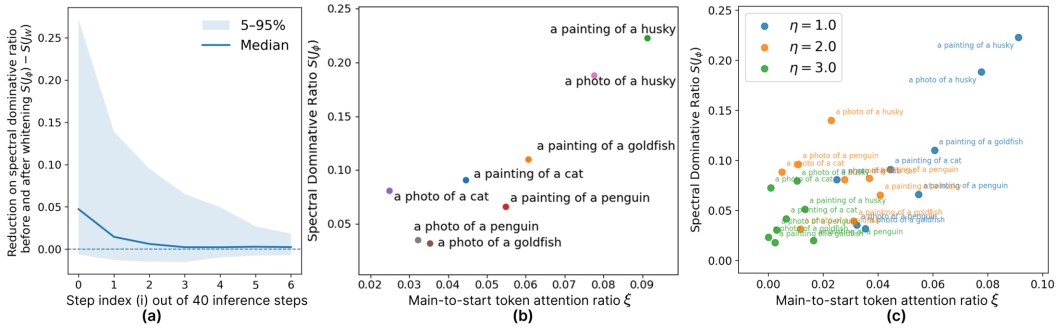

Figure 4: Channel-wise whitening (C1) and prompt reweighting (C2) effectively stabilize energy-guided sampling by reducing early-stage Jacobian anisotropy. (a) Reduction in the spectral dominative ratio after applying C1. (b) Main-to-start token attention ratio ξ versus $S(J_\phi)$ for unguided samples with varied prompts. A correspondence is observed between ξ and $S(J_\phi)$, motivating C2. (c) On the same samples as in (b), increasing the C2 scale η consistently reduces both ξ and $S(J_\phi)$.

4.3 The Relationship Between Spectral Anisotropy and the Prompts

In the previous analysis, we exposed the spectral differences between SDMs and RFMs using empty prompts to minimize prompt-specific effects. In practice, anisotropy also varies with text conditioning, i.e. prompts, as shown in Figure 4b: samples with different object words exhibit different spectral dominative ratios. For example, samples with the word "husky" often have higher $S(J_\phi)$ than others. In this analysis, we take a closer look at how prompts influence spectral anisotropy.

In T2I generation, a prompt c is first prepended with a special start symbol [SOT], appended with padding symbols, and mapped into a sequence of token embeddings E indexed by $j = 0, \ldots, n$, where $j = 0$ denotes the start token. We omit padding tokens in our analysis and refer to the remaining non-special tokens as content tokens, which carry object-level semantic information. In a cross-attention layer, latent queries attend to these token embeddings and produce an attention-weight map. Given a token with index j, we denote its attention weight averaged over query locations and attention heads as $\bar{p}_j$.

We choose the start token as a reference because it usually receives the highest average attention and is known to carry global information (Chen et al., 2024). We further define the main token as the content token with the largest average attention weight among all content tokens. For a prompt such as "a photo of a cat", this main token usually corresponds to the primary semantic word, e.g., "cat". We then define the main-to-start token attention ratio as $\xi = \frac{\bar{p}_m}{\bar{p}_0}$, where m denotes the index of the main token. A higher ξ indicates that the main token receives attention closer to the dominant start token, reflecting stronger dominance of a particular content token in cross-attention.

To study how this ratio influences the spectral anisotropy of RFMs, we perform a small-scale prompt analysis on InstaFlow. We combine two style words, e.g., painting, and four object words, e.g., penguin, to form a prompt template, "a style of an object". We then compute the average ξ for each prompt and their spectral dominative ratio $S(J_\phi)$ at an early timestep. Figure 4b clearly shows a general pattern: samples with a larger main-to-start token attention ratio ξ often exhibit a larger $S(J_\phi)$. This suggests that ξ has an implicit relationship with spectral anisotropy.

5 Spectral Correction Flow

The previous analysis has shown that 1) spectral anisotropy can cause the neglect of spatial control signals (Eqs. 3–4), 2) the spectra of RFMs are inherently anisotropic (Figure 3), and 3) the main-to-start token attention ratio of the prompt has an implicit relationship with spectral anisotropy. A natural way to address these problems and stabilize spatial control is to normalize the spectrum. More precisely, one can replace the singular value matrix Σ with an identity matrix such that the spectrum becomes isotropic and Equation 3 becomes $g_t = \sum_{j=1}^{r} (u_j^T w) v_j$, where each spatial control signal $(u_j^T w)$ can contribute to the generation. However, this straightforward approach, which

requires applying SVD decomposition to the high-dimensional latent-feature Jacobian matrix J_ϕ, is computationally infeasible, especially for practical applications. We instead propose Spectral Correction Flow (ScFlow), which consists of two lightweight components to reduce spectral anisotropy at inference time: channel-wise whitening and prompt reweighting.

5.1 Channel-Wise Whitening

Since directly computing the SVD is infeasible, we seek an alternative approach to ease spectral anisotropy. We observe that the latent-feature Jacobian J_ϕ exhibits a near-block-diagonal pattern (Appendix B.1). Inspired by classic SVD algorithms (Charlier & Van Dooren, 1987), which suggest that, for a strictly block-diagonal matrix, the singular values can be obtained independently from its diagonal blocks, we design a block-wise correction scheme. Moreover, the structural pattern suggests that high input-output sensitivity is concentrated in the channel-to-channel sub-matrix blocks of J_ϕ. This motivates us to consider whether the anisotropic amplification discussed in Sec. 4.1 appears as dominant channel directions in the guidance gradient. We therefore apply channel-wise whitening to g_t, which suppresses such dominant directions by flattening the spectrum of its channel-wise covariance.

For clarity, we reshape g_t into $G_t \in \mathbb{R}^{C\times N}$, where N is the number of spatial locations in the latent, C is the number of channel dimensions, and each column corresponds to a latent channel vector at one spatial location. The channel-wise covariance matrix is estimated as

$$R_t = \frac{1}{N}\hat{G}_t\hat{G}_t^T, \tag{6}$$

where $\hat{G}_t$ denotes the mean-centered version of G_t. The channel-wise whitening transform $W_t = R_t^{-1/2}$ is applied uniformly across spatial locations, giving $\tilde{G}_t = W_t\hat{G}_t$. Finally, $\tilde{G}_t$ is reshaped back to the dimensions of g_t and used to replace it during sampling as in Equation 2.

To evaluate the effect of channel-wise whitening on spectral anisotropy, we view the linear part of the entire whitening operation on g_t as a block-diagonal matrix W_B, with each block given by W_t. We then measure the spectral dominative ratio of the effective post-whitening Jacobian $J_W(x_t, t, c) = W_B J_\phi(x_t, t, c)^\top$ and report the reduction as $S(J_\phi) - S(J_W)$. The measurements are performed on unguided trajectories (Equation 1) so that each Jacobian pair is conditioned on the same input latent x_t, while the guidance gradient is computed only to estimate W_B and not used to update the sampling trajectory. Figure 4a shows that our whitening scheme consistently reduces the spectral dominative ratio, with more prominent reductions in tail cases during the early timesteps. For example, the reduction for the 95th-percentile sample at timestep index 0 is around 0.26.

5.2 Prompt Reweighting

Motivated by subsection 4.3, which shows that the main-to-start token attention ratio ξ is often correlated with spectral anisotropy, we introduce a second method, denoted as prompt reweighting, to alleviate anisotropy by reducing ξ. Given the token embedding sequence E of the text condition, we compute $\tilde{E} = \eta E$, where $\eta > 1$, and $\tilde{E}$ is used in place of E as the input to the cross-attention layers during chosen timesteps. Figure 4c shows that, empirically, increasing η continues to lower ξ in the sampled images and is accompanied by a lower spectral dominative ratio $S(J_\phi)$, consistent with our analysis.

We provide an intuitive explanation for why prompt reweighting can reduce ξ. Consider a standard cross-attention layer with a linear key projection. When all token embeddings are uniformly scaled by η before the key projection, the resulting key vectors, and hence the token attention logits, are approximately scaled by the same factor $\eta > 1$. This has the same effect as lowering the softmax temperature, sharpening the attention distribution toward tokens that already have larger logits. Since the start token typically receives the highest attention, this sharpening further favors the start token and suppresses the relative attention of other tokens, thereby decreasing the main-to-start attention ratio ξ. A detailed discussion is provided in Appendix B.2.

Table 1: **Quantitative results.** We report CLIP, self-similarity, and LPIPS scores. Green, light green, and red mark best, second-best, and worst results. ScFlow achieves the best trade-off among training-free baselines, ranking first on average across metrics and conditions on both RFMs.

Method	Depth			Canny			Normal			HED			Sketch			Rank ↓
	CLIP↑	Self-Sim↓	LPIPS↑	CLIP↑	Self-Sim↓	LPIPS↑	CLIP↑	Self-Sim↓	LPIPS↑	CLIP↑	Self-Sim↓	LPIPS↑	CLIP↑	Self-Sim↓	LPIPS↑	
SDXL (SDM)																
DRF	0.249	0.133	0.918	0.248	0.132	0.931	0.251	0.133	0.827	0.250	0.133	0.796	0.250	0.133	0.800	-
Ctrl-X	0.246	0.090	0.815	0.246	0.079	0.893	0.241	0.089	0.761	0.254	0.079	0.760	0.246	0.087	0.774	-
FreeControl	0.247	0.081	0.753	0.249	0.070	0.773	0.252	0.080	0.609	0.254	0.077	0.682	0.252	0.083	0.718	-
InstaFlow																
ControlNet	0.278	0.071	0.806	0.279	0.055	0.852	0.287	0.074	0.786	0.275	0.040	0.774	0.290	0.080	0.821	-
ControlNet++	0.283	0.053	0.817	0.282	0.053	0.887	-	-	-	0.267	0.033	0.777	-	-	-	-
DRF	0.278	0.122	0.851	0.277	0.122	0.921	0.277	0.122	0.811	0.278	0.121	0.820	0.278	0.122	0.835	3.33
Ctrl-X	0.278	0.135	0.838	0.279	0.135	0.872	0.278	0.137	0.780	0.278	0.137	0.857	0.279	0.136	0.871	4.00
P2P+NTI	0.243	0.122	0.675	0.243	0.127	0.776	0.245	0.130	0.551	0.240	0.095	0.481	0.238	0.141	0.631	6.80
SDEdit-0.85	0.277	0.135	0.484	0.279	0.143	0.762	0.277	0.135	0.512	0.279	0.122	0.583	0.276	0.105	0.698	5.20
SDEdit-0.75	0.239	0.129	0.312	0.266	0.121	0.685	0.256	0.110	0.402	0.261	0.087	0.437	0.251	0.103	0.458	6.67
PNP	0.246	0.082	0.453	0.237	0.060	0.448	0.254	0.084	0.349	0.221	0.075	0.156	0.229	0.132	0.304	6.67
RF-Inversion	0.264	0.094	0.332	0.275	0.097	0.682	0.266	0.082	0.396	0.268	0.076	0.406	0.261	0.086	0.419	5.60
FreeControl	0.265	0.090	0.811	0.270	0.085	0.883	0.267	0.088	0.766	0.270	0.085	0.778	0.268	0.090	0.795	4.13
ScFlow (Ours)	0.277	0.078	0.840	0.281	0.072	0.895	0.277	0.081	0.772	0.278	0.073	0.823	0.278	0.082	0.841	1.80
RD (sdxl)																
ControlNet	0.250	0.082	0.844	0.266	0.052	0.874	-	-	-	-	-	-	-	-	-	-
DRF	0.266	0.124	0.900	0.264	0.132	0.933	0.266	0.124	0.821	0.262	0.130	0.803	0.261	0.130	0.805	2.27
Ctrl-X	0.256	0.097	0.781	0.254	0.082	0.902	0.254	0.093	0.744	0.257	0.092	0.774	0.254	0.090	0.766	2.67
FreeControl	0.227	0.104	0.744	0.215	0.102	0.877	0.221	0.103	0.747	0.225	0.100	0.768	0.224	0.104	0.772	3.53
ScFlow (Ours)	0.272	0.087	0.761	0.259	0.074	0.882	0.271	0.089	0.752	0.269	0.090	0.822	0.260	0.089	0.776	1.53

6 Experiments

6.1 Spatial Control

Baselines. We compare ScFlow to the uncorrected energy-guided method FreeControl (Mo et al., 2024) and other training-free spatial control methods (PNP (Tumanyan et al., 2023), P2P (Hertz et al., 2022) with null-text inversion (NTI) (Mokady et al., 2023), SDEdit (Meng et al., 2022), Ctrl-X (Lin et al., 2024) and DRF (Kim et al., 2025b)), as well as training-based methods ControlNet (Zhang et al., 2023) and ControlNet++ (Li et al., 2024). We also include RF-Inversion (Rout et al., 2024), a flow-native image-editing method that generalizes to stroke-to-image tasks. We evaluate all methods on InstaFlow. To verify ScFlow across backbones, we further compare against ControlNet and state-of-the-art training-free methods (DRF, Ctrl-X, and FreeControl), on the phased-sdxl variant of RD (Wang et al., 2025). In addition, we list the performance of training-free methods on a representative SDM, SDXL (Podell et al., 2024).

Experiment Setup. The evaluation follows the same protocol as FreeControl and uses the ImageNet-R-TI2I dataset (Tumanyan et al., 2023), which contains five condition types with each input paired with five prompts. Implementation details are provided in Appendix C.

Evaluation Metrics. We conduct quantitative evaluation with three standard metrics. DINO self-similarity distance (Tumanyan et al., 2022; Caron et al., 2021) and CLIP score (Radford et al., 2021) reflect the spatial alignment and prompt alignment of the controlled outputs, respectively. We also follow prior work to report LPIPS distance (Zhang et al., 2018) from the input condition as an indication of richer appearance and less appearance leakage from the conditions. To summarize the trade-off among these objectives, we report the average rank of each method across metrics. Additionally, since LPIPS does not fully capture perceptual image quality in this setting, we conduct a human evaluation of the generated outputs.

Comparison with Baselines. Table 1 compares ScFlow with the baselines. Flow-native editing method RF-Inversion does not perform well on non-RGB spatial conditions, and previous spatial control methods transfer unevenly to RFMs. In particular, DRF and Ctrl-X retain competitive prompt alignment but exhibit substantially weaker spatial alignment, while the uncorrected FreeControl baseline suffers from instability. As shown in Figure 5, its outputs can contain severe structural artifacts or alignment failures. ScFlow consistently improves both prompt and spatial alignment over FreeControl, while achieving the best average rank among all training-free methods on both RFMs. The qualitative results in Fig. 5 further show that ScFlow realizes the target structure with less visible artifacts and structural failures compared to other training-free methods.

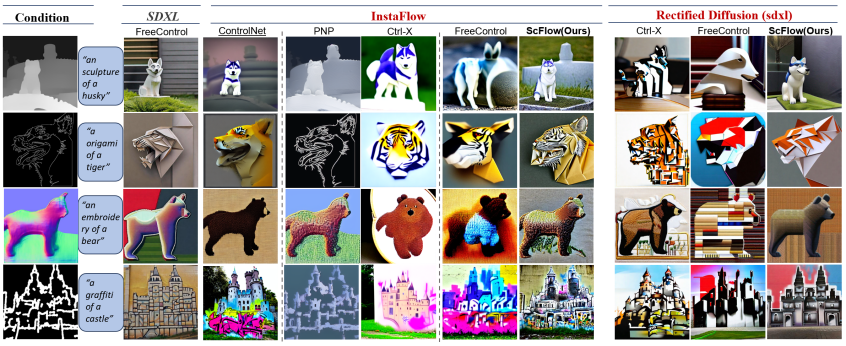

Figure 5: **Qualitative comparison.** ScFlow achieves a better trade-off between semantic and spatial alignment on RFMs compared with other training-free methods, with fewer structural artifacts.

Method	CLIP ↑	Self-Sim ↓	LPIPS ↑
Base	0.268	0.088	0.807
C1	0.271	0.084	0.803
C1 + C2	0.278	0.076	0.808
C1 + C2 + Refinement	0.278	0.077	0.834



Figure 6: **Ablation of components.**

Human Perceptual Evaluation. We conduct a human evaluation on perceptual quality, comparing ScFlow with FreeControl and the strongest baseline DRF. We randomly sample 30 condition–prompt pairs and compare ScFlow with each baseline, collecting 300 votes in total. ScFlow is preferred over both baselines in 62.0% of comparisons. Details are provided in Appendix A.4.

Component Ablation. Figure 6 shows that adding the proposed ScFlow corrections progressively improves the controllability, achieves better prompt alignment and spatial alignment scores, and produces outputs with fewer artifacts. Following Lin et al. (2024), ScFlow additionally includes an appearance refinement stage in the main evaluation, with details in Appendix C.2. Refinement leaves CLIP and self-similarity largely unchanged, while primarily increasing LPIPS, indicating that it affects visual appearance rather than the core control behavior.

7 Conclusion

We introduce ScFlow, a spectrum-aware training-free spatial control framework. ScFlow consistently outperforms the uncorrected baseline and other training-free methods on RFMs. One limitation of ScFlow is that we adopt the original FreeControl energy formulation without modification, and designing better energy objective remains future work. For example, the outputs may inherit mild color biases from the conditions, as shown in Figure 7b; incorporating more refined appearance regulation could help address this issue.



Figure 7: (a) ScFlow generalizes to novel conditions, but (b) may exhibit mild appearance bias.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance. We have not used generative AI tools for hypothesis refinement, experiment design, method implementation and result interpretation. Additionally, we used generative AI tools for formatting tables, literature searching, and proposing keywords for the paper. We have reviewed all AI-assisted work. All AI-assisted manuscript content was repeatedly reviewed by the authors. Literature and bibliographic information suggested by the AI tools are manually checked against original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We make the analytical and experimental procedures used in this work explicit throughout the paper and appendix. The spectral analysis and the two ScFlow correction components are fully specified in Section 3–5, including the definition of the Jacobian spectral statistic, the formulations of the correction methods, and the detailed procedure of the analytical experiments. The task-level evaluation follows the established spatial-control evaluation protocol, which is specified in Section 6. Appendix C then records the exact inference settings, backbone-specific configurations, and hyperparameters used.

REFERENCES

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 51304–51323, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Jean-Pierre Charlier and Paul Van Dooren. On kogbetliantz's svd algorithm in the presence of clusters. *Linear Algebra and its Applications*, 1987.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuiliere, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2174–2183. IEEE, 2023.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning*, 2023.

Jimyeong Kim, Jungwon Park, Yeji Song, Nojun Kwak, and Wonjong Rhee. ReFlex: Text-guided editing of real images in rectified flow via mid-step feature extraction and attention adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Jiwon Kim, Pureum Kim, SeonHwa Kim, Soobin Park, Eunju Cha, and Kyong Hwan Jin. Dual recursive feedback on generation and appearance latents for pose-robust text-to-image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pp. 129–147, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-X: Controlling structure and appearance for text-to-image generation without guidance. *arXiv preprint arXiv:2406.07540*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7465–7475, 2024.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. FlowChef: Steering of rectified flow models for controlled generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15308–15318, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

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*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

Tiancheng Shen, Zilong Huang, Xiangtai Li, Zhijie Lin, Jiyang Liu, Yitong Wang, Jiashi Feng, Ming-Hsuan Yang, and Jun Hao Liew. Qk-edit: Revisiting attention-based injection in mm-dit for image and video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19043–19053, 2025.

Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multi-modal diffusion transformers for enhanced prompt-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Zhicheng Sun, Zhenhao Yang, Yang Jin, Haozhe Chi, Kun Xu, Liwei Chen, Hao Jiang, Yang Song, Kun Gai, and Yadong Mu. Rectifid: Personalizing rectified flow with anchored classifier guidance. *Advances in Neural Information Processing Systems*, 37:96993–97026, 2024.

Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

Fu-Yun Wang, Ling Yang, Zhaoyang Huang, Mengdi Wang, and Hongsheng Li. Rectified diffusion: Straightness is not your need in rectified flow. In *International Conference on Learning Representations*, volume 2025, pp. 31420–31445, 2025.

Tianyi Wei, Yifan Zhou, Dongdong Chen, and Xingang Pan. Freeflux: Understanding and exploiting layer-specific roles in rope-based mmdit for versatile image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16745–16754, 2025.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Chen Xu, Weiwei Xu, and Kaili Jing. Fast algorithms for singular value decomposition and the inverse of nearly low-rank matrices. *National Science Review*, 10(6):nwad083, June 2023. doi: 10.1093/nsr/nwad083. URL `https://doi.org/10.1093/nsr/nwad083`.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

A Additional Results

A.1 More Instability Examples Across RFMs

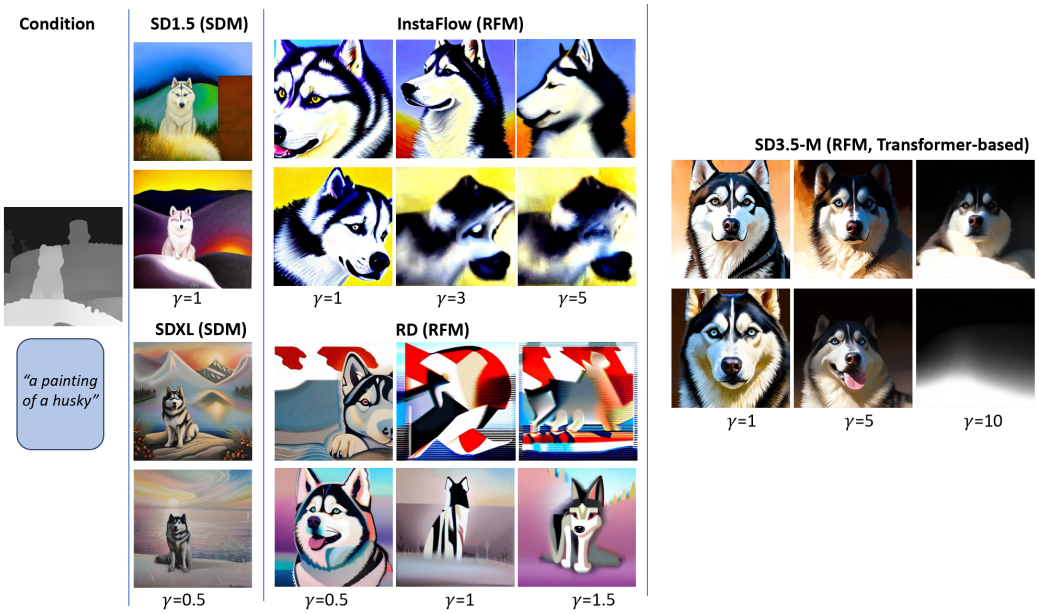

Figure 8: Same instability issue as illustrated in Figure 1 exists consistently across tested RFMs, including RD and SD3.5-M. Energy-guided sampling can not reliably perform well across RFMs even with a range of different weights (γ in Equation 2). In contrast, the method is stable across SDMs (SD1.5, SDXL).

We investigate other models using the same protocol as in Figure 1 and show that the instability issue of energy-guided sampling on the spatial control task is consistent across RFMs, including InstaFlow, Rectified Diffusion (the phased-sdxl variant) and SD3.5 (the medium variant). It is worth noting that the three models have different architectures: InstaFlow is architecture-wise similar to SD1.5, the RD variant is similar to SDXL 1.0, and SD3.5 is a transformer-based model. Nevertheless, we observe consistent instability over all three RFMs, while the performance is stable across SDMs such as SD1.5 and SDXL.

A.2 Result of Spectral Correction on MMDiT-based RFMs

We further show that one of our spectral correction methods, channel-wise whitening (C1), also yields improvements on two MMDiT-based models, SD3 and SD3.5.

C1 operates on the already-computed guidance gradient rather than architecture-specific modules and can therefore be applied directly to MMDiT-based models. Prompt reweighting (C2), on the other hand, was built on the assumption of separate cross-attention modules, unlike the joint-attention modules required by MMDiT architectures. We therefore apply C1 alone to the MMDiT models under the same protocol as the main evaluation.

Table 2: Result of channel-wise whitening (C1) on MMDiT-based RFMs.

Model	Method	CLIP ↑	LPIPS ↑	Self-Sim ↓
SD3	FreeControl	0.2630	0.8223	0.1418
	ScFlow (C1 only)	0.2702	0.8110	0.1377
SD3.5	FreeControl	0.2733	0.8190	0.1724
	ScFlow (C1 only)	0.2772	0.8229	0.1542

Across both MMDiT backbones, C1 consistently improves CLIP and reduces self-similarity distances relative to FreeControl. These improvements are consistent with the trends observed on the

two UNet backbones in the main evaluation, extending evidence that spectral correction can improve controllability beyond U-Net-based architectures and suggesting its potential applicability to modern MMDiT-based RFMs.

A.3 Full Qualitative Comparison

We show the full qualitative comparison of all evaluated methods on the two RFMs and the representative SDM, SDXL, in Figure 9 and Figure 10.

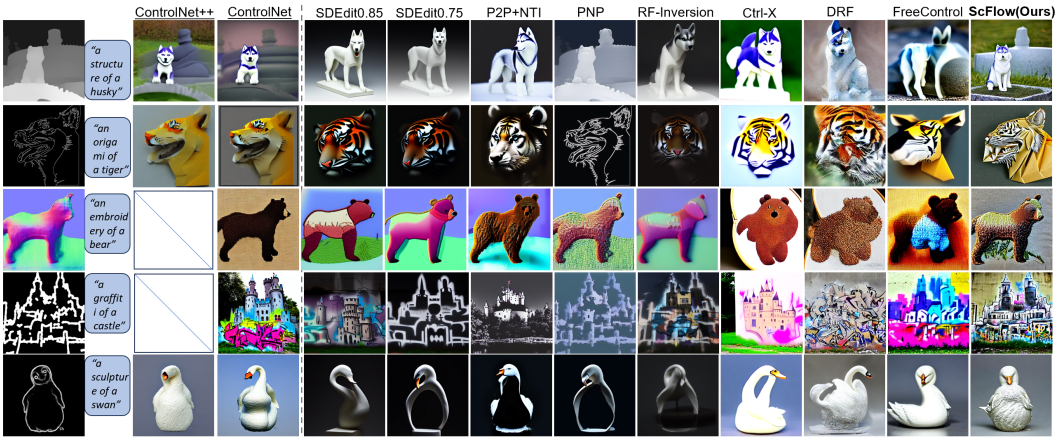

Figure 9: Full qualitative comparison on InstaFlow.

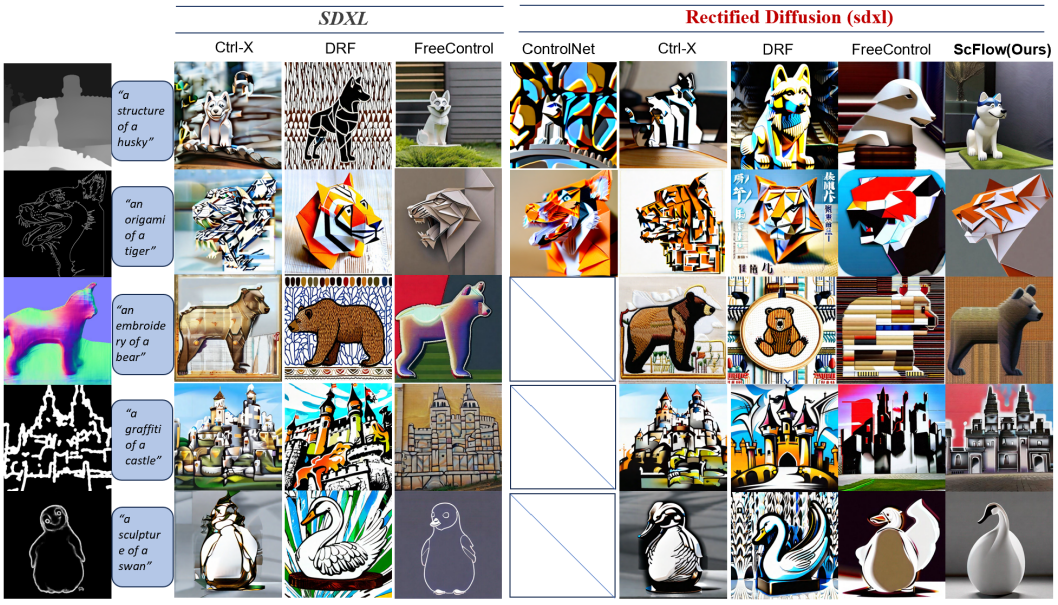

Figure 10: Full qualitative comparison on SDXL and Rectified Diffusion (phased-sdxl variant).

Table 3: Human preference study.

Statistic	vs. FreeControl	vs. DRF
ScFlow preferences	**93/150 (62.0%)**	**93/150 (62.0%)**
Conditions with a ScFlow majority ($\geq$3/5 votes)	**21/30 (70.0%)**	**18/30 (60.0%)**

A.4 Human Perceptual Evaluation

As discussed in Section 6, we conducted a blinded pairwise human preference study to supplement the automated perceptual metric, as LPIPS alone does not reflect human perceptual preferences of image quality. Our study finds that participants prefer ScFlow over both the uncorrected baseline FreeControl and the strongest baseline competitor DRF in 62.0% of comparisons. We report the full results in Table 3 and the experiment details below.

We randomly sampled 30 condition–prompt pairs from the main evaluation set and generated one output for each pair using ScFlow, DRF, and FreeControl. This yielded 60 pairwise comparisons: 30 between ScFlow and DRF and 30 between ScFlow and FreeControl. Each comparison received five independent judgments, resulting in 300 valid responses from 30 Prolific participants, with each participant completing 10 trials. In each trial, one ScFlow result and one baseline result were presented side by side with randomized left–right placement and no method labels. To evaluate perceptual image quality independently of spatial or textual alignment, we did not show the condition image or text prompt and asked participants to select the image with better overall visual quality.

A.5 Results on novel Condition Types

As shown in Figure 7a, ScFlow works on a wide range of novel conditions and diverse prompts. The outputs show both reasonable alignment to the spatial conditions and the prompt content.

A.6 Sensitivity Study

ScFlow does not use instance-specific or modality-specific tuning during evaluation and remains stable across the evaluated condition types and RFM backbones. The main evaluation already covers five condition types and two RFM backbones using fixed settings within each backbone. Regarding hyperparameters, the first correction component, channel-wise whitening, is parameter-free, while the second component, prompt reweighting, introduces one scalar hyperparameter, the reweighting scale η, which is fixed to 1.5 throughout evaluation experiments. We show the robustness of ScFlow by dedicated sensitivity analyses for (a) prompt variation, (b) spatial guidance steps, (c) reweighting scale η.

A.6.1 Sensitivity Analysis on Prompt Variations

We stratify 100 image–prompt pairs from the original evaluation set and construct two additional prompt variants for each pair: a long descriptive prompt and a multi-object prompt. The prompts are generated with an LLM and manually verified. The multi-object setting evaluates whether additional semantic content interferes with ScFlow. We report the details of these prompt groups in Table 4 and the sensitivity experiment results in Table 5.

Table 4: Details of prompt groups used for prompt sensitivity test of ScFlow. Mean, median, and range refer to prompt length, measured in words.

Setting	Mean	Median	Range
Short (base)	5.00	5	5–5
Multi-object	11.78 ± 0.84	12	10–13
Long descriptive	19.76 ± 0.75	20	17–21

ScFlow exhibits improved CLIP and DINO self-similarity scores with long descriptive prompts. Multi-object prompts produce only small changes across three metrics. Overall, the metric variations remain small across the tested prompt settings, indicating that ScFlow is not brittle to different prompt length or additional prompt contents.

Table 5: Sensitivity of ScFlow on different prompt groups.

Setting	CLIP ↑	Δ vs. base	LPIPS ↑	Δ vs. base	Self-sim ↓	Δ vs. base
Short (base)	0.2752	–	0.7917	–	0.0790	–
Multi-object	0.2707	-1.64%	0.7965	+0.61%	0.0818	+3.54%
Long descriptive	0.2899	+5.34%	0.7705	-2.68%	0.0777	-1.65%

Table 6: Sensitivity of ScFlow with different spatial guidance steps.

Setting	CLIP ↑	LPIPS ↑	Self-sim ↓
Steps-5	0.2884	0.8426	0.1263
Steps-10	0.2841	0.8129	0.0938
Steps-15 (base)	0.2752	0.7917	0.0790

A.7 Sensitivity Analysis on Guidance Steps

We denote the number of diffusion steps to which energy guided sampling is applied as guidance steps. It is worth noticing that the energy guided sampling on spatial control task is usually applied starting from the first diffusion step. We vary the number of guidance steps to show that the performance of our method is not brittle under different guidance step settings. The results are reported in Table 6. Fewer spatial guidance steps favor semantic alignment, whereas increasing the number of spatial guidance steps progressively improves structural alignment, which is expected. The default 15-step setting prioritizes spatial correspondence, while the overall trend shows a predictable trade-off rather than a narrow operating range requiring precise tuning.

A.8 Sensitivity Analysis on prompt-reweighting scale

We stratify 100 image–prompt pairs from the evaluation set to analyze the sensitivity of ScFlow over the hyperparameter η of the second component, prompt reweighting. We report the result in Table 7. ScFlow remains stable around the default setting, with η producing a smooth trade-off between semantic and structural alignment rather than abrupt performance changes. Larger η typically improves spatial alignment while decreasing prompt alignment.

Table 7: Sensitivity of ScFlow with different prompt reweighting scales η.

Setting	CLIP ↑	LPIPS ↑	Self-sim ↓
$\eta = 1.0$	0.2698	0.7992	0.0855
$\eta = 1.3$	0.2774	0.7910	0.0843
$\eta = 1.4$	0.2776	0.7887	0.0809
$\eta = 1.5$ (base)	0.2752	0.7917	0.0790
$\eta = 1.6$	0.2746	0.7949	0.0766
$\eta = 1.8$	0.2699	0.8013	0.0735
$\eta = 2.0$	0.2652	0.8088	0.0700

B Method Details

B.1 More Details on Channel-Wise Whitening

Structural Pattern of Latent-Feature Jacobian $J_\phi(x_t, t, c)$. We probe and visualize latent-feature Jacobian matrix $J_\phi(x_t, t, c)$ to examine its structural pattern. Intuitively, $J_\phi(x_t, t, c)$ characterizes the local response of the chosen self-attention key feature $\phi(x_t, t, c)$ to infinitesimal perturbations in the latent input x_t. Although $\phi(x_t, t, c)$ is represented as a sequence of attention features, these features are projected from a UNet feature map, which maintains a coarse spatial correspondence to the input latent map x_t. One may therefore expect the input-output sensitivity to be spatially localized.

Our visualization supports this intuition. Since the output-channel dimension is too high, directly displaying all output-channel responses would be infeasible. Therefore, we aggregate the Jacobian over the output-channel dimensions using squared energy, while preserving the input-index axis.

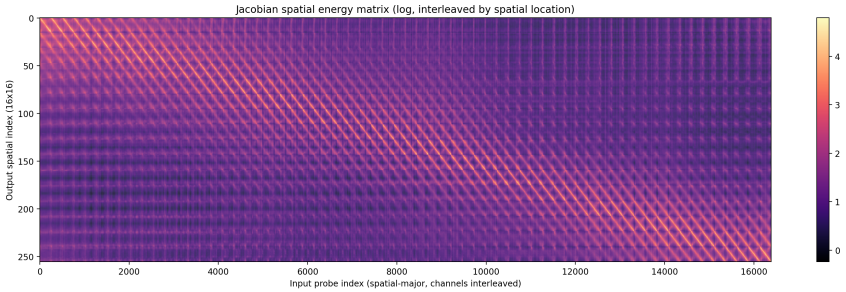

Figure 11: Illustration of the structural pattern of $J_\phi(x_t, t, c)$.

The resulting plot shows the sensitivity energy from each latent input entry to each output spatial location.

This aggregated view reveals a clear near-diagonal concentration pattern: perturbations at input latent entries belonging to one spatial location mainly affect a small neighborhood near the corresponding spatial location in the output feature map. Such spatial locality suggests that the underlying $J_\phi(x_t, t, c)$ exhibits high spatial locality and a near-block-diagonal pattern. This structural observation motivates our first correction method, channel-wise whitening.

B.2 More Details on Prompt Reweighting

Effect of Uniform Embedding Scaling on Token Attention Ratios. We provide an explanation for why prompt reweighting, which uniformly scales all token embeddings by a scalar η before they enter the cross-attention layers, can reduce the main-to-start token attention ratio ξ defined in Sec. 4.3. For clarity, we first analyze a single attention head of a standard cross-attention layer with only linear projection layers.

Let e_j denote the embedding of token j, and let $k_j = W_K e_j$ be the corresponding key vector. For a query vector q_i at query location i, the attention logit and attention weight assigned to token j are

$$\ell_{ij} = \frac{q_i^\top k_j}{\sqrt{d_k}}, \qquad p_{ij} = \frac{\exp(\ell_{ij})}{\sum_r \exp(\ell_{ir})}.$$

Here, d_k is the key dimension, and the factor $\sqrt{d_k}$ is the standard attention scaling factor. With prompt reweighting, we uniformly scale the token embeddings by $\eta > 1$, i.e., $\tilde{e}_j = \eta e_j$. Ignoring the projection bias for clarity, the linear key projection gives

$$\tilde{k}_j = W_K \tilde{e}_j = \eta W_K e_j = \eta k_j.$$

Therefore, the attention logits are also scaled:

$$\tilde{\ell}_{ij} = \frac{q_i^\top \tilde{k}_j}{\sqrt{d_k}} = \eta \ell_{ij}.$$

The resulting attention weight is thus

$$\tilde{p}_{ij} = \frac{\exp(\eta \ell_{ij})}{\sum_r \exp(\eta \ell_{ir})}.$$

For the start token 0 and the main token m, we define the local main-to-start attention ratio at query location i as

$$\xi_i = \frac{\tilde{p}_{im}}{\tilde{p}_{i0}} = \frac{\exp(\eta \ell_{im})}{\exp(\eta \ell_{i0})} = \exp\left(\eta(\ell_{im} - \ell_{i0})\right).$$

This shows that increasing η directly sharpens the softmax distribution: tokens with larger logits become relatively more dominant, while tokens with smaller logits are suppressed more strongly. Empirically, the start token receives the largest attention weight. Therefore, we have

$$\frac{\partial}{\partial \eta} \log \xi_i = \ell_{im} - \ell_{i0} < 0.$$

Thus, increasing η tends to reduce the main-to-start token attention ratio at the query level. Our empirical proxy ξ, however, is defined as the ratio between the averaged main-token attention and the averaged start-token attention, rather than the direct average of ξ_i. Therefore, the above query-level derivation does not strictly prove the monotonicity of ξ, but it suggests that ξ should tend to decrease when the same sharpening effect is aggregated across queries and heads. This trend is supported empirically in Figure 4c.

C Implementation Details

C.1 More Details on Energy-Guided Sampling Objective

In the main paper, we write the spatial-control objective as a general feature-level energy $\mathcal{E}(\phi(x_t, t, c), y_t)$, where $\phi(x_t, t, c)$ denotes an intermediate feature of the current sample and y_t denotes the corresponding reference feature. This notation emphasizes the guidance dynamics rather than the exact form of the detailed energy objective.

In our experiments, we instantiate this energy with the original FreeControl objective (Mo et al., 2024). We next provide details of this energy formulation for completeness. FreeControl first calculates and saves a time-dependent semantic basis B_t from self-attention features by sampling a small number of seed images of the target content. It then projects both the current feature and the reference feature y_t into this basis, e.g.,

$$A_t = B_t^\top \Phi_t, \qquad A_t^y = B_t^\top y_t. \tag{7}$$

Its structural energy term further combines foreground matching between A_t and A_t^y inside a binary mask M with background suppression outside the mask, while its appearance term matches semantic-component-wise statistics with a sibling sample generated without spatial control, denoted as I_a. It is worth noting that M is directly estimated during the re-generation process of the spatial condition image I_s from its inverted noise s_T and does not require additional tools or user inputs. Therefore, the full objective can be viewed more explicitly as

$$\mathcal{E}_{\mathrm{FC}} = \mathcal{E}(\phi(x_t, t, c); y_t, B_t, M, I_a), \tag{8}$$

where B_t, M, and I_a are auxiliary quantities. In the main paper, we omit these auxiliary quantities for notational simplicity.

C.2 More Details on Inference

Ctrl-X (Lin et al., 2024) and DRF (Kim et al., 2025b) both apply an additional self-recurrence refinement stage after structural and appearance control to enhance low-level appearance. Therefore, for fair comparison purpose, we also introduce a refinement stage to the full pipeline in main evaluation. Empirically, we find that the self-recurrence correction used in prior diffusion-based methods does not transfer well to RFMs, and we instead adopt standard Score Distillation Sampling (SDS) as our refinement method. We apply SDS over the discretized timestep indices from T_s to T_e. At each selected timestep index, we run K_{sds} refinement steps with learning rate γ_{sds}.

Empirically, this refinement stage enriches low-level appearance details. However, it cannot fully remove color biases inherited from the condition images, as shown in Figure 7b. Integrating an enhanced appearance transfer method as a more refined energy-guided sampling objective therefore remains a promising direction for future work.

C.3 More Details on Baselines

We use the official implementations and recommended hyperparameters for all baselines. We only modify protocol-level settings required for fair comparison, including image resolution, sampling steps, and backbone-specific scheduler settings. On the SDXL backbone, all images are sampled with their original inference steps, which are 50, 50, and 200 for Ctrl-X, DRF, and FreeControl, respectively. On RFMs, all images are sampled with 40 inference steps.

C.4 Experiment Hyperparameters

For ScFlow, we apply 15 guidance steps out of the total 40 inference steps. The gradient guidance scale γ is set to 0.1 and 0.05 on InstaFlow and Rectified Diffusion, respectively. Channel-wise whitening is applied to all guidance steps, while prompt reweighting is only applied to the first 10 steps with $\eta = 1.5$. For the refinement method (SDS), $T_s = 5$, $T_e = 15$, $K_{\text{sds}} = 2$, and $\gamma_{\text{sds}} = 0.01$.

D Limitations

Speed Limitation. The original FreeControl method has identified inference time as a key limitation of their work, requiring 25 additional seconds on average compared to normal DDIM sampling under their settings of 120 guidance steps and 200 inference steps. This issue is moderately alleviated because our method is able to work with fewer guidance and inference steps (15 and 40, respectively, in our experiment settings) due to RFMs' straighter trajectories. Overall, ScFlow requires 67% longer inference time (6 seconds more) than standard sampling on a common RTX 3090 GPU without code optimization. If the refinement method in Appendix C.2 is also included, the inference time increases to 178% longer (16 seconds more). We report the quality-speed tradeoff in Table 8.

Table 8: Quality-speed tradeoff of ScFlow.

Configuration	CLIP ↑	Self-sim ↓	LPIPS ↑	Time/image (RTX 3090)
FreeControl	0.268	0.088	0.807	15 s
ScFlow w/o refinement	0.278	0.076	0.808	15 s
ScFlow with refinement	0.278	0.077	0.834	25 s

Scope Limitation. We examine our analytical findings on two UNet-based RFMs of different architectures and one MMDiT-based RFM, SD3.5-M. The results show that high early-stage anisotropy (Figure 3) and instability in energy-guided sampling (Figure 8) are consistently observed across the evaluated RFMs, despite architectural differences. Our core spectral correction method, channel-wise whitening, is able to improve both control alignment metrics (CLIP as prompt alignment metric and DINO self-similarity as spatial alignment metric) when applied to MMDiT-based RFMs as shown in A.2, demonstrating the efficiency of spectral correction as a general method for improving spatial control stability on RFMs across architectures. However, our second correction method, prompt reweighting, is currently motivated by assumptions and empirical observations from UNet-based models, on which text conditioning is implemented through separate cross-attention layers. In contrast, MMDiT-based architectures use joint attention over image and text tokens (Esser et al., 2024), making a direct extension of our full framework to MMDiT non-trivial. We leave this extension for future work.