FLAT FEATURES, DETAILED IMAGES? SIMPLE DETAIL CONTROL FOR DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

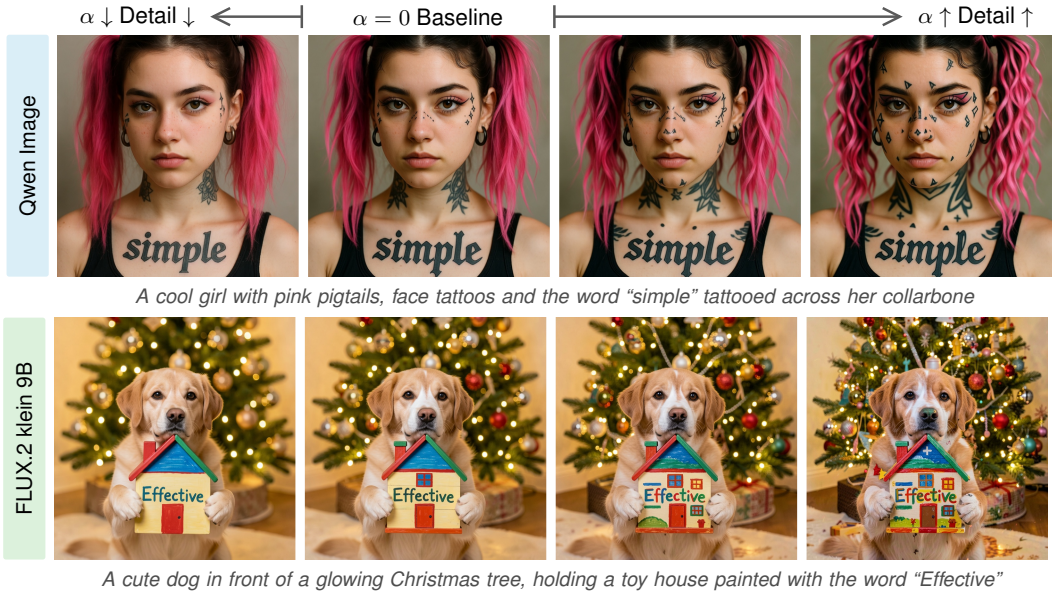

Figure 1: Qualitative examples of our method across two representative diffusion transformers. Our method controls detail richness through a user-specified scalar α as a lightweight plugin.

ABSTRACT

Prompt engineering is commonly used to encourage text-to-image models to generate more detailed images. However, it does not provide reliable, continuous, or disentangled control. We analyze feature tokens produced with and without detail-enhancing prompts across six diffusion transformers and observe a common pattern: these prompts tend to increase the entropy of each token's energy distribution while largely preserving its ordering. That is, flattening the distribution increases the level of detail. Motivated by this counter-intuitive finding, we introduce a simple training-free approach that continuously adjusts the level of detail according to a scalar. Specifically, it shifts the squared channel magnitudes of each token toward a higher entropy distribution with a time-aware scheduler to concentrate the intervention at intermediate sampling steps. Across extensive experiments on different types of diffusion transformers, we show that our approach provides continuous control over detail metrics while largely preserving image quality and semantic structure. It requires only a few lines of code and adds negligible inference time overhead.

1 INTRODUCTION

Prompt engineering is a straightforward approach that can be used to enhance visual detail in images generated by text-to-image (T2I) models (Liu & Chilton, 2022; Mo et al., 2024), for example, by adding phrases such as "highly detailed" or "fine texture". However, such instructions provide vague

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.

control, making their strength difficult to specify precisely. Their effects are also unpredictable and model-dependent, as different models may interpret the same phrase differently. Moreover, prompt-based control is not localized and may alter the global semantic structure, complicating iterative refinement. We aim to understand how such detail-focused prompt engineering impacts feature tokens internally, and whether universal patterns exist that can be leveraged for finer control.

Flow-matching models (Lipman et al., 2023) with diffusion transformers (DiTs) (Peebles & Xie, 2023) have emerged as a dominant paradigm for T2I generation. We first investigate how such prompts affect feature tokens in pretrained DiTs. Given the same image embedding, we compare the features obtained with and without detail instructions in the prompt. Across six variants of DiTs, we analyze all tokens at every layer and inference timestep. We find that detail-based instructions consistently increase the entropy of the token's energy distribution, that is, the distribution of its squared channel amplitudes, while largely preserving their rank ordering. In contrast, adding content descriptions rather than detail descriptions exhibits this pattern weakly or not at all. This observation suggests that detail richness could be controlled in a training-free manner by directly rescaling channel magnitudes. It may seem counter-intuitive that flattening a token's feature distribution should make the associated image patch *less* flat. We speculate that activating a sparse set of feature dimensions encodes a specific texture, so activating a larger set (i.e., a higher entropy distribution) leads to a mixture of textures.

Motivated by this finding, we introduce T♭ (Token Flattening), a simple mechanism for controlling the level of detail in the generated images according to a user-specified scalar α. T♭ linearly interpolates the per-token distribution of squared channel amplitudes (i.e., its energy distribution) with the uniform distribution, increasing its entropy. To improve stability, we design a time-aware scheduler that emphasizes the intervention at intermediate sampling stages. T♭ requires only a few lines of code and can be applied directly to DiTs as a plug-in.

We evaluate T♭ on three representative diffusion transformers: FLUX.2 Klein 9B, FLUX.2 Klein 9B Base, and Qwen-Image (Black Forest Labs, 2026; Wu et al., 2025), covering distilled and undistilled variants with both single and double stream architectures. Varying α produces continuous changes across all evaluated detail metrics while largely preserving text alignment, perceptual quality, and semantic structure. Some qualitative examples are demonstrated in fig. 1. Compared with existing training-free detail-enhancement methods, T♭ achieves greater structure-preserving detail improvements and increased image quality, with negligible additional inference time overhead.

2 Related Work

Diffusion transformers. Previously, diffusion models were based on U-Net architectures (Rombach et al., 2022; Podell et al., 2024; Li et al., 2024). Nowadays, Diffusion Transformers (DiTs) (Peebles & Xie, 2023) are dominant due to their better performance and scalability. Text-conditioned DiTs have evolved into several architectural variants. For example, text embeddings can be incorporated into the image stream via cross-attention (Chen et al., 2024). In MMDiT architectures, text and image embeddings are processed in two separate streams, with joint attention enabling interaction between the two modalities (Esser et al., 2024; Wu et al., 2025). In single-stream architectures, text and image embeddings are concatenated into a unified sequence and processed jointly through self-attention (Team et al., 2025). Some models combine both designs, using dual-stream blocks followed by single-stream blocks (Black Forest Labs, 2024; 2026). Furthermore, some alternatives leverage linear attention instead to speed up inference (Xie et al., 2025).

Manipulating T2I models at inference time. Given a pretrained text-to-image model, its inference process can be manipulated to achieve various desired effects. For example, some methods scale up test-time computation through search or optimization to obtain more desirable generations according to a preference model (Kim et al., 2025; Ma et al., 2025a; Singhal et al., 2025). Other methods improve generation quality and mitigate sampling artifacts by constructing degraded internal predictions as negative references (Hong et al., 2023; Ahn et al., 2024; Zhang et al., 2026; Hong, 2024). Ma et al. (2026); Bu et al. (2025) enable models to generate images beyond their training resolution. Prior work has also adapted T2I models for image editing (Hertz et al., 2023), inpainting (Lugmayr et al., 2022), super-resolution (Saharia et al., 2023), interpolation (Yu et al., 2025), 3D generation (Poole et al., 2023), and personalized generation (Ruiz et al., 2023).

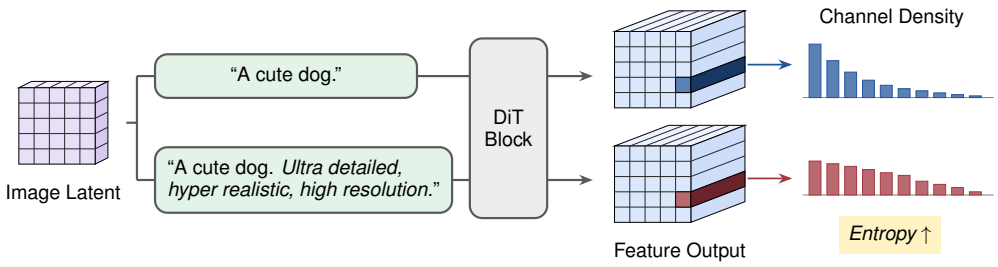

Figure 2: Overview of the analysis experiment described in section 3. Given the same image latent, we compare DiT features obtained using an original prompt and its detail-guided version. For corresponding token pairs, detail instructions produce higher entropy over squared channel magnitudes.

Our work aims to provide a knob for controlling the detail level of generated images. Some closely related methods enhance the detail of images by amplifying high-frequency energy, including FreeU (Si et al., 2024), Fresca (Huang et al., 2025), and FDG (Sadat et al., 2025). However, FreeU is designed for U-Nets, while Fresca and FDG target undistilled models that use CFG (Ho & Salimans, 2021). Furthermore, Gan et al. (2026) enhances detail through negative guidance constructed by masking the massively activated channels, but requires an additional model forward pass, which causes large extra inference computation. Chen et al. (2026); Wang et al. (2026) do this through prompt manipulation but do not provide continuous control. In contrast, our method generalizes across DiT architectures while being efficient and explicitly controllable.

3 The Effect of Detail-Enhancing Prompts

We first analyze feature tokens in pretrained DiTs under additional detail-oriented text conditioning. Given an image VAE latent with $H \times W$ patches, a DiT first up-projects it into a hidden feature of shape $(H \times W) \times D$, where D is the token dimension. For the same input image latent, we extract the output features after each transformer block under text conditions with and without detail-enhancing instructions as suffix and compare their differences. This process is illustrated in fig. 2. To obtain appropriate detail instructions, we first perform a voting-based search for each model to obtain phrases that produce more detail, as measured by the Fourier high-frequency band energy ratio. The searching process is described in section A.1. We evaluate a total of 1,000 prompts from the COCO2017 validation set (Lin et al., 2014) across six commonly used open-weight undistilled and distilled DiT models (Black Forest Labs, 2026; Esser et al., 2024; Team et al., 2025). We use $\mathbf{q}$ and $\hat{\mathbf{q}} \in \mathbb{R}^D$ to denote the produced feature token without and with detail instructions, respectively. The main comparison results are summarized in table 1 and discussed further in section A.

Detail instructions tend to increase the entropy of token energy distributions. We compute the normalized energy distribution of the token using the squared amplitudes, analogous to the Born rule from quantum physics (Born, 1926):

$$\mathbf{p}_i = \frac{\mathbf{q}_i^2}{\|\mathbf{q}\|^2}, \tag{1}$$

where $i = 1, \ldots, D$ is the channel index. Denoting $\mathbf{p}$ and $\hat{\mathbf{p}}$ as the densities computed from the corresponding token pair $\mathbf{q}$ and $\hat{\mathbf{q}}$, we compare their entropies $H(\mathbf{p})$ and $H(\hat{\mathbf{p}})$, averaging over all tokens, layers, timesteps, and prompts. We found that detail-enhancing instructions mildly increase the mean entropy for five of the six models. To provide stronger statistical evidence, we quantify the agreement ratio of this entropy increase. Let L, S, and K denote the numbers of transformer layers, inference steps, and prompts. We calculate two aggregated pairwise agreements:

$$A_{\text{all}} = \frac{1}{N} \sum_{n=1}^{N} 1 \left[H(\hat{\mathbf{p}}^{(n)}) > H(\mathbf{p}^{(n)}) \right], \; N = HWLSK, \tag{2}$$

$$\text{and } A_{\text{prompt}} = \frac{1}{K} \sum_{k=1}^{K} 1 \left[\sum_{n=1}^{N} (H(\hat{\mathbf{p}}^{(n,k)}) - H(\mathbf{p}^{(n,k)})) > 0 \right], \; N = HWLS, \tag{3}$$

Model	$\overline{H}(\mathbf{p})$	Added text	$\overline{H}(\hat{\mathbf{p}})$	A_{all}	A_{prompt}	Spearman
FLUX.2 Klein 9B Base	4.463	content	4.460	46.0%	45.6%	0.956
		detail	**4.530**	**78.0%**	**100.0%**	**0.965**
FLUX.2 Klein 9B	4.307	content	4.292	39.7%	37.6%	0.799
		detail	**4.433**	**83.7%**	**100.0%**	**0.878**
SD3.5 Large	3.065	content	3.070	59.2%	66.6%	0.929
		detail	**3.080**	**71.4%**	**98.2%**	**0.966**
SD3.5 Large Turbo	3.006	content	3.013	57.7%	66.4%	0.881
		detail	**3.027**	**70.0%**	**96.2%**	**0.928**
Z-Image Base	0.057	content	0.057	50.8%	55.8%	0.859
		detail	0.057	46.6%	**70.8%**	**0.895**
Z-Image Turbo	1.087	content	1.088	59.0%	58.0%	0.868
		detail	**1.096**	**62.2%**	**94.4%**	**0.903**

Table 1: For each model, we first report the baseline mean entropy ($\overline{H}(\mathbf{p})$) of the token density defined in eq. (1). Under each type of added text condition, we report the mean entropy ($\overline{H}(\hat{\mathbf{p}})$), together with pairwise statistics computed between spatially corresponding tokens of baseline: the entropy increasing agreement ratios (eqs. (2) and (3)) and the mean Spearman correlation. Bold values indicate the patterns observed under detail instructions as discussed in section 3.

A_{all} measures the fraction of token pairs where $\hat{\mathbf{p}}$ has higher entropy than $\mathbf{p}$, which substantially exceeds 50% for five out of six models. At the prompt level, A_{prompt} measures the fraction of prompts where $\hat{\mathbf{p}}$ has higher summed entropy. This agreement exceeds 94% for five out of six models, which shows that the trend is consistent across prompts. Therefore, we see that flatter (higher entropy) per-token energy distributions correspond to richer image details. This might seem counter-intuitive, since a flatter distribution of RGB values within a patch would imply less variation and fewer details. Nonetheless, we find that a flatter distribution across channels decodes to a less uniform RGB distribution across pixels. We speculate that activating a sparse set of feature dimensions encodes a specific texture, while activating a larger set leads to a mixture of textures.

Detail instructions largely preserve the rank of channel magnitudes. For each pair of corresponding tokens, we measure the Spearman correlation between their channel-magnitude ranks and average it over all token pairs. Under detail instructions, the correlation ranges from 0.878 to 0.966 across models, where 1.0 indicates perfect rank preservation and 0.0 indicates no correlation. This result suggests that detail instructions tend not to alter the relative importance of different channels, which provides a useful cue to allow a simple training-free control.

These findings are specific to detail instructions. Does this effect arise simply from longer prompts or specifically from detail-oriented instructions? To distinguish these factors, we add a content-adding baseline generated by an LLM that appends additional semantic information to the original prompt. Under additional content, neither the entropy change nor the agreement ratio shows a consistent trend. Its Spearman correlation is also lower than that of detail conditioning, suggesting that extra semantic content changes the channel scale ordering more strongly. Thus, longer prompts alone do not explain the consistent effects of adding detail instructions. We summarize all the findings in this section as follows:

Summary: *Detail instructions tend to increase the entropy of squared channel magnitudes within each feature token while largely preserving their rank.*

4 Simple Detail Control in Diffusion Transformers

To mimic the above observed patterns, we propose a simple training-free mechanism named T♭ that allows users to explicitly control the detail richness of generated images through a scalar parameter α on DiTs. The overall procedure is outlined in fig. 3.

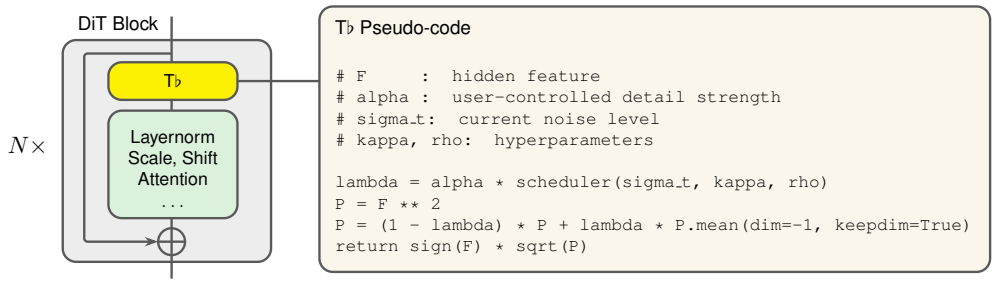

Figure 3: Illustration of the Tþ pseudo-code corresponding to eqs. (1) and (4) to (6). Tþ is injected into the residual branch of each DiT block.



Figure 4: Qualitative effects of applying Tþ at a single sampling step with the same positive weights using FLUX.2 Klein 9B. At early steps, our method tends to enhance low-frequency details, such as adding objects; but it may also alter the global semantic structure if applying it too early. At later steps, Tþ introduces more localized, high-frequency details, such as subtle freckles on faces and coarse fabric textures. That motivates our time-aware scheduler design in section 4.

Token Entropy Amplifier. Following the definition in eq. (1), we compute the normalized channel density $\mathbf{p}_{h,w}$ from each token feature $\mathbf{F}_{h,w}$. The goal is to increase the entropy of $\mathbf{p}_{h,w}$. Since entropy is strictly concave and the uniform distribution is its unique maximiser, we interpolate it towards $1/D$, which raises entropy monotonically:

$$\mathbf{p}'_i = (1 - \lambda(\alpha, \sigma_t))\, \mathbf{p}_i + \lambda(\alpha, \sigma_t)\, \frac{1}{D}, \tag{4}$$

where λ is a scalar determined by the user-specified guidance strength α and the noise level σ_t of the current sampling step. Finally, we revert the redistributed density $\mathbf{p}'$ to amplitude space, rescaling and restoring original per-channel signs:

$$\mathbf{q}'_i = \mathrm{sign}(\mathbf{q}_i) \sqrt{\mathbf{p}'_i \textstyle\sum_{j=1}^{D} \mathbf{q}_j^2}. \tag{5}$$

By default, the amplifier is applied to all feature tokens. For greater stability, we apply it only within the residual branch. In other words, it affects only the current block's residual update while leaving the accumulated hidden states from preceding blocks unchanged.

Time-aware Scheduler We model the interpolation strength λ in eq. (4) by scaling α according to the current σ_t. A flow-matching or diffusion model progressively samples from a standard Gaussian distribution ($\sigma_t = 1$), toward a high-density data point ($\sigma_t = 0$). We observe that applying our operation at early timesteps alters more low-frequency components, which may undesirably affect the global semantic structure. At late timesteps, it mainly alters high-frequency components, causing very localized changes. A qualitative example is shown in fig. 4. We therefore aim to concentrate the guidance on intermediate timesteps, where it primarily affects mid-frequency components that correspond to meaningful details. Specifically, we define

$$\lambda(\alpha, \sigma_t) = \alpha \left(\frac{\sigma_t}{\rho}\right)^{\kappa\rho} \left(\frac{1 - \sigma_t}{1 - \rho}\right)^{\kappa(1-\rho)}, \tag{6}$$

where $\rho \in (0, 1)$ sets the peak location and κ its concentration. Thus, λ reaches its maximum α at $\sigma_t = \rho$ and remains below α elsewhere.



Figure 5: A qualitative example of T♭ from the experiment in table 2 across three models. Although the prompt already contains detail instructions, T♭ further enhances visual details, demonstrating that it is orthogonal to prompt engineering. Further examples are provided in section B.8.

5 Experiment

In this section, we first demonstrate the performance of T♭ on three DiTs in section 5.1. Then, we compare T♭ with other related methods in section 5.2. Finally, we provide further discussion of the choice of density formulation, the time-wise effect, and the limitations in section 5.3.

Models and implementation details We use three representative diffusion transformers: FLUX.2 Klein 9B Base, its distilled variant FLUX.2 Klein 9B (Black Forest Labs, 2026), and Qwen-Image (20B) (Wu et al., 2025). The Base and Qwen models use 25 inference steps with a classifier-free guidance (CFG) scale of 4, whereas the distilled model uses 4 steps without CFG. All models generate 768×768 images. For the time-aware scheduler, we set $\kappa = 40$ for all models and $\rho = 0.9$, 0.7, and 0.8 for the distilled, Base, and Qwen models. More details are provided in section B.2.

Evaluation Metrics To assess image quality, we use CLIP (Radford et al., 2021) (Long-CLIP-L, ViT-L/14) and HPSv3 (Ma et al., 2025b) (based on Qwen2-VL-7B) to evaluate text-image alignment and human preference. We quantify image detail using four metrics: edge strength (mean gradient magnitude), sharpness (Laplacian variance), spectral edge frequency f_{95} (radial frequency enclosing 95% of Hann-windowed Fourier energy, in cycles per pixel), and wavelet detail energy (Mallat, 1989). Detailed definitions of them are provided in section B.7. Higher values across these metrics indicate stronger detail components. We assess structure preservation by comparing generated images and the baseline, using DINO (Oquab et al., 2024) cosine distance between the CLS tokens and LPIPS (Zhang et al., 2018) computed after downsampling both images to 128×128.

5.1 Detail Control Results

We conduct the experiment on 800 prompts obtained by combining the HPDv2 (Wu et al., 2023) and HPDv3 (Ma et al., 2025b) test sets, since these prompts span diverse categories and complexity levels. We evaluate T♭ over a reasonable range of α values for each model and report the results in table 2. Compared with the baseline ($\alpha = 0$), T♭ consistently increases or decreases image detail metrics as α becomes more positive or negative, as indicated by all four averaged detail metrics For

Input α	Image Quality		Detail Statistics				Structure Difference	
	CLIP↑	HPSv3↑	Edge	Sharpness	f_{95}	Wavelet	DINO	LPIPS_{128}
Qwen Image								
-0.15	0.2771	11.212	0.0170	0.0048	0.0336	0.0196	0.1146	0.1517
-0.1	**0.2774**	11.474	0.0171	0.0051	0.0357	0.0209	0.0922	0.1251
-0.05	0.2773	11.749	0.0173	0.0055	0.0381	0.0227	0.0645	0.0846
0 (baseline)	0.2769	12.011	0.0182	0.0063	0.0435	0.0268	0.0000	0.0000
$+0.05$	0.2765	**12.013**	0.0194	0.0072	0.0494	0.0317	0.0544	0.0713
$+0.1$	0.2762	11.880	0.0203	0.0076	0.0536	0.0358	0.0828	0.1110
$+0.15$	0.2759	11.603	0.0211	0.0079	0.0575	0.0402	0.1133	0.1524
$\nearrow_\alpha$	-	-	✓	✓	✓	✓	-	-
FLUX.2 Klein 9B								
-0.03	0.2748	11.908	0.0313	0.0192	0.0553	0.0317	0.0532	0.1007
-0.02	0.2747	11.986	0.0322	0.0200	0.0566	0.0322	0.0446	0.0857
-0.01	**0.2749**	12.097	0.0335	0.0210	0.0585	0.0331	0.0315	0.0601
0 (baseline)	0.2748	12.271	0.0360	0.0230	0.0634	0.0354	0.0000	0.0000
$+0.01$	0.2745	**12.434**	0.0403	0.0265	0.0723	0.0391	0.0403	0.0719
$+0.02$	0.2738	12.315	0.0448	0.0304	0.0823	0.0429	0.0759	0.1309
$+0.03$	0.2723	11.948	0.0500	0.0355	0.0952	0.0476	0.1178	0.1924
$\nearrow_\alpha$	-	-	✓	✓	✓	✓	-	-
FLUX.2 Klein Base 9B								
-0.0045	0.2732	8.535	0.0275	0.0214	0.0511	0.0330	0.0743	0.1267
-0.003	0.2741	9.586	0.0294	0.0231	0.0560	0.0353	0.0455	0.0780
-0.0015	0.2745	10.397	0.0317	0.0247	0.0615	0.0377	0.0214	0.0340
0 (baseline)	0.2750	11.041	0.0347	0.0268	0.0682	0.0407	0.0000	0.0000
$+0.0015$	**0.2751**	11.405	0.0386	0.0299	0.0762	0.0434	0.0189	0.0284
$+0.003$	0.2750	**11.520**	0.0424	0.0336	0.0845	0.0455	0.0394	0.0624
$+0.0045$	0.2749	11.480	0.0465	0.0382	0.0948	0.0477	0.0616	0.0978
$\nearrow_\alpha$	-	-	✓	✓	✓	✓	-	-

Table 2: T♭ results across three diffusion transformer models and three metric groups over a range of α values. $\nearrow_\alpha$ indicates a monotonic increase with respect to the given αs. T♭ provides continuous detail control while maintaining image quality and causing only small structural changes.

image quality assessment, CLIP scores barely change with α, indicating that T♭ largely preserves text-image alignment. HPSv3 scores even improve with slightly positive α, suggesting that a moderate increase in detail may benefit human preference. DINO and LPIPS_{128} increase with $|\alpha|$, but they remain relatively low within the evaluated ranges, indicating limited structural changes. Some qualitative examples are illustrated in fig. 5.

5.2 Comparison with others

Several existing training-free methods may also enhance image details. First, we compare against methods that introduce additional guidance, including SAG (Hong et al., 2023), PAG (Ahn et al., 2024), SEG (Hong, 2024), TPG (Rajabi et al., 2025), DG (Gan et al., 2026), and SSG (Zhang et al., 2026). These methods employ CFG-like operations to steer the sampling process away from negative predictions constructed by perturbing intermediate features in different ways. We also compare with FreSca (Huang et al., 2025) and FDG (Sadat et al., 2025), which apply different guidance weights to the high and low frequency components of the CFG term. Finally, we compare with HiGS (Sadat et al., 2026), which enhances details using a weighted average of historical predictions.

For a fair comparison, we apply all methods to FLUX.2 Klein 9B Base using identical settings and the same prompt set. We progressively increase the guidance strength of each method until its structural deviation from the baseline exceeds that of ours. Even at this point, our T♭ achieves higher detail metrics. This indicates that our method provides more effective, detail-focused control. In terms of image quality, while all the methods maintain a similar level of CLIP score, T♭ delivers the highest HPSv3 preference. In addition, most guidance methods require about $0.5\times$ more inference

Method	Latency over Baseline↓	Image Quality		Detail Statistics				Structure Difference	
		CLIP↑	HPSv3↑	Edge↑	Sharpness↑	f_{95} ↑	Wavelet↑	DINO↓	LPIPS_{128} ↓
Baseline	1.00×	0.2750	11.0413	0.0347	0.0268	0.0682	0.0407	0.0000	0.0000
SAG	1.56×	0.2740	11.0230	0.0389	0.0241	0.0776	**0.0491**	0.0690	0.1286
PAG	1.53×	0.2743	10.9532	0.0363	0.0315	0.0761	0.0464	0.0663	0.1012
SEG	1.53×	0.2749	10.5784	0.0372	0.0293	0.0713	0.0436	0.1081	0.1763
TPG	1.53×	0.2738	10.5830	0.0336	0.0218	0.0606	0.0385	0.0752	0.1303
DG	1.53×	0.2747	10.8166	0.0338	0.0252	0.0702	0.0413	0.0881	0.1503
SSG	1.58×	0.2742	10.2274	0.0288	0.0132	0.0533	0.0398	0.0993	0.1533
FreSca	1.03×	**0.2752**	11.2964	0.0379	0.0294	0.0763	0.0459	0.0719	0.1194
FDG	**1.02**×	0.2744	10.9110	0.0368	0.0261	0.0652	0.0414	0.1774	0.2724
HiGS	**1.02**×	0.2749	11.1414	0.0360	0.0277	0.0707	0.0418	0.0637	0.1026
Tþ (Ours)	1.03×	0.2749	**11.4796**	**0.0465**	**0.0382**	**0.0948**	0.0477	**0.0616**	**0.0978**

Table 3: Comparison with training-free detail enhancement methods on FLUX.2 Klein 9B Base. Tþ achieves the strongest detail statistics and quality with smallest structural changes. It adds negligible inference time compared to the baseline.



Figure 6: Comparison with the best-performing competing methods in table 3. Tþ better enhances local details over the baseline. Zoom in to better view the differences. More examples in section B.9.

computation, while Tþ only adds a negligible overhead of 3%. We report the results and the added computation over baseline in table 3. A qualitative comparison is shown in fig. 6.

5.3 Further Discussion

Time-aware effects of Tþ. We apply Tþ at one sampling step at a time with a fixed guidance weight $\alpha > 0$ using the four-step FLUX.2 Klein 9B model. In fig. 7, we report the percentage change in spectral energy across five frequency bands, whose definitions are provided in section B.4. Early interventions distribute the energy change relatively evenly across frequencies, whereas later interventions increasingly amplify high-frequency energy. Figure 8 shows that early interventions cause larger structural deviations from the baseline. A qualitative example is provided in fig. 4. Together, these observations motivate the time-aware scheduler in section 4, which emphasizes guidance at in-

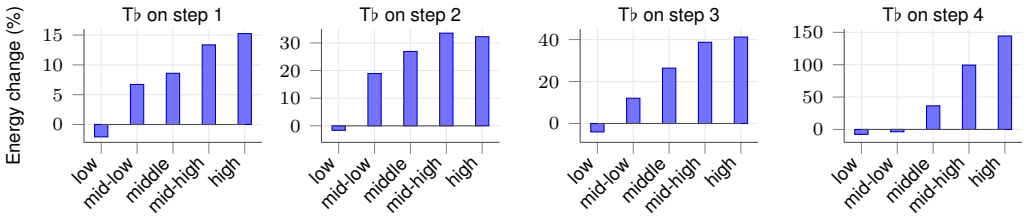

Figure 7: Percentage change in spectral energy of different frequency bands when Tþ is applied at a single sampling step using the same weight with FLUX.2 Klein 9B.

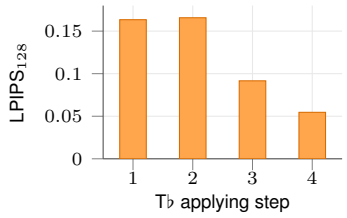

Figure 8: Structural difference from the baseline when Tþ is applied at a single sampling step.

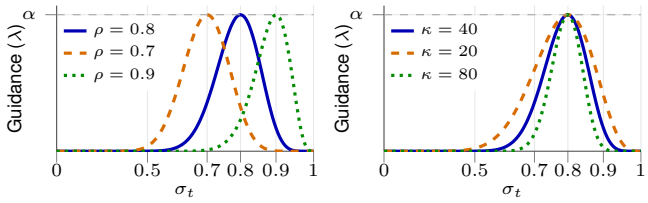

Figure 9: Time-aware scheduler curves. Left: varying ρ shifts the peak of the curve. Right: varying κ controls the width of the curve. The peak is always equal to α.

termediate sampling steps. Figure 9 visualizes the scheduler curves: ρ determines the peak location, while κ controls the width. A larger κ produces a sharper and more concentrated schedule. Both parameters can be adjusted to accommodate different user preferences.

Why not define density using absolute magnitudes? In eq. (1), we define channel density using squared magnitudes. A straightforward alternative is to use absolute magnitudes. We repeat the analysis in fig. 2 with this definition and report the results in table 4. Compared with squared magnitudes, absolute ones yield lower entropy-increase ratios. We attribute this difference to the amplification effect of squaring. Squaring emphasizes changes among higher-magnitude channels while suppressing contributions from lower ones. This reveals that detail instructions redistribute density primarily among highly active channels. This observation is consistent with prior findings on massive activations, where tokens are often sparse and dominated by few channels (Gan et al., 2026; 2025; Yehezkel et al., 2026). The effect is more pronounced for models with sparser tokens.

Tþ relies on the model capabilities. We further evaluate Tþ on the other DiT families. As reported in table 7, Tþ retains monotonic control over most detail metrics but is less effective than on stronger FLUX and Qwen models, and image quality and structural similarity decline faster for the same change. We hypothesize that stronger models learn a more densely populated feature space and tolerate token perturbations better, whereas perturbing weaker models may more easily push features outside the learned distribution, degrading image quality. This dependence on the inherent capability of the pretrained model represents a limitation of Tþ as a training-free method.

Failure cases and future work. Tþ operates within a limited effective range and may come with some side effects. Additional examples are provided in section B.8. A fixed pushing toward a uniform target is unlikely to be the optimal direction for every token, but characterizing a per-token optimal direction is hard for a training-free method. We hope this observation motivates future work that enables fine-grained control over detail, which is further discussed in section C.

6 Conclusion

We analyze how detail-guided prompts alter feature tokens across diffusion transformers and identify a consistent pattern. Motivated by this observation, we propose Tþ, a simple training-free method for controlling detail richness with a user-specified scalar. Our work provides insight into the mechanisms behind prompt engineering and an effective method for precise detail control.

AI USE STATEMENT

In this work, we used generative AI tools to aid the implementation of the experiments in sections 3 and 5. We also used generative AI tools to discuss ideas and lightly polish the writing. The AI tools we used include GPT-5.6 (OpenAI, 2026) and Claude Opus (Anthropic, 2026). We carefully reviewed all AI-assisted work to ensure the code is correctly implemented and the text is accurate and appropriate. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide sufficient implementation details in sections A.1, B.1, B.2, 3 and 5 to ensure that our results can be reproduced without issues, including pseudocode, models used, experimental settings, hyperparameters, and datasets.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, 2024.

Anthropic. Claude opus. `https://www.anthropic.com/claude/opus`, 2026.

Black Forest Labs. FLUX.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024.

Black Forest Labs. Flux.2 [klein]. `https://bfl.ai/models/flux-2-klein`, 2026.

Max Born. Quantenmechanik der stoßvorgänge. *Zeitschrift für Physik*, 1926.

Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. In *Advances in Neural Information Processing Systems*, 2025.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, 2024.

Lifeng Chen, Jiner Wang, Zihao Pan, Beier Zhu, Xiaofeng Yang, and Chi Zhang. Detail++: Training-free detail enhancer for t2i diffusion models. *IEEE Transactions on Image Processing*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *Advances in Neural Information Processing Systems*, 2025.

Chaofan Gan, Zicheng Zhao, Yuanpeng Tu, Xi Chen, Ziran Qin, Tieyuan Chen, Mehrtash Harandi, and Weiyao Lin. Massive activations are the key to local detail synthesis in diffusion transformers. In *International Conference on Learning Representations*, 2026.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems*, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Chao Huang, Susan Liang, Yunlong Tang, Jing Bi, Li Ma, Yapeng Tian, and Chenliang Xu. Fresca: Scaling in frequency space enhances diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *International Conference on Learning Representations*, 2025.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014.

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

Vivian Liu and Lydia B. Chilton. Design guidelines for prompt engineering text-to-image generative models. In *CHI Conference on Human Factors in Computing Systems*, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Yiyang Ma, Feng Zhou, Xuedan Yin, Pu Cao, Yonghao Dang, and Jianqin Yin. Resdit: Evoking the intrinsic resolution scalability in diffusion transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *IEEE/CVF International Conference on Computer Vision*, 2025b.

Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M. Weber. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales. *arXiv preprint arXiv:2506.19713*, 2025.

Seyedmorteza Sadat, Farnood Salehi, and Romann M. Weber. Higs: History-guided sampling for plug-and-play enhancement of diffusion models. In *International Conference on Learning Representations*, 2026.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning*, 2025.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Linqing Wang, Zhiyong Xu, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Zhenxi Li, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. Promptenhancer: Taming your rewriter for text-to-image generation via fine-grained reward. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025.

Shai Yehezkel, Shahar Yadin, Noam Elata, Yaron Ostrovsky-Berman, and Bahjat Kawar. Accelerating text-to-video generation with calibrated sparse attention. In *European Conference on Computer Vision*, 2026.

Qingtao Yu, Jaskirat Singh, Zhaoyuan Yang, Peter Henry Tu, Jing Zhang, Hongdong Li, Richard Hartley, and Dylan Campbell. Probability density geodesics in image diffusion latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

Weijia Zhang, Yuehao Liu, Shanyan Guan, Wu Ran, Yanhao Ge, Wei Li, and Chao Ma. Guiding a diffusion model by swapping its tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

APPENDIX

A MORE ABOUT THE OBSERVATION EXPERIMENT

A.1 DETAIL INSTRUCTION SEARCH

Before the analysis in section 3, we search for a model-specific prompt suffix that consistently elicits more visual details from each model. Specifically, we first did a detail-guided prompt search for each model to get phrases that can enable more detail, measured by the Fourier high-frequency energy ratio. We build a phrase bank from an LLM: e.g. "ultra detailed", "fine textures", "micro-contrast", and selected phrases as a suffix for each prompt and did a voting based on the increase of the detail measurement. Each model gets the highest-voted phrases to use as a suffix. In the end, we get the phrases for each model and use them as suffix. Then we input the same image VAE latent to the diffusion transformer with different text conditions, one with the original prompt and the other with the detail-guided suffixes.

After the search, we use the following model-specific suffixes in the final analysis experiments: "hyper realistic, ultra detailed, high resolution" for FLUX.2, "intricate surface detail, fine textures, ultra detailed" for SD3.5, and "micro-contrast, studio lighting, fine textures" for Z-Image.

During this prompt search, we also have some interesting findings. Effective detail-guided suffixes are strongly model-dependent. For example, "ultra detailed" increases the FFT-based detail score for FLUX.2 but decreases it for Z-Image. Across our search, the preferred suffixes vary more across models than across input prompts, indicating that detail instructions are more model-dependent than prompt-dependent. Longer suffixes provide no consistent benefit and can even reduce detail; empirically, two to four phrases are sufficient. It suggests that prompt engineering is not a very reliable way to control details, and T♭ provides a more robust alternative.

A.2 COMPARISON OF ENTROPY TRENDS BETWEEN ABSOLUTE AND SQUARED MAGNITUDES

We further examine how our observation changes when token density is defined using absolute rather than squared channel magnitudes. The complete comparison is reported in table 4. The absolute-magnitude formulation yields lower entropy-increase agreement ratios, especially for the sparser representations of SD3.5 and Z-Image. As discussed in section 5.3, squaring emphasizes changes among highly active channels and suppresses weaker responses, making the density redistribution under detail instructions more evident.

Model	Magnitude	$\overline{H}(\mathbf{p})$	$\overline{H}(\hat{\mathbf{p}})$	A_{all}	A_{prompt}
FLUX.2 Klein 9B Base	square	4.463	4.530	78.0%	100.0%
	absolute	7.824	7.827	68.2%	100.0%
FLUX.2 Klein 9B	square	4.307	4.433	83.7%	100.0%
	absolute	7.825	7.830	68.9%	100.0%
SD3.5 Large	square	3.065	3.080	71.4%	98.2%
	absolute	7.150	7.149	47.6%	37.4%
SD3.5 Large Turbo	square	3.006	3.027	70.0%	96.2%
	absolute	7.151	7.152	59.1%	83.4%
Z-Image Base	square	0.057	0.057	46.6%	70.8%
	absolute	4.252	4.251	45.8%	36.0%
Z-Image Turbo	square	1.087	1.096	62.2%	94.4%
	absolute	6.700	6.700	43.6%	51.6%

Table 4: Comparison of token densities constructed from squared and absolute channel magnitudes. For each model, $\overline{H}(\mathbf{p})$ and $\overline{H}(\hat{\mathbf{p}})$ denote the mean entropy without and with detail instructions, respectively. We also report the token-level and prompt-level entropy-increase agreement ratios, A_{all} and A_{prompt}.

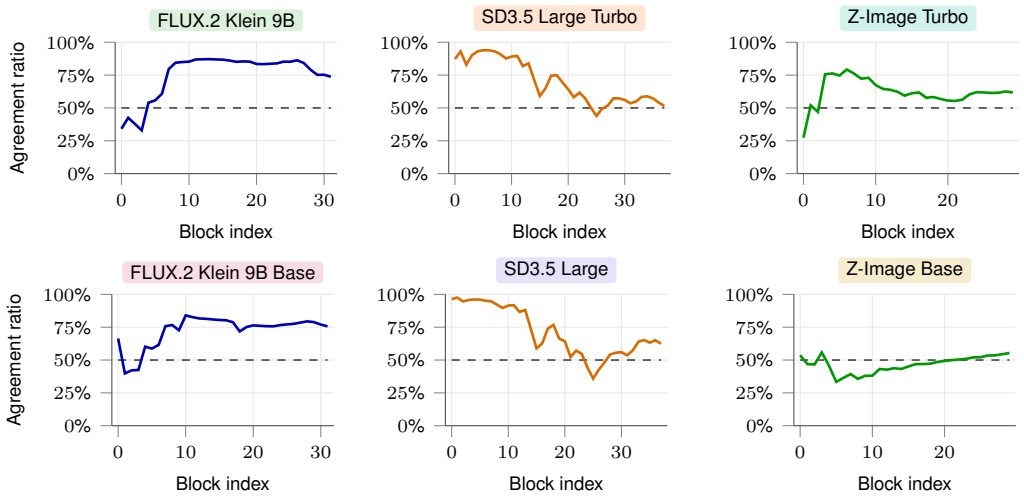

Figure 10: Block-wise entropy-increase agreement ratios computed from squared channel magnitudes. Each point is averaged over timesteps and image tokens.

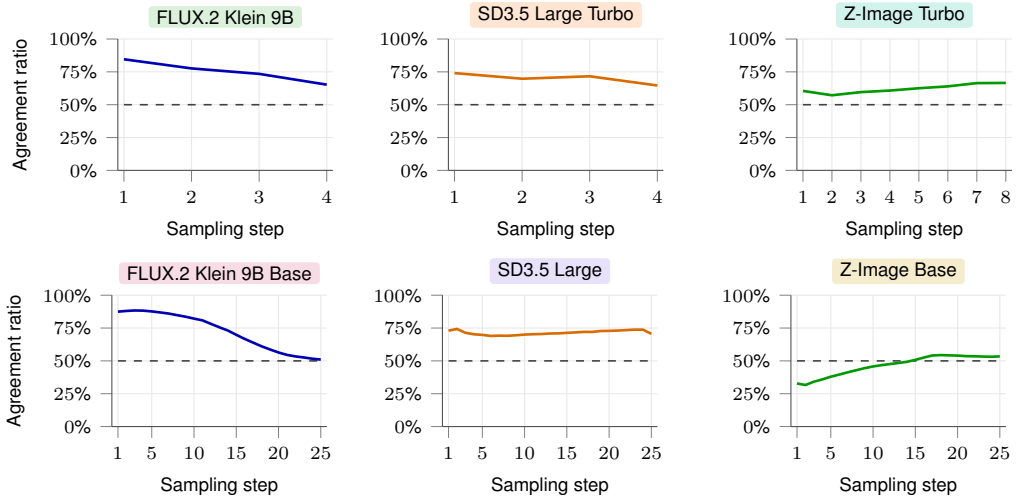

Figure 11: Time-wise entropy-increase agreement ratios computed from squared channel magnitudes. Each point is averaged over blocks and image tokens.

A.3 Block-wise Trend

We further examine how the entropy-increase agreement ratio varies across DiT blocks. For each block, we compute the fraction of spatially corresponding token pairs whose squared-magnitude entropy increases under detail instructions, averaged over prompts, timesteps, and image tokens. As shown in fig. 10, the block-wise trends vary across models and reveal no consistent advantage for weighting specific blocks more strongly. For both FLUX.2 variants, the last 24 single-stream blocks show a markedly stronger entropy-increase trend than the first eight double-stream blocks. We therefore use a uniform weight across block indices in $\mathbb{T}\flat$ as a simple default.

A.4 Time-wise Trend

We also measure the entropy-increase agreement ratio at each sampling step, averaging over prompts, blocks, and image tokens. As shown in fig. 11, the trend is relatively smooth within each model, although both FLUX.2 variants show a clearer decrease over time. However, applying $\mathbb{T}\flat$ at different sampling steps produces substantially different frequency and structural effects. We therefore determine the time-aware guidance weights using the analyses in section 5.3.

Model	$\overline{H}(\mathbf{p})$	Added text	$\overline{H}(\hat{\mathbf{p}})$	A_{all}	A_{prompt}	Spearman
FLUX.2 Klein 9B Base	4.510	content	4.507	48.7%	45.6%	0.967
		detail	**4.561**	**71.0%**	**100.0%**	**0.969**
FLUX.2 Klein 9B	4.355	content	4.328	38.8%	18.4%	0.846
		detail	**4.442**	**70.8%**	**100.0%**	**0.902**
SD3.5 Large	3.115	content	3.118	58.8%	64.0%	0.950
		detail	**3.131**	**69.8%**	**98.8%**	**0.971**
SD3.5 Large Turbo	3.013	content	3.020	57.5%	67.4%	0.908
		detail	**3.037**	**68.8%**	**98.0%**	**0.939**
Z-Image Base	0.055	content	0.056	51.5%	70.4%	0.869
		detail	0.056	51.9%	**84.4%**	**0.905**
Z-Image Turbo	1.083	content	1.083	56.5%	55.4%	0.888
		detail	**1.090**	**62.0%**	**92.8%**	**0.920**

Table 5: Same comparison as table 1, computed at 1024×1024 resolution instead of 512×512. The same pattern holds: detail instructions increase mean entropy and agreement ratios more than content instructions across models, confirming the conclusion is stable across resolutions.

Detail instruction	Top eight shares among 64 bins							
	1	2	3	4	5	6	7	8
	0.6438	0.0416	0.0317	0.0262	0.0226	0.0198	0.0177	0.0159
✓	0.6310	**0.0430**	**0.0328**	**0.0272**	**0.0234**	**0.0205**	**0.0183**	**0.0165**

Table 6: Top eight shares of the averaged 64-bin squared channel-magnitude distribution for the distilled FLUX.2 Klein 9B model, without and with a detail instruction.

A.5 Different resolution conclusion is stable

We repeat the observation experiment in section 3 at 1024×1024 resolution instead of the 512×512 resolution used in table 1, using the same prompts, models, and detail-instruction suffixes. Table 5 reports the same statistics as table 1. The conclusion is unchanged: detail instructions increase the mean entropy of squared channel magnitudes and the entropy-agreement ratios more consistently than content instructions across resolutions.

A.6 Distribution across channel-magnitude bins

We further examine how detail instructions redistribute the density (squared channel magnitudes) by computing their accumulated density. We divide the ranked channel magnitudes into 64 equal-sized bins and average each bin's share over prompts, blocks, timesteps, and image tokens. As an example, table 6 reports the eight largest shares for the distilled FLUX.2 Klein 9B model. With a detail instruction, the first bin contributes less, while the following bins contribute more, yielding a flatter distribution.

B More about the T♭

B.1 A clarification about the token amplifier

For negative values of α, the interpolation in eq. (4) may produce negative entries in $\mathbf{p}'$, violating the nonnegativity requirement of a density. In practice, we apply ReLU to $\mathbf{p}'$ and then renormalize it to sum to one for the implementation. For presentation simplicity, we omit these operations from the main formulation.

B.2 FURTHER IMPLEMENTATION DETAILS.

For FLUX.2 Klein 9B Base and its distilled variant, we inject T♭ into all single-stream blocks (last 24 blocks). Qwen-Image adopts a double-stream MMDiT architecture, and we apply T♭ to its last 40 blocks. For models distilled without CFG, we apply T♭ only to the conditional branch.

For the comparison in table 3, we use the following method-specific settings: T♭ with $\alpha = +0.0045$; FDG with $s_{\text{low}} = 1.7$ and $s_{\text{high}} = 8$; HiGS with $s = 1$; FreSca with $s = 3$ and $r_0 = 0.5$ where r_0 is the low-frequency cutoff; SEG with $s = 3$; PAG with $s = 1$; DG with $w = 0.5$; TPG with $s = 0.2$; SAG with $s = 0.75$; and SSG with $s = 0.2$. s means the guidance weight.

B.3 MORE ABOUT THE COMPETING METHOD DG

Our findings share some similarities with DG (Gan et al., 2026): both study token-channel magnitudes, and both relate to image detail. We clarify the differences below.

Mechanistically, the two guidance signals differ: DG requires an extra forward pass to construct its negative-guidance prediction, while T♭ does not. This is reflected in table 3, where DG costs $1.53\times$ the baseline latency versus $1.03\times$ for T♭.

The underlying findings also differ. DG observes massive activations: a few channels dominate a token, and masking them out causes only local distortion rather than structural change. DG therefore uses the masked prediction as negative guidance. This is effective at fixing local corruptions such as blur or unclear detail. Their method primarily serves as a safeguard against local detail degradation rather than a mechanism for further enhancing already well-rendered details. In contrast, we compare model behavior with and without a detail instruction and use that difference to construct a direction that increases detail continuously.

B.4 MORE ABOUT TIME-AWARE EFFECTS.

For the analysis in fig. 7, we partition the radial Fourier spectrum into five bands: 0–0.10π, 0.10π–0.25π, 0.25π–0.45π, 0.45π–0.70π, and 0.70π–π. Because natural images contain substantially more low-frequency than high-frequency energy, we report the relative energy change within each band rather than its absolute energy. For the four-step FLUX.2 Klein 9B example analyzed in figs. 7 and 8, we place the scheduler peak between the second and third steps to balance structural preservation and high-frequency detail enhancement.

Would a low-level post-processing operation on the baseline, such as sharpening or boosting high frequencies, achieve the same effect? This is essentially shown by the timestep comparison in figs. 7 and 8: applying such a low-level operation resembles pushing ρ to a very late step. However, low-level operations cannot generate new detail; they can only strengthen detail that is already present. T♭ can do both, and both regimes remain controllable through ρ and κ.

B.5 HYPERPARAMETER ADJUSTMENT

The scheduler parameters ρ and κ in the experiments for different models are not fixed to a single choice. As with the timestep analysis in figs. 7 and 8, moving the peak earlier or later can be a user choice: both directions are valid, and which one to pick depends on whether low-frequency structure or high-frequency detail should be emphasized.

We set $\rho = 0.9$ for FLUX.2 Klein 9B because it is a distilled, four-step model whose final sampling step already sits at $\sigma_t \approx 0.75$, so $\rho = 0.9$ falls roughly between its first two steps. FLUX.2 Klein 9B Base and Qwen-Image instead use 25-step, non-distilled sampling, where image structure solidifies more slowly, so we push the peak slightly later, to $\rho = 0.7$ and $\rho = 0.8$ respectively. In general, we find $\rho \in [0.85, 0.95]$ works best for step-distilled models, and $\rho \in [0.7, 0.8]$ works best for non-distilled models.

B.6 PERFORMANCE OF T♭ ON SD3.5 AND Z-IMAGE FAMILIES

We extend our evaluation to the SD3.5 and Z-Image families, including both base and accelerated variants. As discussed in section 5.3, the effectiveness of T♭ depends on the capability of the pre-

Input α	Image Quality		Detail Statistics $\nearrow_\alpha$				Structure Difference	
	CLIP↑	HPSv3↑	Edge↑	Sharpness↑	f_{95} ↑	Wavelet↑	DINO	LPIPS_{128}
SD3.5 Large								
-0.03	0.2722	11.2652	0.0317	0.0188	0.0377	0.0219	0.0933	0.0922
-0.02	0.2724	11.5866	0.0324	0.0193	0.0395	0.0232	0.0760	0.0767
-0.01	0.2726	11.8132	0.0331	0.0195	0.0416	0.0248	0.0600	0.0627
0 (baseline)	0.2726	**12.0507**	0.0340	0.0194	0.0441	0.0268	0.0000	0.0000
$+0.01$	0.2727	11.9897	0.0351	0.0195	0.0461	0.0287	0.0607	0.0644
$+0.02$	0.2730	11.9015	0.0360	0.0194	0.0476	0.0301	0.0763	0.0784
$+0.03$	0.2724	11.4555	0.0369	0.0192	0.0490	0.0318	0.1002	0.1037
$\nearrow_\alpha$	-	-	✓	-	✓	✓	-	-
SD3.5 Large Turbo								
-0.15	0.2676	10.6599	0.0232	0.0108	0.0182	0.0108	0.0897	0.0699
-0.10	0.2678	11.5693	0.0238	0.0111	0.0200	0.0120	0.0536	0.0475
-0.05	0.2678	11.8290	0.0242	0.0112	0.0208	0.0127	0.0329	0.0339
0 (baseline)	0.2685	**12.0571**	0.0247	0.0112	0.0214	0.0134	0.0000	0.0000
$+0.05$	0.2689	11.9711	0.0249	0.0109	0.0214	0.0136	0.0311	0.0325
$+0.10$	0.2685	11.7018	0.0250	0.0107	0.0213	0.0136	0.0476	0.0458
$+0.15$	0.2679	11.1279	0.0252	0.0104	0.0213	0.0137	0.0799	0.0655
$\nearrow_\alpha$	-	-	✓	-	-	✓	-	-
Z-Image Turbo								
-0.003	0.2628	13.1598	0.0196	0.0075	0.0350	0.0171	0.0468	0.0385
-0.002	0.2630	13.1764	0.0198	0.0076	0.0355	0.0174	0.0301	0.0250
-0.001	0.2634	**13.2295**	0.0201	0.0079	0.0366	0.0180	0.0196	0.0148
0 (baseline)	0.2641	13.1989	0.0206	0.0083	0.0382	0.0190	0.0000	0.0000
$+0.001$	0.2643	13.1321	0.0211	0.0086	0.0397	0.0201	0.0188	0.0136
$+0.002$	0.2647	12.9700	0.0216	0.0089	0.0412	0.0213	0.0397	0.0276
$+0.003$	0.2645	12.8665	0.0224	0.0092	0.0430	0.0230	0.0650	0.0460
$\nearrow_\alpha$	-	-	✓	✓	✓	✓	-	-
Z-Image Base								
-3×10^{-5}	0.2716	11.5622	0.0261	0.0181	0.0533	0.0284	0.0785	0.0724
-2×10^{-5}	0.2719	11.6882	0.0270	0.0187	0.0549	0.0295	0.0675	0.0614
-1×10^{-5}	0.2722	**11.7037**	0.0280	0.0191	0.0577	0.0311	0.0543	0.0467
0 (baseline)	0.2725	11.6860	0.0290	0.0187	0.0602	0.0328	0.0000	0.0000
$+1 \times 10^{-5}$	0.2726	11.6299	0.0302	0.0183	0.0623	0.0347	0.0575	0.0538
$+2 \times 10^{-5}$	0.2729	11.3411	0.0313	0.0177	0.0631	0.0364	0.0869	0.0872
$+3 \times 10^{-5}$	0.2731	10.9888	0.0326	0.0176	0.0641	0.0386	0.1100	0.1286
$\nearrow_\alpha$	-	-	✓	-	✓	✓	-	-

Table 7: T♭ results for the SD3.5 and Z-Image families over a range of α values.

trained model. These models generally perform worse with T♭ than the three models evaluated in the main paper. We examine whether changing α still provides monotonic detail control and how it affects image quality and structural consistency. The complete quantitative results are reported in table 7.

B.7 Detailed definitions of detail metrics

Let $I \in [0,1]^{H\times W\times 3}$ denote an RGB image at its native resolution. All four detail metrics are computed from its Rec.709 luminance,

$$Y_{ij} = 0.2126R_{ij} + 0.7152G_{ij} + 0.0722B_{ij}. \tag{7}$$

Edge strength. This metric measures the average strength of local intensity transitions and image boundaries. We compute horizontal and vertical image gradients using central differences for interior pixels and one-sided differences at the boundary:

$$(\partial_x Y)_{ij} = \frac{Y_{i,j+1} - Y_{i,j-1}}{2}, \qquad (\partial_y Y)_{ij} = \frac{Y_{i+1,j} - Y_{i-1,j}}{2}. \tag{8}$$

Edge strength is the mean gradient magnitude,

$$E(I) = \frac{1}{HW} \sum_{i,j} \sqrt{(\partial_x Y)_{ij}^2 + (\partial_y Y)_{ij}^2}. \tag{9}$$

We use direct differences rather than Sobel filters to avoid attenuating the high-frequency components being measured.

Sharpness. This metric measures the strength of fine-scale intensity variation associated with image sharpness. We apply the four-neighbor Laplacian to the $(H-2)(W-2)$ interior pixels,

$$(\Delta Y)_{ij} = Y_{i-1,j} + Y_{i+1,j} + Y_{i,j-1} + Y_{i,j+1} - 4Y_{ij}, \tag{10}$$

and define sharpness as its variance:

$$S(I) = \mathrm{Var}_{i,j}[(\Delta Y)_{ij}]\,. \tag{11}$$

Spectral edge frequency. This metric measures at which frequency of the spectral domain, the cumulative energy extends 95% of the total cumulative energy. To reduce boundary discontinuities in the Fourier transform, we first apply a separable Hann window

$$w_{ij} = h_i^{(H)} h_j^{(W)}, \qquad h_n^{(N)} = \frac{1}{2}\left(1 - \cos\frac{2\pi n}{N-1}\right), \tag{12}$$

and compute the centered power spectrum and normalized radial frequency:

$$P = |\mathrm{fftshift}\,\mathcal{F}(Y \odot w)|^2\,, \tag{13}$$

$$r_{ij} = \sqrt{\left(\frac{i - H/2}{H/2}\right)^2 + \left(\frac{j - W/2}{W/2}\right)^2}. \tag{14}$$

Let

$$C(r) = \frac{\sum_{r_{ij} \le r} P_{ij}}{\sum_{i,j} P_{ij}} \tag{15}$$

denote the cumulative fraction of spectral energy within radius r. The spectral edge frequency is the smallest radius containing 95% of the energy:

$$f_{95}(I) = \inf\{r : C(r) \ge 0.95\}. \tag{16}$$

Normalizing each axis by its Nyquist frequency makes f_{95} comparable across image resolutions and aspect ratios.

Wavelet detail energy. This metric measures the fraction of image energy contained in localized details across multiple spatial scales. We apply a three-level orthonormal Haar transform to Y, starting from $A_0 = Y$. At each level, horizontal filtering gives

$$u_{ij} = \frac{A_{i,2j} + A_{i,2j+1}}{\sqrt{2}}, \qquad v_{ij} = \frac{A_{i,2j} - A_{i,2j+1}}{\sqrt{2}}, \tag{17}$$

followed by vertical filtering:

$$LL_{ij} = \frac{u_{2i,j} + u_{2i+1,j}}{\sqrt{2}}, \qquad LH_{ij} = \frac{u_{2i,j} - u_{2i+1,j}}{\sqrt{2}}, \tag{18}$$

$$HL_{ij} = \frac{v_{2i,j} + v_{2i+1,j}}{\sqrt{2}}, \qquad HH_{ij} = \frac{v_{2i,j} - v_{2i+1,j}}{\sqrt{2}}. \tag{19}$$

We recursively set $A_l = LL_l$ and measure the fraction of energy in all detail subbands:

$$R_{\mathrm{wav}}(I) = \frac{\sum_{l=1}^{3} \left(\|LH_l\|_2^2 + \|HL_l\|_2^2 + \|HH_l\|_2^2\right)}{\|A_3\|_2^2 + \sum_{l=1}^{3} \left(\|LH_l\|_2^2 + \|HL_l\|_2^2 + \|HH_l\|_2^2\right)}. \tag{20}$$

Because the transform is orthonormal, the denominator equals the total luminance energy up to odd-size boundary cropping.

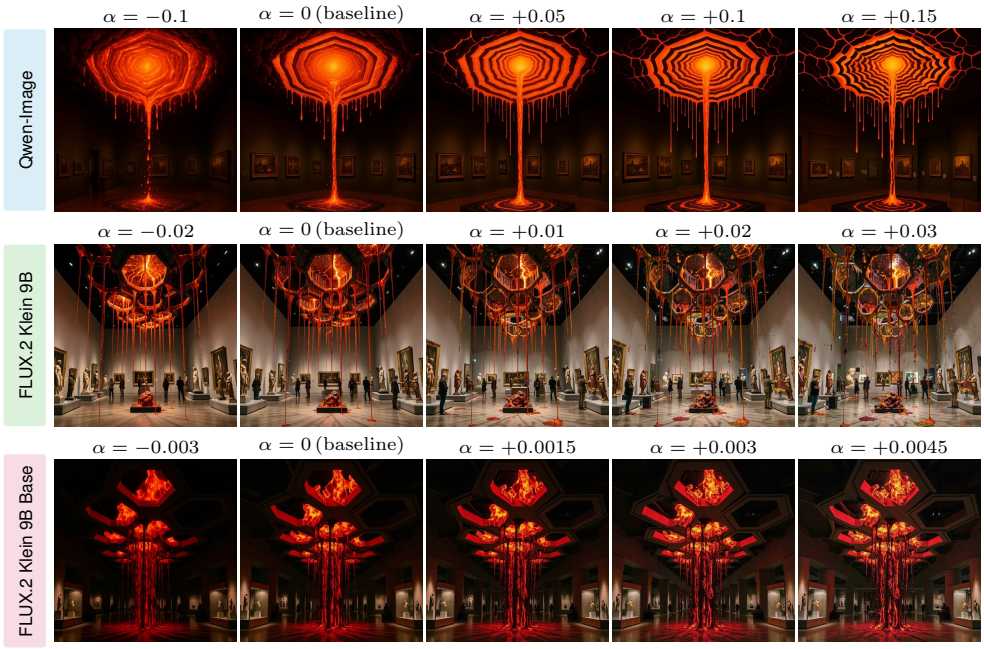

Figure 12: Additional qualitative examples following the α ranges in table 2.

B.8 More qualitative examples

Image quality assessment metrics may not fully reflect human preference, and whether an image looks better is always subjective and biased. Hence, we show some additional qualitative examples in this section. We show some examples of T♭ under different α selections and different models in figs. 12 to 16. Instead of showing only the well-performing ones, we also show some representative failure cases.

One limitation of T♭ is that it operates within a limited effective range. An excessively large α may distort images, producing cartoon-like appearances, sharp artifacts, or unnatural background clutter. The figures below provide additional qualitative examples to illustrate it more intuitively. In addition, T♭ may also shift the overall color tone of the images, not solely controlling the local details.

B.9 More comparisons with competing methods

Table 3 reports quantitative results against the competing training-free detail-enhancement methods. Here we provide additional qualitative comparisons, in the same format as fig. 6, but with two rows per prompt: the top row is unchanged (Baseline, T♭, DG, FreSca, SAG), and the bottom row adds every remaining method except FDG (PAG, SEG, TPG, SSG, HiGS).

C Future work

Table 1 also shows that entropy does not increase for every token: the agreement ratio A_{all} stays well below 100% for all models. This suggests that each token likely has its own optimal direction for detail enhancement, rather than sharing a single one. T♭ instead pushes every token's channel density toward a more uniform distribution, which is unlikely to be optimal for every token. However, characterizing a per-token optimal direction is difficult for a training-free method. We hope this observation motivates future work on training a network that enables fine-grained, per-token control over detail.



Figure 13: Additional qualitative examples following the α ranges in table 2. Failure example: the overall color tone in the last row (FLUX.2 Klein 9B Base) also shifts slightly.



Figure 14: Additional qualitative examples following the α ranges in table 2. Failure examples: Qwen-Image becomes more cartoonish at larger α, and FLUX.2 Klein 9B Base shows some local sharp color artifacts.

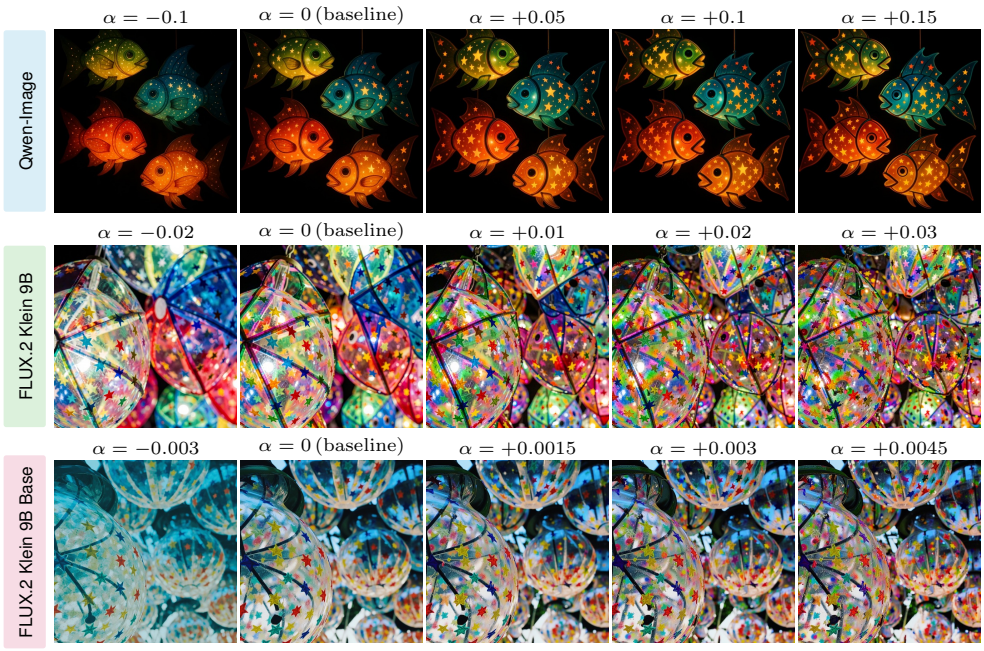

Figure 15: Additional qualitative examples following the α ranges in table 2. When the image is already highly complex with abundant detail, applying our method further can look somewhat inharmonious.

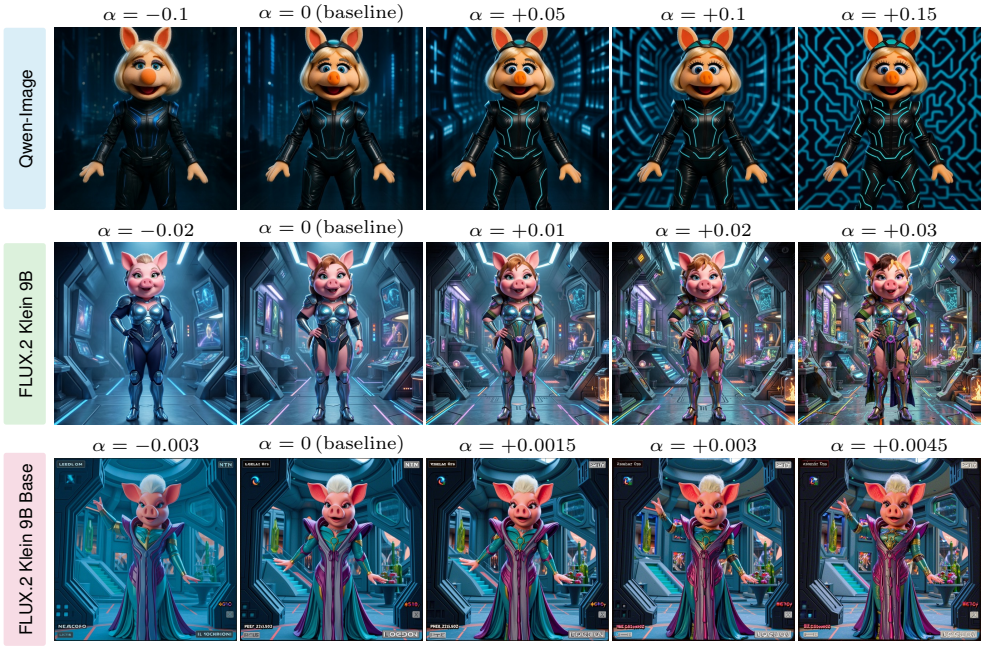

Figure 16: Additional qualitative examples following the α ranges in table 2. Failure example: in the last row (FLUX.2 Klein 9B Base), the arm pose changes across α, showing that structure preservation is not always guaranteed.

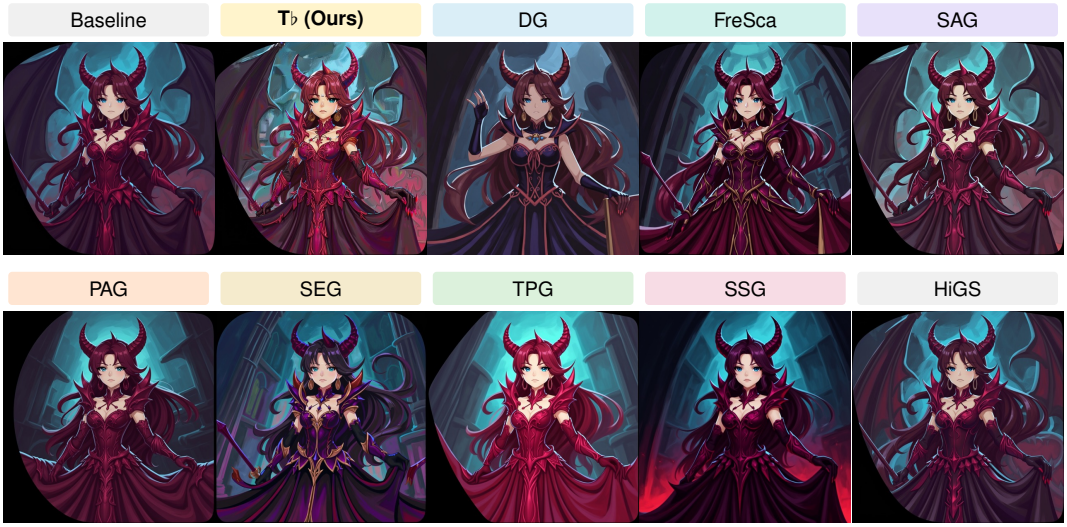

Figure 17: Additional comparison with other competing methods in table 3 on FLUX.2 Klein 9B Base. Zoom in to better view the differences.

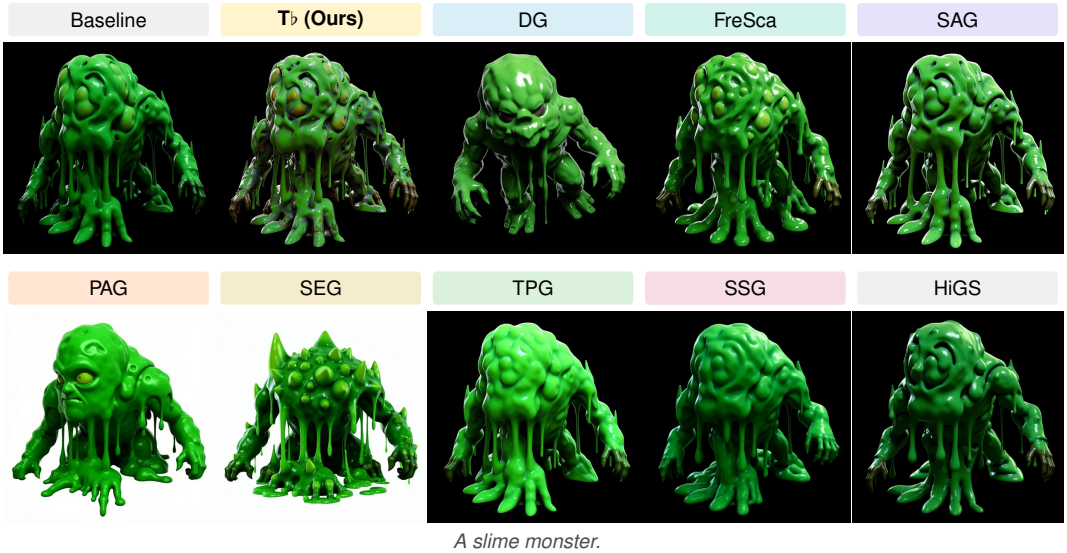

Figure 18: Additional comparison with other competing methods in table 3 on FLUX.2 Klein 9B Base. Zoom in to better view the differences.



Man jumping and holding up blanket in front of building.

Figure 19: Additional comparison with other competing methods in table 3 on FLUX.2 Klein 9B Base. Zoom in to better view the differences.



The image depicts a muscular blonde butch tomboy engineer wearing a patched flight suit, set in a detailed sci-fi environment with cinematic lighting and crepuscular rays.

Figure 20: Additional comparison with other competing methods in table 3 on FLUX.2 Klein 9B Base. Zoom in to better view the differences.

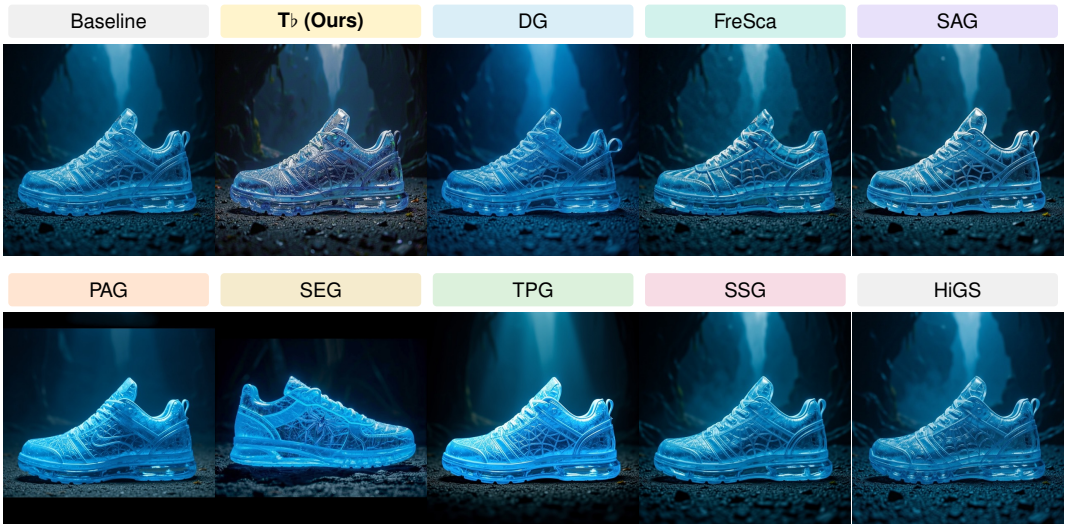

Figure 21: Additional comparison with other competing methods in table 3 on FLUX.2 Klein 9B Base. Zoom in to better view the differences.