TokenRepel: Generating Diverse Image Sets Without Compromising Quality

Qingtao Yu **Zhen Wang** **Ning Lai** **Boyang Hu** **Hongdong Li** **Dylan Campbell**
The Australian National University
{terry.yu, zhen.wang, ning.lai, boyang.hu, hongdong.li, dylan.campbell}@anu.edu.au

Abstract

We present an inference-time approach that enhances the visual diversity of image sets generated by text-to-image models, without compromising the average image quality. Our method leverages the intrinsic geometry of hidden representations in pretrained models to diversify images generated from the same prompt. First, we introduce a token repulsion mechanism that increases the angle between feature tokens across samples. It takes a negative Riemannian gradient step on a von Mises–Fisher mixture distribution of tokens on the unit hypersphere. Second, to preserve the quality of each output image, we propose a token density preservation mechanism that projects the repulsion-induced prediction change onto a half-space aligned with the model's score direction. This encourages the updates to remain within the high-density regions learned by the generative model. We further extend this mechanism to the output of each intermediate layer, enabling finer-grained iterative guidance. We evaluate our method on three types of pretrained text-to-image models. Experiments show substantial gains in visual diversity under identical prompts, while maintaining a comparable level of image fidelity.

1 Introduction

Recent advances in text-to-image (T2I) generation have led to unprecedented levels of visual quality [1–6]. However, a challenge persists in the inherent trade-off between image quality and sample diversity [7–11]. Diversity, defined as a T2I model's ability to generate varied outputs for a fixed condition, is a critical component of generative modeling. Enhancing diversity can foster creativity and provide designers with a broader range of visual choices [11, 12], support data augmentation [13] and test-time scaling [14, 15], and increase the representation of ethnicities and genders underrepresented in the data [16]. Although diffusion and flow-matching models are capable of broad mode coverage, they may suffer from "mode gravity" during inference [17]. In these scenarios, models tend to generate highly similar outputs regardless of the random initial noise, focusing on a narrow subset of the data distribution to generate only the highest confidence samples.

Existing inference-time methods diversify samples by changing prompts or text conditioning [18, 11], optimizing initial noise [19, 20], or repelling image latents [21, 22]. More recent methods intervene in intermediate features, which carry richer information about the generated content [23, 24]. However, they treat each feature representation as a whole rather than explicitly modeling interactions between individual tokens. Token-level repulsion can act more selectively, pushing highly similar cross-image tokens apart while leaving unrelated tokens untouched. For example, a token representing a dog's paw should be repelled from a similar paw token in another image, but not from a sky token. NegToMe [25] follows this token-level approach, but its linear extrapolation has undesirable mathematical properties, limiting its effectiveness. It also lacks an explicit fidelity-preserving mechanism, which can introduce artifacts.

To address this problem, we present TokenRepel, a training-free algorithm for generating diverse image sets while preserving fidelity. To model cross-image token interactions, we represent the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

normalized tokens from other images as a von Mises–Fisher (vMF) mixture on the unit hypersphere. We then update each token with a negative Riemannian gradient step on this density, increasing the angular separation between similar tokens. To prevent these updates from degrading fidelity, we introduce a token density preservation mechanism. It projects the repulsion-induced prediction change onto a half-space aligned with the model's score direction, encouraging tokens to stay in high-density regions. We further extend this mechanism to intermediate layers, allowing for more frequent and fine-grained corrections. Together, token repulsion and token density preservation increase the diversity of generated images while largely maintaining their fidelity.

Our method applies to different architectures without requiring external models or additional training. We evaluate our approach on three representative T2I models: FLUX.1-dev [1] (flow-matching), SDXL [4] (diffusion), and FLUX.2-Klein [26] (few-step distilled flow-matching). We achieve a more favorable diversity–quality trade-off than existing inference-time methods. Our approach is simple and effective, with moderate computational overhead. Our contributions are:

1. the TokenRepel algorithm, a batch-wise approach for enhancing generative diversity;
2. a high-density preservation strategy that regularizes diversity-promoting perturbations to stay within the valid data distribution, maintaining high visual fidelity; and
3. a layerwise residual preservation strategy that mitigates against error accumulation.

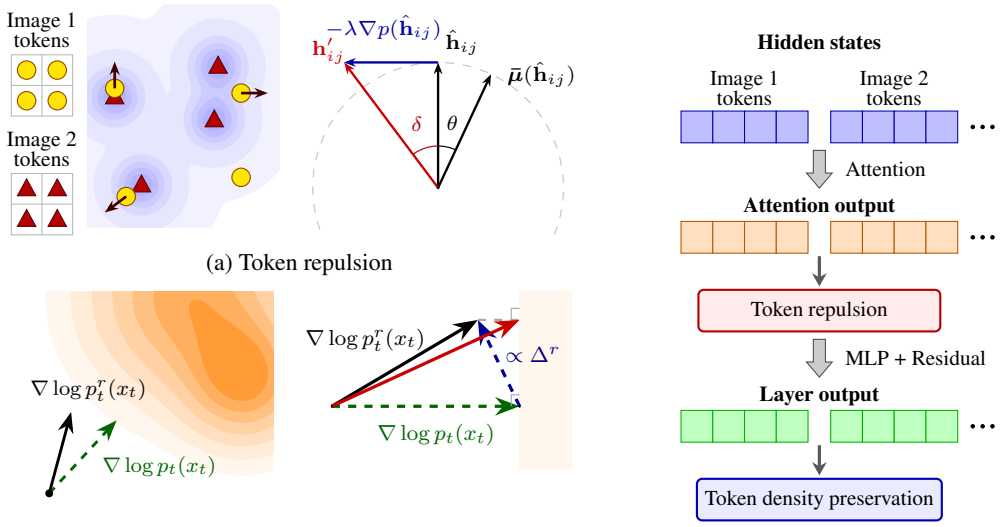
2 Related Work

Although diffusion and flow-matching models improve both diversity and fidelity over GANs [27–29], further enhancing the sampling diversity remains a critical challenge [17]. This issue is further compounded by recent trends in model optimization, as both post-training [9, 8, 30] and distillation [31] have been shown to diminish diversity. To mitigate this, Datta et al. [18] use specialized LLMs to generate diverse descriptions. Harrington et al. [19] repeatedly generate image batches and update their initial noise using diversity and quality objectives. GASS decomposes image embeddings into orthogonal axes for independent prompt-dependent and prompt-independent variation [10]. Group Inference generates a large candidate pool and selects a subset by jointly considering image quality and group diversity [32]. These approaches may require external models, additional decoding, or a much larger candidate pool.

Other inference-time interventions have also been proposed. Um and Ye [16] introduced guidance that moves samples beyond high-density regions. Contrastive Noise Optimization (CNO) optimizes batches of initial noise latents with a contrastive loss on their Tweedie estimates [33]. CADS perturbs text embeddings early in sampling according to a schedule [11]. SPARKE and Particle Guidance (PG) optimize batch latents with dissimilarity objectives based on inverse Rényi kernel entropy and pairwise RBF kernels, respectively [21, 22]. Shielded Diffusion applies sparse repulsion to nearby Tweedie estimates [34]; DSM disperses intermediate features through feature self-guidance [23]; and Contextual Repulsion repels contextual embeddings in multimodal attention [24]. These methods generally do not model fine-grained interactions between image feature tokens or explicitly correct diversity-promoting updates to preserve fidelity. NegToMe linearly extrapolates attention tokens away from a reference image within the batch [25], but this update can push less similar tokens farther apart. Motion Modes [35] also uses repulsion to enhance diversity in video generation. It guides each new object motion away from previously generated motion fields. TokenRepel instead repels similar hidden tokens across images using angular feature geometry, then corrects the induced predictions toward a high-density half-space to preserve fidelity.

3 Enhancing Image Diversity

In this section, we present our algorithm for enhancing image diversity, as depicted in Figure 1a. Given a set of image token features, our goal is to have similar features repel each other. This can be visualized as an electron cloud (feature set) in which electrons (features) repel one another when too close but are otherwise unaffected. Like electrons around a nucleus, the features cannot move arbitrarily; they remain on spherical shells. This reflects the notion that the angle between two feature vectors captures their semantic similarity more faithfully than Euclidean distance [36, 25, 37]; thus, repulsion should change their angle rather than their magnitudes.



(a) Token repulsion

(b) Token density preservation

(c) Overall pipeline

Figure 1: Illustration of our TokenRepel algorithm. (a) Left: The tokens of Image 1 are repelled by taking a negative gradient step induced by the distribution formed from the tokens of the other images. (a) Right: Update formulation for $\mathbf{h}_{ij}$, where $\bar{\mu}(\hat{\mathbf{h}}_{ij})$ denotes the weighted mean of the other image tokens with respect to their cosine similarity to $\mathbf{h}_{ij}$. (b) Left: Contours representing the learned data distribution of the generative model, a green dashed arrow representing the original score, and a black solid arrow representing the score after applying token repulsion. (b) Right: The refinement process applied to $\nabla \log p_t^r(x_t)$ to project it into the higher-density half-space (red solid arrow). (c) Locations of these two operations inside a transformer layer.

Modeling cross-image token interactions. We model the normalized token features from all other images as a probability density on the unit sphere using a von Mises–Fisher (vMF) mixture model. Let $\mathbf{h}_{ij} \in \mathbb{R}^D$ denote the hidden state of the jth token in the ith image after a given attention layer, and let $\hat{\mathbf{h}}_{ij} = \mathbf{h}_{ij}/\|\mathbf{h}_{ij}\|$ be its unit-norm counterpart. Here, $i \in [1, B]$ and $j \in [1, N]$ for a batch of B images with N tokens per image. We model the density of the normalized hidden state features as an isotropic equal-weight von Mises–Fisher (vMF) mixture with probability density function

$$p(\hat{\mathbf{h}}_{ij}) = \frac{C_D(\kappa)}{(B-1)N} \sum_{k \neq i,\, \ell} \exp\big(\kappa\, \hat{\mathbf{h}}_{ij}^{\mathsf{T}} \hat{\mathbf{h}}_{k\ell}\big), \tag{1}$$

where $C_D(\kappa)$ is the vMF normalizing constant in D dimensions, κ is the concentration parameter, and the mixture means $\hat{\mathbf{h}}_{k\ell}$ are the $(B-1)N$ features from images other than I_i. This corresponds to a kernel density estimate of the distribution. The Riemannian gradient of the probability density with respect to $\hat{\mathbf{h}}_{ij}$ is

$$\nabla p(\hat{\mathbf{h}}_{ij}) = p(\hat{\mathbf{h}}_{ij}) \nabla \log p(\hat{\mathbf{h}}_{ij}) = \kappa p(\hat{\mathbf{h}}_{ij}) T(\bar{\boldsymbol{\mu}}(\hat{\mathbf{h}}_{ij}), \hat{\mathbf{h}}_{ij}) \tag{2}$$

where $T(\mathbf{a}, \mathbf{b}) = \mathbf{a} - (\mathbf{a}^{\mathsf{T}}\mathbf{b})\,\mathbf{b}$ is the tangent space projection operator and

$$\bar{\boldsymbol{\mu}}(\hat{\mathbf{h}}_{ij}) = \sum_{k \neq i, \ell} \frac{\exp\left(\kappa \hat{\mathbf{h}}_{ij}^{\mathsf{T}} \hat{\mathbf{h}}_{k\ell}\right)}{\sum_{m \neq i, n} \exp\left(\kappa \hat{\mathbf{h}}_{ij}^{\mathsf{T}} \hat{\mathbf{h}}_{mn}\right)} \hat{\mathbf{h}}_{k\ell} \tag{3}$$

is the softmax-weighted mean of the other features. Geometrically, $\nabla p(\hat{\mathbf{h}}_{ij})$ is proportional to the projection of $\bar{\boldsymbol{\mu}}(\hat{\mathbf{h}}_{ij}) - \hat{\mathbf{h}}_{ij}$ onto the tangent space at $\hat{\mathbf{h}}_{ij}$.

Token repulsion. In the high-dimensional token spaces of generative models (e.g., $D = 3072$ for FLUX.1-dev), the vMF density $p(\hat{\mathbf{h}}_{ij})$ is sharply concentrated and decays rapidly as the angular separation θ increases. Retaining interactions over a useful angular range therefore requires a large concentration κ. For example, placing the transition near $\theta = 45°$ in $D = 3072$ requires κ on the

order of $50{,}000$. At this concentration, the density is effectively step-like, falling from a large value to nearly zero, while the softmax-weighted mean $\bar{\boldsymbol{\mu}}(\hat{\mathbf{h}}_{ij})$ approaches argmax. We therefore closely approximate the update by a uniform repulsion from the most similar cross-image token when their angular separation is below a threshold $\bar{\theta}$, and no update otherwise.

To do so, we leverage two geometrically meaningful parameters: the token rotation angle δ and the angular threshold $\bar{\theta}$, instead of the step size λ and concentration κ. The token update is formulated as:

$$\mathbf{h}'_{ij} = \hat{\mathbf{h}}_{ij} - \lambda \nabla p(\hat{\mathbf{h}}_{ij}) \approx \hat{\mathbf{h}}_{ij} - \tan(\delta)\mathbf{1}[\hat{\mathbf{h}}_{ij}^{\mathsf{T}}\hat{\mathbf{h}}_{k^\star \ell^\star} > \cos(\bar{\theta})]T(\hat{\mathbf{h}}_{k^\star \ell^\star}, \hat{\mathbf{h}}_{ij})/\|T(\hat{\mathbf{h}}_{k^\star \ell^\star}, \hat{\mathbf{h}}_{ij})\|, \quad (4)$$

where $k^\star, \ell^\star = \arg\max_{k \neq i,\, \ell} \hat{\mathbf{h}}_{ij}^{\mathsf{T}}\hat{\mathbf{h}}_{k\ell}$. If $\theta = 0$, the two tokens coincide, the norm is zero and we replace this term with a random unit direction in the tangent space at $\hat{\mathbf{h}}_{ij}$. We also cap the rotation so that $\bar{\theta} + \delta \leq 90^\circ$, preventing the update from overshooting orthogonality.

4 Preserving Image Fidelity

An unconstrained token repulsion can improve diversity but may move predictions away from the model's learned distribution, reducing image quality and prompt adherence. To preserve fidelity, we introduce a token density preservation mechanism that prevents repulsion-induced updates from pointing away from high-density regions of latent space. The pretrained flow-matching model implicitly encodes a direction toward these regions [38–40], which we infer from its model predictions and use to restrict the token movement.

Token density preservation. Ideally, we would compare the score of the original distribution, $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c)$, with the score after token repulsion, $\nabla_{\mathbf{x}_t} \log p_t^r(\mathbf{x}_t \mid c)$. Here, $\mathbf{x}_t$ is the latent at time t and c is the text condition. We seek to prevent the repulsion-induced change $\mathbf{s}^r - \mathbf{s}$ from opposing a direction of increasing density. However, a flow-matching model predicts a velocity $\mathbf{v}_t(\mathbf{x}_t, c)$ rather than either score directly. For the model's probability path, the score can be expressed as a time-dependent linear combination of the velocity and the latent,

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c) \propto \mathbf{v}_t(\mathbf{x}_t, c) + a_t \mathbf{x}_t, \quad (5)$$

where a_t depends only on time. At the same $\mathbf{x}_t$, the latent term cancels when we subtract two predictions, making velocity differences proportional to score differences. We therefore define

$$\Delta^r \triangleq \mathbf{v}_t^r(\mathbf{x}_t, c) - \mathbf{v}_t(\mathbf{x}_t, c) \propto \nabla_{\mathbf{x}_t} \log p_t^r(\mathbf{x}_t \mid c) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c), \quad (6)$$

$$\Delta^h \triangleq \mathbf{v}_t(\mathbf{x}_t, c) - \mathbf{v}_t(\mathbf{x}_t, c_{\text{neg}}) \propto \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c_{\text{neg}}), \quad (7)$$

where $\mathbf{v}_t^r$ is the velocity prediction with TokenRepel and $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid c_{\text{neg}})$ is the score under a negative prompt c_{neg}. Thus, Δ^r captures the repulsion-induced change, while Δ^h provides a direction toward higher density under the desired condition relative to the negative prompt. At the token level, Δ^r and Δ^h become more negatively aligned as the repulsion angle increases, as shown in Figure 12, motivating the following correction mechanism.

For the jth token of the ith image, we constrain the repulsion-induced change to the half-space $\mathcal{R}_{ij} = \{\mathbf{d} : \langle \mathbf{d}, \Delta_{ij}^h \rangle \geq 0\}$. Its boundary is orthogonal to Δ_{ij}^h, as illustrated in Figure 1b. As derived in Section E, if Δ_{ij}^r points outside this half-space, we remove its component opposing Δ_{ij}^h; otherwise, we leave it unchanged:

$$\bar{\Delta}_{ij}^r = \begin{cases} \Delta_{ij}^r - \langle \Delta_{ij}^r, \widehat{\Delta}_{ij}^h \rangle \, \widehat{\Delta}_{ij}^h & \text{if } \langle \Delta_{ij}^r, \Delta_{ij}^h \rangle < 0, \\ \Delta_{ij}^r & \text{otherwise,} \end{cases} \quad (8)$$

where $\widehat{\Delta}_{ij}^h = \Delta_{ij}^h / \|\Delta_{ij}^h\|$. This is the Euclidean projection of Δ_{ij}^r onto $\mathcal{R}_{ij}$. Denoting this constraint by ϕ, the updated velocity is

$$\tilde{\mathbf{v}}^r = \phi(\mathbf{v}^r(\mathbf{x}, c), \mathbf{v}(\mathbf{x}, c), \mathbf{v}(\mathbf{x}, c_{\text{neg}})) = \mathbf{v}(\mathbf{x}, c) + \bar{\Delta}^r, \quad (9)$$

We use $\tilde{\mathbf{v}}^r$ to obtain the next latent $\mathbf{x}_{t-1}$. Thus, the velocity prediction after applying token repulsion is constrained to the high-density half-space. For ϵ-prediction models such as SDXL, the same constraint formulation applies. It better steers the token updates to the high-density region while enabling sufficient diversity.

Layerwise density preservation. Applying token density preservation only at the final block output (velocities) preserves image quality, but repulsion-induced errors can accumulate excessively across layers. We therefore correct the intermediate hidden states as repulsion proceeds, using smaller, more frequent adjustments to guide them toward high-density regions.

Given a model with M transformer layers, the latent embedding $\mathbf{x}$ is first projected into a higher-dimensional hidden space, passed through the M layers, and finally projected back to the dimension of $\mathbf{x}$ to predict the velocity. The process can be written as

$$\mathbf{u}_1 = f_{\uparrow}(\mathbf{x}), \qquad \mathbf{u}_m = \mathbf{u}_{m-1} + g_m(\mathbf{u}_{m-1}), \qquad \mathbf{v} = f_{\downarrow}(\mathbf{u}_M) \tag{10}$$

where $f_{\uparrow}$ and $f_{\downarrow}$ are the up-projection and down-projection transformations, and g_m represents the mth transformer layer. Each transformer layer produces the next hidden state through a residual connection. Thus, each successive hidden state can be viewed as a closer approximation of $\mathbf{u}_M$, which is mapped to the velocity $\mathbf{v}$.

Therefore, for each layer, we maintain three parallel streams: one using the provided prompt, another using the negative prompt, and a third using the provided prompt with TokenRepel enabled. We denote their hidden states after the mth layer by $\mathbf{u}_m^+$, $\mathbf{u}_m^-$, and $\mathbf{u}_m^r$, respectively. Using the constraint operator ϕ defined above, we replace $\mathbf{u}_m^r$ with

$$\tilde{\mathbf{u}}_m^r = \phi(\mathbf{u}_m^r, \mathbf{u}_m^+, \mathbf{u}_m^-). \tag{11}$$

This layerwise preservation better steers token updates towards high-density regions while preserving sufficient diversity.

5 Experiments

We evaluate TokenRepel on three representative T2I models: FLUX.1-dev [1] (flow matching), FLUX.2-Klein-9B [26] (few-step distilled flow matching), and SDXL [4] (diffusion). We report the main results on the GenEval dataset [41] and conduct ablation studies and analyses on the DrawBench dataset [42].

Implementation details. The token repulsion process is applied to the attention outputs inside each transformer layer, while the token density preservation is applied to the output of each layer. For FLUX.1-dev and FLUX.2-Klein-9B, we use 35 and 4 inference steps, respectively, and apply TokenRepel in single-stream layers during the first 5 steps and the first step, respectively. For SDXL, we adopt the RunDiffusion/Juggernaut-XI-v11 checkpoint with 50 steps, applying TokenRepel in the first 5 steps of the conditional branch. We generate 6 images per prompt under the same seed for each prompt across all the experiments. Further details on parameter settings are provided in Section C.1.

Evaluation metrics. For diversity evaluation, we use three metrics: (1) DreamSim [43], trained on human-labeled data to measure visual differences, where we use the ensemble version (CLIP, DINO, OpenCLIP; all ViT-B/16) and report averaged pairwise distances per prompt batch; (2) Vendi Score [44], which computes the Shannon entropy of eigenvalues from the cosine similarity matrix of InceptionV3 features; and (3) LPIPS [45] (VGG backbone), measuring pairwise perceptual distances within each batch. For image quality, we use multiple reward models capturing different aspects: (1) HPSv3 [46] (Qwen7B-based), evaluating overall quality and prompt alignment; (2) ImageReward [47] (BLIP-based), assessing overall quality and alignment; (3) CLIP [48] (ViT-H/14), focusing on prompt adherence; (4) PickScore [49], a CLIP-based human preference model; and (5) GenEval score [41], targeting compositionality and constraint satisfaction using a detection model.

5.1 Results

Diversity–quality trade-off analysis on DrawBench. We adapt existing open-source methods to all three generative models for comparison, including CADS [11], SPARK [21], PG [22], NegToMe [25], Group-Inference [32] and Contextual-Repulsion [24]. All of the above have a diversity–quality trade-off hyperparameter. In Figure 2, we plot DreamSim against HPSv3 as the hyperparameters vary for each method on the DrawBench dataset.

For SDXL, SPARK and PG improve diversity while maintaining good quality, but cannot go beyond a certain diversity level without the image quality dropping sharply. Other methods are more stable,

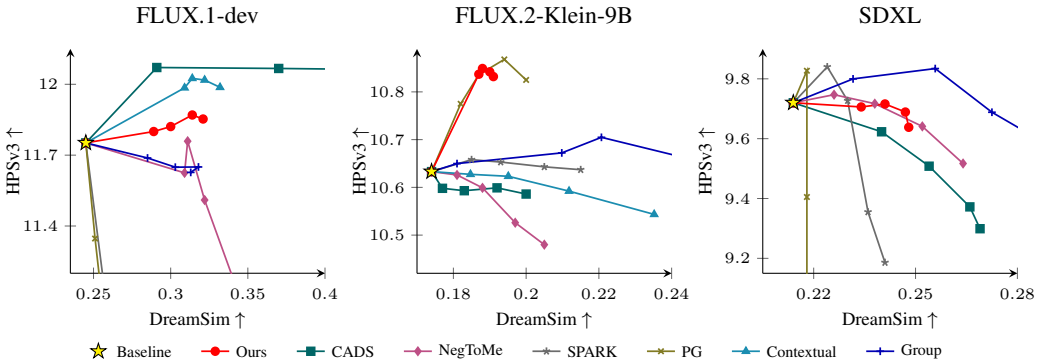


Figure 2: DreamSim versus HPSv3 on DrawBench for three T2I models, under varying diversity–quality trade-off hyperparameters for each method. The same trade-off comparison on the larger GenEval dataset is shown in Figure 10.

with ours and NegToMe outperforming CADS. For FLUX.2-Klein-9B, all methods remain relatively stable, but ours and PG achieve visibly better trade-offs. For FLUX.1-dev, SPARK and PG are less effective; our method outperforms NegToMe but is below CADS and Contextual-Repulsion. However, we observe that CADS and Contextual-Repulsion often weaken prompt adherence and introduce off-prompt content. Since it perturbs text embeddings rather than image features, outputs remain within the natural image domain, but alignment with the prompt is degraded. When isolating prompt alignment using CLIP instead of HPSv3, CADS shows a significant drop, whereas our method maintains performance close to the baseline. In general, our method performs consistently across different settings and is not overly sensitive to hyperparameter variations.

GenEval test results. Based on the DrawBench validation results, we select the best hyperparameter setting for each method using the heuristic score $3 \times \text{DreamSim} + \text{HPSv3}$. We then evaluate each method on the GenEval test dataset [41]. The results are presented in Table 1. Our method achieves substantially higher diversity than the baseline while maintaining strong image quality compared to prior approaches. For each model, we rank methods with reported results on each of the three diversity and five fidelity metrics, with higher values receiving higher ranks. We then average the ranks within each group and sum the two averages, giving diversity and fidelity equal weight in the merged rank. The results show that, for SDXL, our TokenRepel method without fidelity preservation achieves the best overall trade-off. For the FLUX models, the layerwise fidelity-preserving variant achieves the highest merged rank. In Figures 7 to 9, we present qualitative results generated from identical initial noise samples. The results demonstrate that our method effectively reduces duplicated visual modes and produces more diverse images while maintaining a comparable level of image quality. We also report per-image averaged latency relative to the baseline. The correction variants incur additional computation because they need extra forward passes during the early sampling steps where TokenRepel is applied.

5.2 Ablations and Analysis

Effect of hyperparameters. We analyze the impact of the token rotation angle δ and the angular threshold $\bar{\theta}$. The former controls the repulsion strength, where larger δ increases the angular separation between similar features, and $\delta = 0$ reduces to the baseline. The latter sets the angle below which features are repelled, where $\bar{\theta} = 0$ reduces to the baseline. We report sweeps over both parameters for our method without fidelity preservation in Figure 3. Increasing δ and $\bar{\theta}$ increases diversity at the cost of quality. In addition, a higher concentration κ in Equation (3) increases diversity with a slight drop in quality. In this paper, we use the zero-temperature setting, which converts softmax to argmax to select the most similar token, yielding a simpler design.

On fidelity preservation. We compare TokenRepel without token density preservation, with blockwise preservation applied only to the final block output (B), and with layerwise preservation (L) across all three models in Figure 5. For the FLUX models, at comparable diversity levels, layerwise preservation generally achieves higher HPSv3 than the other variants. For SDXL, this effect is less pronounced. In the appendix, Figure 11 further shows that correcting more layers improves fidelity

Table 1: Comparison of diversity and image fidelity metrics on GenEval, using the best hyperparameters selected on DrawBench. The last column is the merged rank, summing the equal-weighted average diversity and fidelity ranks. 'Ours w/o P' denotes TokenRepel without density preservation, while 'Ours-B' and 'Ours-L' refer to the blockwise and layerwise preservation strategies. A quality-matched comparison with settings selected on GenEval is given in Table 4.

Method	Latency over baseline↓	Diversity			Image Fidelity					Merged rank↑
		DreamSim↑	Vendi↑	LPIPS↑	HPSv3↑	CLIP↑	ImR↑	PickS↑	GenEval↑	
FLUX.1-dev:										
Baseline	1.00×	0.265	2.166	0.612	8.944	0.359	0.919	**23.406**	**0.652**	9.80
CADS	1.00×	**0.446**	**2.634**	0.755	7.864	0.310	0.338	22.623	0.488	11.67
SPARK	1.00×	0.355	2.415	**0.796**	5.258	0.328	0.562	21.906	0.476	9.57
PG	1.00×	0.271	2.203	0.625	8.498	**0.361**	0.899	23.291	0.648	10.00
NegToMe	1.02×	0.336	2.367	0.707	8.939	0.356	0.890	23.260	0.629	10.73
Group	2.89×	0.347	2.440	0.674	7.362	0.360	0.925	23.255	0.646	12.37
Contextual	1.23×	0.342	2.344	0.701	7.501	0.360	**0.932**	23.268	0.645	11.47
Ours w/o P	1.01×	0.400	2.543	0.742	8.653	0.344	0.722	23.088	0.589	12.27
Ours-B	1.25×	0.342	2.388	0.689	8.853	0.352	0.837	23.249	0.621	9.53
Ours-L	1.25×	0.355	2.368	0.720	**9.046**	0.352	0.866	23.251	0.642	**12.60**
FLUX.2-Klein-9B:										
Baseline	1.00×	0.196	1.892	0.581	9.092	0.382	1.337	23.757	0.860	8.67
CADS	1.00×	0.209	1.944	0.600	9.107	0.381	1.323	23.758	0.840	9.07
SPARK	1.00×	0.245	2.001	0.673	9.123	0.378	1.295	23.596	0.838	11.27
PG	1.00×	0.218	1.966	0.636	9.290	0.376	1.327	23.613	0.843	10.73
NegToMe	1.00×	0.209	1.932	0.620	9.170	0.382	1.329	23.766	0.854	11.07
Group	1.99×	0.219	1.959	0.591	8.962	**0.386**	1.379	23.649	0.864	12.13
Contextual	3.52×	0.214	1.820	0.541	8.952	**0.386**	**1.381**	23.653	**0.870**	9.00
Ours w/o P	1.01×	**0.262**	**2.093**	**0.675**	9.036	0.380	1.275	23.616	0.824	12.20
Ours-B	1.49×	0.235	2.019	0.654	9.236	0.381	1.309	23.713	0.839	**13.53**
Ours-L	1.50×	0.217	1.950	0.639	**9.387**	0.381	1.331	**23.786**	0.850	13.47
SDXL:										
Baseline	1.00×	0.229	2.077	0.633	7.613	0.385	1.069	23.340	0.616	7.80
CADS	1.00×	0.250	2.153	0.635	7.525	0.384	1.042	23.303	0.617	8.00
SPARK	1.00×	0.250	2.098	0.696	7.851	0.384	1.037	23.251	0.619	9.80
PG	1.00×	0.241	2.098	**0.709**	**7.920**	0.381	1.082	23.014	0.614	8.53
NegToMe	1.00×	0.261	2.184	0.677	7.460	0.383	1.048	23.234	0.615	9.20
Group	10.38×	**0.267**	2.176	0.669	7.488	**0.388**	**1.155**	**23.360**	**0.659**	**13.93**
Ours w/o P	1.01×	0.266	**2.185**	0.685	7.539	0.384	1.061	23.266	0.617	12.60
Ours-B	1.08×	0.260	2.167	0.676	7.566	0.384	1.060	23.289	0.620	11.67
Ours-L	1.09×	0.256	2.157	0.672	7.561	0.384	1.058	23.290	0.608	8.47

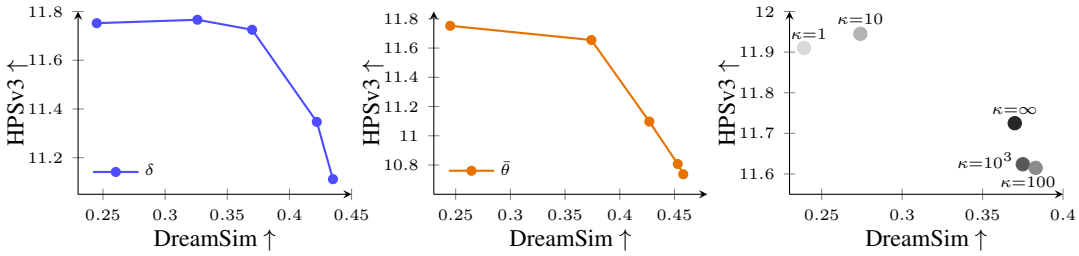


Figure 3: Analysis of the diversity–quality trade-off as the token rotation angle δ, the angular threshold $\bar{\theta}$, and the concentration κ vary, where $\kappa = \infty$ corresponds to the argmax formulation used in the main results, for TokenRepel without fidelity preservation on DrawBench using FLUX.1-dev.

but reduces diversity. Correcting only the early layers can better balance fidelity and diversity while saving computation, as the two auxiliary branches can stop early.



Figure 4: One example of fidelity preservation on FLUX.2-Klein-9B. Left: generated images of baseline, TokenRepel without preservation (w/o P), and TokenRepel with layerwise preservation (L). Without preservation, TokenRepel increases diversity but introduces artifacts. Right: heatmaps of the top principal component of the features at the 15th layer. Layerwise preservation keeps the feature distribution closer to the baseline.

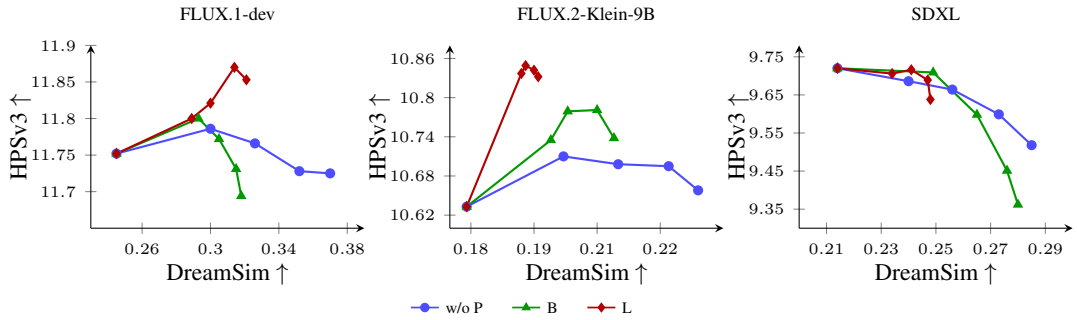


Figure 5: Ablation study comparing the diversity (DreamSim) and quality (HPSv3) trade-off on DrawBench for our method without token density preservation (w/o P), with blockwise preservation (B), and with layerwise preservation (L) across all three models, respectively, under increasing values of the repulsion angle δ.

Table 2: Batch-size ablation of TokenRepel on DrawBench using FLUX.1-dev. Percentage increases over the baseline at the same batch size, using matched seeds, are shown in parentheses.

Batch size	Diversity			Image fidelity			
	DreamSim↑	Vendi↑	LPIPS↑	HPSv3↑	CLIP↑	ImR↑	PickS↑
2	0.317 (+25.3%)	1.483 (+3.1%)	0.655 (+9.3%)	11.936	0.348	0.993	22.828
4	0.312 (+18.2%)	2.030 (+4.7%)	0.655 (+8.1%)	12.216	0.345	1.002	22.783
6	0.306 (+18.6%)	2.352 (+6.2%)	0.654 (+8.6%)	12.315	0.347	1.016	22.824
8	0.309 (+19.8%)	2.601 (+6.9%)	0.662 (+8.3%)	12.253	0.345	0.990	22.787
10	0.304 (+17.8%)	2.802 (+7.4%)	0.660 (+7.8%)	12.231	0.346	0.993	22.787
12	0.305 (+18.2%)	2.955 (+7.3%)	0.666 (+7.9%)	12.230	0.345	0.993	22.785

Batch size analysis. We evaluate TokenRepel at batch sizes from 2 to 12 on DrawBench using FLUX.1-dev and matched seeds. Table 2 reports all seven metrics, with diversity gains measured against the baseline at each batch size. The gains in pairwise DreamSim and LPIPS remain positive and generally decrease slightly as the batch size grows, while the relative gain in the set-level Vendi Score rises. The image-quality metrics remain relatively stable across batch sizes. DreamSim is a pairwise metric; larger batches include more image pairs, increasing the likelihood that some pairs converge to similar modes after a single-step repulsion. The Vendi Score estimates the effective number of distinct samples in a set; therefore, its absolute value increases with set size.

Number of repulsion steps. We examine the effectiveness of TokenRepel across different repulsion windows. As shown in the left and middle panels of Figure 6, extending this window consistently improves diversity, but the gains quickly saturate while HPSv3 remains stable. To explain this

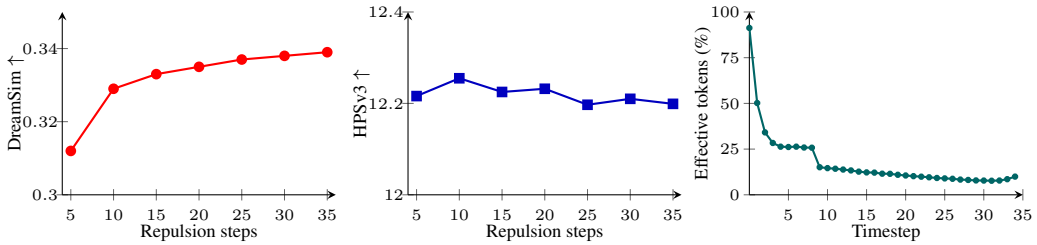


Figure 6: Effect of the number of repulsion steps (from 5 to 35) on DreamSim (left) and HPSv3 (middle), and the fraction of tokens eligible for repulsion at each timestep with $\cos\bar{\theta} = 0.7$ (right), measured with FLUX.1-dev on DrawBench.



(a) A photo of a bear above a clock.



(b) A photo of a pink parking meter.



(c) A photo of an orange TV.

Figure 7: FLUX.1-dev qualitative results under the same initial noise. Left: baseline; right: ours.

Table 3: Image-editing results on KontextBench using FLUX Kontext.

Method	DreamSim↑	Vendi↑	LPIPS↑	HPSv3↑	CLIP↑	ImageReward↑	PickScore↑
Baseline	0.098	1.425	0.320	**9.384**	**0.364**	**0.582**	**21.777**
Ours	**0.209**	**1.699**	**0.502**	9.301	0.357	0.509	21.591

saturation, we measure the fraction of image tokens whose cross-image similarity exceeds the repulsion threshold at each timestep, averaged across layers and prompts. As shown in the right panel of Figure 6, this fraction falls from around 90% at the first step and remains around 8% near the end of sampling. Early steps shape coarse global structure and contain more similar cross-image patches, so repulsion affects more tokens. Later steps refine details, leaving fewer similar tokens to repel. We therefore apply repulsion only during the first few steps, capturing most of the diversity gain while reducing the additional computation.

5.3 Discussion

Generalization to image editing. We evaluate TokenRepel on KontextBench, which contains 1,026 image–instruction pairs from 108 source images. For evaluation, an LLM generates the target prompt from the source image and editing instruction. We generate four images for each image–instruction pair using FLUX Kontext. As shown in Table 3, TokenRepel substantially improves output diversity, with only a small drop in editing quality. Its effectiveness depends on the instruction. It offers limited benefit for deterministic edits, such as object removal, but works well for open-ended edits, such as "add a cowboy hat" or generating content in the reference style.



(a) A photo of a bench.



(b) A photo of a handbag.



(c) A photo of a frisbee and a vase.

Figure 8: FLUX.2-Klein qualitative results under the same initial noise. Left: baseline; right: ours.



(a) A photo of a blue clock.



(b) A photo of a red giraffe.



(c) A photo of a horse on the couch.

Figure 9: SDXL qualitative results under the same initial noise. Left: baseline; right: ours.

Limitations. Although our method enhances diversity, the upper bound of achievable diversity is still fundamentally limited by the support of the training data distribution learned by the model. For example, our method is significantly more effective on the FLUX models than on SDXL. Quality preservation also varies across metrics (Table 1): it is more pronounced for HPSv3 but less effective for GenEval and PickScore. In addition, the proposed quality-preserving mechanism is not directly applicable to autoregressive image generative models [6], since their predictions correspond to probabilities rather than gradient-like score estimates.

6 Conclusion

We have presented TokenRepel, a test-time approach for enhancing the visual diversity of generated image sets under identical prompts. First, we outlined a token repulsion mechanism for encouraging patch-wise image diversity by taking a negative Riemannian gradient step on a distribution of features on the unit hypersphere. Second, we showed how to correct these perturbations so that they remain in the high-density region learned by the generative model and extend this strategy within the transformer layers. Our approach demonstrates substantial improvements in diversity of the output images across three T2I models, while largely preserving the average sample fidelity.

Acknowledgements. We thank Xingjian Leng, Jaskirat Singh, and Zhengyang Yu for valuable discussions. Dr Campbell is the recipient of an Australian Research Council Discovery Early Career Award (project number DE250100542) funded by the Australian Government.

References

[1] B. F. Labs. Flux.1 [dev]. `https://blackforestlabs.ai/`, 2024.

[2] 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.

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

[4] 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.

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

[6] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

[8] Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. *arXiv preprint arXiv:2512.21514*, 2025.

[9] Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. *arXiv preprint arXiv:2510.22319*, 2025.

[10] Ye Zhu, Kaleb S Newman, Johannes F Lutzeyer, Adriana Romero-Soriano, Michal Drozdzal, and Olga Russakovsky. Gass: Geometry-aware spherical sampling for disentangled diversity enhancement in text-to-image generation. *arXiv preprint arXiv:2602.17200*, 2026.

[11] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024.

[12] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[13] Yunhao Chen, Zihui Yan, and Yunjie Zhu. A comprehensive survey for generative data augmentation. *Neurocomputing*, 2024.

[14] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[15] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

[16] Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[17] Lianghe Shi, Meng Wu, Huijie Zhang, Zekai Zhang, Molei Tao, and Qing Qu. A closer look at model collapse: From a generalization-to-memorization perspective. In *Advances in Neural Information Processing Systems*, 2025.

[18] Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. Prompt expansion for adaptive text-to-image generation. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[19] Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[20] Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. In *International Conference on Learning Representations*, 2026.

[21] Mohammad Jalali, LEI Haoyu, Amin Gohari, and Farzan Farnia. Sparke: Scalable prompt-aware diversity and novelty guidance in diffusion models via rke score. In *Advances in Neural Information Processing Systems*, 2025.

[22] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. In *International Conference on Learning Representations*, 2024.

[23] Pradhaan S. Bhat, Rishubh Parihar, Abhijnya Bhat, and R. Venkatesh Babu. Don't settle at the mode! mitigating diversity collapse in pretrained flow models via feature self-guidance. In *European Conference on Computer Vision*, 2026.

[24] Omer Dahary, Benaya Koren, Daniel Garibi, and Daniel Cohen-Or. On-the-fly repulsion in the contextual space for rich diversity in diffusion transformers. In *ACM SIGGRAPH Conference Papers*, 2026.

[25] Jaskirat Singh, Lindsey Li, Weijia Shi, Ranjay Krishna, Yejin Choi, Pang Wei Koh, Michael F Cohen, Stephen Gould, Liang Zheng, and Luke Zettlemoyer. Negative token merging: Image-based adversarial feature guidance. *arXiv preprint arXiv:2412.01339*, 2024.

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

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.

[28] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

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

[30] Dmitrii Sorokin, Maksim Nakhodnov, Andrey Kuznetsov, and Aibek Alanov. Imagerefl: Balancing quality and diversity in human-aligned diffusion models. *arXiv preprint arXiv:2505.22569*, 2025.

[31] Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

[32] Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. In *International Conference on Learning Representations*, 2026.

[33] Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. In *International Conference on Learning Representations*, 2026.

[34] Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *International Conference on Machine Learning*, 2025.

[35] Karran Pandey, Yannick Hold-Geoffroy, Matheus Gadelha, Niloy J. Mitra, Karan Singh, and Paul Guerrero. Motion modes: What could happen next? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[36] Liao Shen, Tianqi Liu, Huiqiang Sun, Xinyi Ye, Baopu Li, Jianming Zhang, and Zhiguo Cao. Dreammover: Leveraging the prior of diffusion models for image interpolation with large motion. In *European Conference on Computer Vision*, 2024.

[37] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, 2023.

[38] 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.

[39] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[40] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

[41] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[43] Stephanie Fu, Netanel Y Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, 2023.

[44] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

[45] 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.

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

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

[48] 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.

[49] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Appendix: TokenRepel: Generating Diverse Image Sets Without Compromising Quality

A Full GenEval Trade-off Comparison

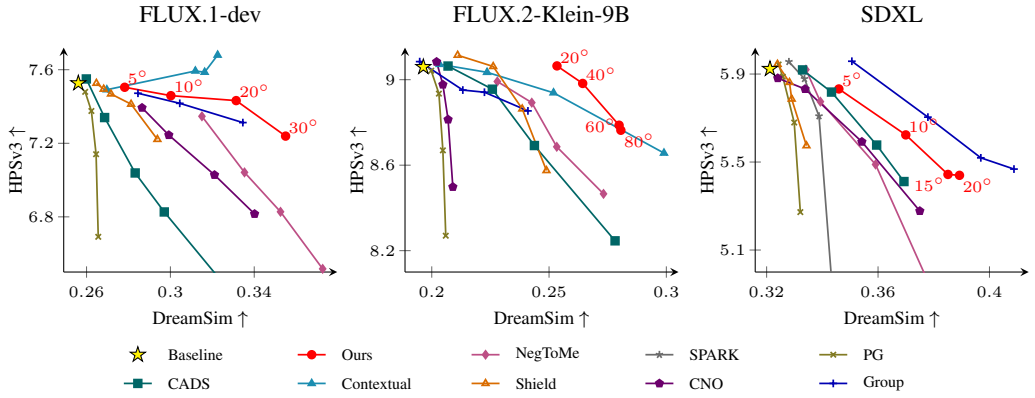


Figure 10: Diversity–quality trade-offs on GenEval across three models. Each curve varies one method-specific hyperparameter; the yellow star marks the model baseline, and red labels give the TokenRepel rotation angle δ.

We provide the full sweep comparison between TokenRepel and related methods on GenEval, including CADS [11], SPARKE [21], Particle Guidance (PG) [22], NegToMe [25], Contextual Repulsion [24], Group Inference [32], Shielded Diffusion [34], and Contrastive Noise Optimization (CNO) [33]. For each method, we sweep the hyperparameter that controls the diversity–quality trade-off on GenEval for FLUX.1-dev, FLUX.2-Klein-9B, and SDXL. The results are shown in Figure 10.

Implementation details. We sweep δ and fix the remaining settings for each model. For FLUX.1-dev, we use $\bar{\theta} = 45°$, apply repulsion in steps 1–3, and apply preservation in the first 30 layers. For FLUX.2-Klein-9B, we use $\bar{\theta} = 60°$, apply repulsion in step 1, and apply preservation in the first 20 layers. For SDXL, we use $\bar{\theta} = 45°$, apply repulsion in steps 1–5, and apply preservation after the down and middle layers. We generate 4 images per prompt at 512×512 with seed 0 and share the initial noise across methods. FLUX.1-dev uses 25 sampling steps. This setup differs slightly from the main experiments due to our limited computational resources, which generate 6 images per prompt, use 1024×1024 resolution for SDXL, and use 35 sampling steps for FLUX.1-dev.

Comparison results. Figure 10 shows the trade-off curves using DreamSim and HPSv3, our most comprehensive diversity and image quality metrics, respectively. We sweep four hyperparameter values for each method, with the HPSv3 drop roughly capped at 1.0. TokenRepel gives the best trade-off on FLUX.2-Klein-9B and the second-best trade-off on FLUX.1-dev and SDXL. Its performance is also the most stable across the three models.

Quality-matched comparison. To compare diversity at matched image quality directly at Geneval, we select for each method the sweep point in Figure 10 that maximizes DreamSim while keeping HPSv3 above a model-specific threshold. The threshold reflects the variation of baseline HPSv3 across random seeds. Table 4 reports all metrics at the selected settings. TokenRepel achieves the highest DreamSim on both FLUX models and the second-highest on SDXL, behind Group Inference

Table 4: Quality-matched comparison on GenEval. For each method, we select the sweep point in Figure 10 with the highest DreamSim whose HPSv3 stays above 7.4, 8.8, and 5.5 for FLUX.1-dev, FLUX.2-Klein-9B, and SDXL, respectively.

Method	Latency over baseline↓	Diversity			Image Fidelity				
		DreamSim↑	Vendi↑	LPIPS↑	HPSv3↑	CLIP↑	ImR↑	PickS↑	GenEval↑
FLUX.1-dev:									
Baseline	1.00×	0.2561	1.8747	0.5821	7.5266	0.3613	0.9579	**23.3245**	0.6653
CADS	1.00×	0.2599	1.8694	0.5869	7.5498	0.3608	0.9690	23.3233	0.6551
PG	1.00×	0.2593	1.8786	0.6019	7.4796	0.3606	0.9798	23.2855	0.6598
NegToMe	1.01×	0.3151	1.9829	0.6572	7.3467	0.3616	1.0055	23.1479	0.6670
CNO	1.13×	0.2865	1.9288	0.6416	7.3931	**0.3642**	**1.0084**	23.3135	0.6671
Shield	1.00×	0.2813	1.9112	0.6221	7.4127	0.3624	0.9734	23.2872	0.6633
Group	1.54×	0.3045	1.9727	0.6325	7.4173	0.3620	0.9264	23.2908	**0.6723**
Contextual	1.28×	0.3225	1.9793	**0.6994**	**7.6795**	0.3533	0.9296	23.2449	0.6404
Ours	1.15×	**0.3314**	**2.0090**	0.6519	7.4321	0.3619	0.9603	23.2617	0.6619
FLUX.2-Klein-9B:									
Baseline	1.00×	0.1964	1.6791	0.5495	**9.0595**	0.3876	**1.3912**	**23.6764**	**0.8813**
CADS	1.01×	0.2258	1.7567	0.5711	8.9555	0.3859	1.3712	23.6665	0.8324
PG	1.01×	0.2031	1.6799	0.5787	8.9348	0.3855	1.3889	23.4989	0.8632
NegToMe	1.00×	0.2426	1.7955	0.6373	8.8926	0.3857	1.3454	23.6210	0.8504
CNO	1.97×	0.2070	1.7079	0.6032	8.8135	**0.3891**	1.3840	23.5130	0.8672
Shield	1.01×	0.2386	1.7401	**0.6735**	8.8634	0.3887	1.3902	23.3728	0.8698
Group	1.99×	0.2410	1.7915	0.6157	8.8540	0.3861	1.3517	23.6074	0.8540
Contextual	3.52×	0.2519	1.8037	0.6093	8.9387	0.3859	1.3736	23.6401	0.8573
Ours	1.32×	**0.2643**	**1.8280**	0.6599	8.9825	0.3860	1.3430	23.6246	0.8525
SDXL:									
Baseline	1.00×	0.3212	1.9565	0.7243	**5.9234**	0.3828	0.9481	**22.9723**	0.5539
CADS	1.00×	0.3595	2.0310	0.7264	5.5772	0.3786	0.8433	22.8572	0.5464
SPARKE	1.00×	0.3388	1.9795	**0.7713**	5.7089	0.3797	0.9398	22.8326	0.5477
PG	1.00×	0.3299	1.9598	0.7531	5.6796	0.3811	**0.9589**	22.7121	0.5605
NegToMe	1.00×	0.3393	2.0042	0.7459	5.7745	0.3822	0.9311	22.8969	0.5638
CNO	1.09×	0.3542	2.0364	0.7587	5.5920	0.3820	0.9154	22.8427	0.5626
Shield	1.00×	0.3343	1.9852	0.7604	5.5742	0.3819	0.9180	22.7516	0.5587
Group	10.38×	**0.3969**	**2.1468**	0.7583	5.5191	**0.3838**	0.9303	22.8413	**0.5784**
Ours	1.05×	0.3700	2.0863	0.7480	5.6233	0.3804	0.8938	22.7996	0.5745

with $10.38\times$ latency, while keeping the other image fidelity metrics close to the baseline. The absolute values differ from Table 1 because the GenEval sweep uses a slightly different setup.

B Pseudocode

Algorithm 1 gives PyTorch-style pseudocode for token repulsion (Section 3) and token density preservation (Equation (8)).

C Additional Implementation Details

The negative prompt used for all models is: *"blurry, low quality, unrealistic, distorted, deformed, bad anatomy, ugly, watermark, oversaturated, flat background, simple environment, cartoon, anime, low detail, disfigured, disgusting, noise, overexposed, underexposed, compressed artifacts, plastic skin, airbrushed, doll-like, fake, low resolution, illustration, out of focus."* The SDXL checkpoint is RunDiffusion/Juggernaut-XI-v11.

C.1 Parameter Settings

We denote the TokenRepel repulsion strength by $\alpha = \tan(\delta)$ and its cosine-similarity threshold by $\beta = \cos(\bar{\theta})$.

Default settings. For the hyperparameter analysis in Figure 3, we sweep $\alpha \in \{0, 0.4, 0.8, 1.2, 1.6\}$. The selected TokenRepel settings are:

Algorithm 1: TokenRepel with layerwise token density preservation

```
def TokenRepel(h, delta, theta_bar):
    # h: [B, N, D] image tokens after an attention layer
    # delta: token rotation angle
    # theta_bar: maximum angle for receiving repulsion
    B, N, D = h.shape
    h = h.reshape(B * N, D)
    h_hat = F.normalize(h, dim=-1)

    # 1) Most similar token from the other images
    image_id = torch.arange(B).repeat_interleave(N)
    cosine = h_hat @ h_hat.T
    cosine[image_id[:, None] == image_id[None, :]] = -inf
    max_cos, idx = cosine.max(dim=-1)
    max_cos = max_cos[:, None]
    h_ref = h_hat[idx]

    # 2) Unit tangent direction pointing away from h_ref
    # (a random tangent direction if h_hat and h_ref coincide)
    t = F.normalize(max_cos * h_hat - h_ref, dim=-1)

    # 3) Rotate on the unit sphere without passing 90 degrees
    h_new = h_hat + math.tan(delta) * t
    overshoot = (h_new * h_ref).sum(-1, keepdim=True) < 0
    h_new = torch.where(overshoot, h_hat - max_cos * h_ref, h_new)
    h_new = F.normalize(h_new, dim=-1) * h.norm(dim=-1, keepdim=True)

    # 4) Repel only tokens above the similarity threshold
    h_new = torch.where(max_cos > math.cos(theta_bar), h_new, h)
    return h_new.reshape(B, N, D)

def Preserve(u_r, u_pos, u_neg):
    # u_r, u_pos, u_neg: [B, N, D] outputs with TokenRepel,
    # with prompt c, and with negative prompt c_neg
    d_r = u_r - u_pos # repulsion-induced change
    d_h = F.normalize(u_pos - u_neg, dim=-1) # high-density direction
    proj = (d_r * d_h).sum(-1, keepdim=True)
    d_r = torch.where(proj < 0, d_r - proj * d_h, d_r) # half-space projection
    return u_pos + d_r
```

- FLUX.1-dev: $\alpha = 0.8$, $\beta = 0.7$.
- FLUX.2-Klein-9B: $\alpha = 0.8$, $\beta = 0.5$.
- SDXL: $\alpha = 0.2$, $\beta = 0.7$.

Fidelity-preservation ablation. For Figure 5, we hold β at each model's default value and sweep α:

- FLUX.1-dev: $0.2, 0.4, 0.6, 0.8$ without preservation; $0.4, 0.8, 1.2, 1.6$ with blockwise or layerwise preservation.
- FLUX.2-Klein-9B: $0.2, 0.4, 0.6, 0.8$ without preservation; $0.4, 0.8, 1.2, 1.6$ with blockwise or layerwise preservation.
- SDXL: $0.05, 0.1, 0.15, 0.2$ without preservation; $0.1, 0.2, 0.3, 0.4$ with blockwise or layerwise preservation.

Method comparison. For Figure 2, we sweep the following parameters:

- FLUX.1-dev: CADS noise scale $0.002, 0.004, 0.006, 0.008$ ($\psi = 1$); SPARK scale $0.05, 0.1, 0.15, 0.2$; PG coefficient $0.01, 0.02, 0.03, 0.04$; NegToMe $\alpha = 1, 2, 3, 4$; and TokenRepel $\alpha = 0.4, 0.8, 1.2, 1.6$.
- FLUX.2-Klein-9B: CADS noise scale $0.005, 0.01, 0.015, 0.02$ ($\psi = 1$); SPARK scale $0.025, 0.05, 0.075, 0.1$; PG coefficient $0.1, 0.2, 0.3, 0.4$; NegToMe $\alpha = 0.1, 0.4, 0.6, 0.8$; and TokenRepel $\alpha = 0.4, 0.8, 1.2, 1.6$.
- SDXL: CADS noise scale $0.01, 0.02, 0.03, 0.04$ ($\psi = 1$); SPARK scale $0.05, 0.1, 0.15, 0.2$; PG coefficient $0.005, 0.006, 0.007, 0.008$; NegToMe $\alpha = 0.05, 0.1, 0.3, 0.4$; and TokenRepel $\alpha = 0.1, 0.2, 0.3, 0.4$.

Image editing. For Table 3, we use FLUX.1-Kontext-dev with 28 inference steps and generate four images for each image–instruction pair. Fidelity metrics are computed against the LLM-generated target prompts, and diversity metrics are computed among the four edits of each pair. Token density preservation is applied in all layers.

D Additional Results

D.1 Number of Corrected Layers

As shown in Figure 11, correcting more layers increases HPSv3 while decreasing DreamSim.

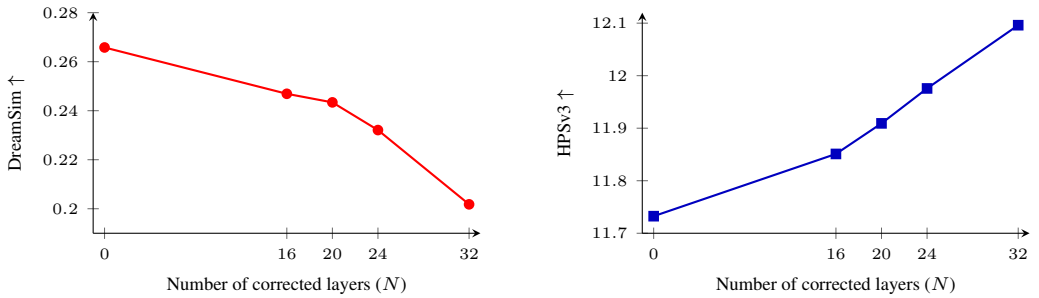


Figure 11: Effect of correcting the first N transformer layers on DreamSim (left) and HPSv3 (right) for FLUX.2-Klein-9B on DrawBench with $\alpha \approx 0.84$ and $\beta = 0.5$. TokenRepel is applied at the first sampling step; $N = 0$ uses no correction.

D.2 Alignment of Repulsion and Density Directions

To examine when token density preservation acts, we measure the cosine similarity between the repulsion displacement $\mathbf{u}_m^r - \mathbf{u}_m^+$ and the high-density direction $\mathbf{u}_m^+ - \mathbf{u}_m^-$ before correction. We average over tokens with nonzero repulsion displacement across the layers at the first sampling step of FLUX.2-Klein-9B, including text tokens in single-stream layers. As α increases, the mean cosine becomes more negative and the fraction of negatively aligned tokens rises (Figure 12). These tokens oppose the estimated high-density direction and would be modified by the half-space correction in Equation (8).

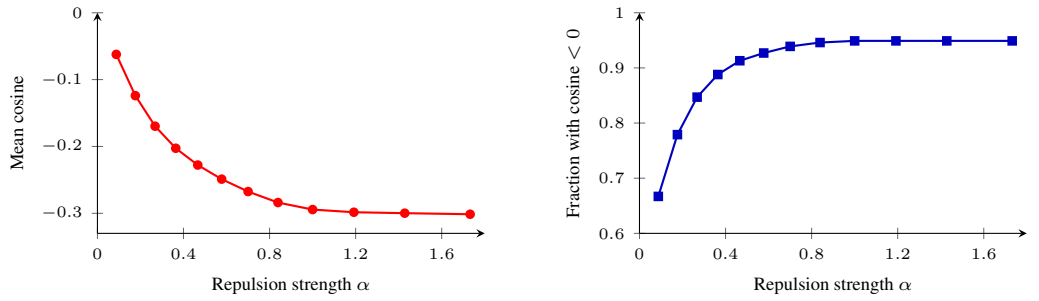
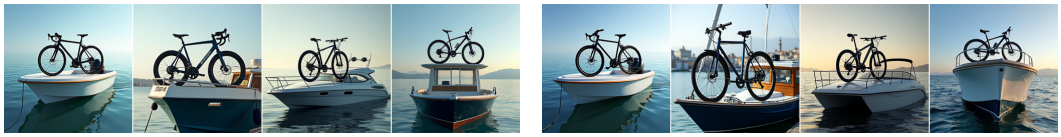


Figure 12: Alignment between the repulsion displacement and the high-density direction before correction for FLUX.2-Klein-9B on DrawBench with $\beta = 0.5$. The plots show the mean token-wise cosine (left) and the fraction of negatively aligned tokens (right) as α varies.



(a) A bicycle on top of a boat.



(b) A small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws. It is widely kept as a pet or for catching mice, and many breeds have been developed.



(c) An elephant is behind a tree. You can see the trunk on one side and the back legs on the other.

Figure 13: FLUX.1-dev qualitative results under the same initial noise. Left: baseline; right: ours.



(a) Refrigerator.



(b) A sheep to the right of a wine glass.



(c) A photo of a black teddy bear.

Figure 14: FLUX.2-Klein qualitative results under the same initial noise. Left: baseline; right: ours.

E Derivation of Token Density Preservation

We seek the smallest change to the repulsion-induced velocity difference Δ^r_{ij} that places it in the half-space $\mathcal{R}_{ij}$:

$$\min_{\bar{\Delta}^r_{ij}} \quad \frac{1}{2}\|\bar{\Delta}^r_{ij} - \Delta^r_{ij}\|^2 \quad \text{s.t.} \quad \langle \bar{\Delta}^r_{ij}, \Delta^h_{ij} \rangle \geq 0. \tag{12}$$

If $\langle \Delta^r_{ij}, \Delta^h_{ij} \rangle \geq 0$, the original update is feasible and already minimizes the objective. Otherwise, the optimum lies on the boundary $\langle \bar{\Delta}^r_{ij}, \Delta^h_{ij} \rangle = 0$. The stationarity condition gives $\bar{\Delta}^r_{ij} = \Delta^r_{ij} + \lambda \Delta^h_{ij}$, where $\lambda = -\langle \Delta^r_{ij}, \Delta^h_{ij} \rangle / \|\Delta^h_{ij}\|^2$. Substituting λ yields the projection in Equation (8).

F Additional Qualitative Results

We present additional qualitative results for FLUX.1-dev and FLUX.2-Klein-9B.

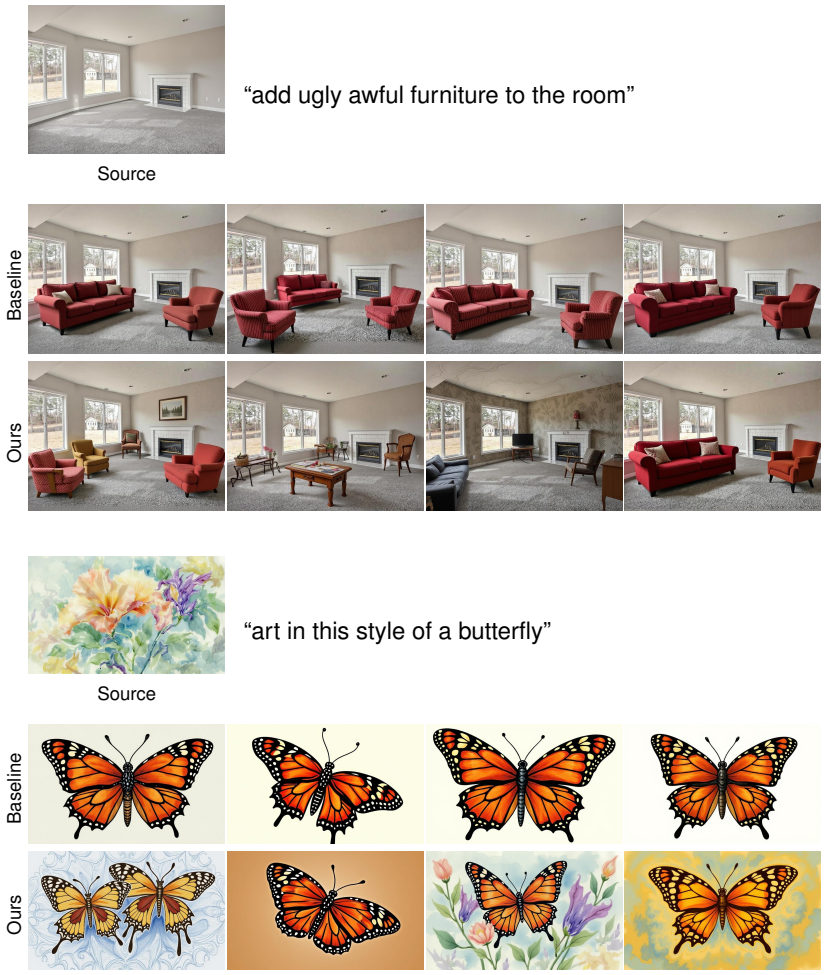


Figure 15: Qualitative image-editing results on KontextBench using FLUX Kontext under the same initial noise. Each example shows the source image and the instruction (top), followed by four edits from the baseline and TokenRepel.

Image editing. Figures 15 and 16 show additional image-editing results on KontextBench using FLUX Kontext.

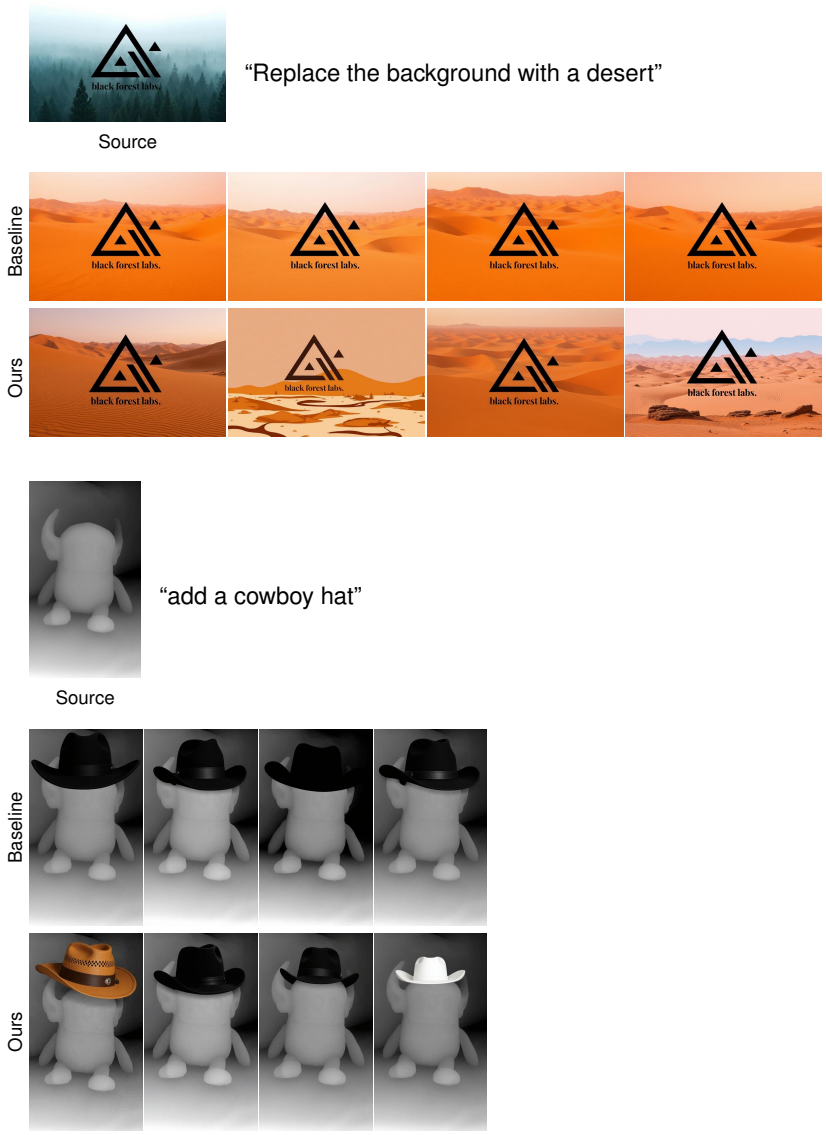


Figure 16: Qualitative image-editing results on KontextBench using FLUX Kontext under the same initial noise. Each example shows the source image and the instruction (top), followed by four edits from the baseline and TokenRepel.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Experiment section and method section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: section 4.3

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: section 3

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: appendix

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: code will be released after acceptence.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: First part of the experiment

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the experiment setting part.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: opensource datasets, clear reference

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: It will be benifitial in the Image generation communities. No negative impacts I can think of.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the models and datasets used in the paper are open-sourced.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: answerYes

    Justification: We commit to releasing the fully documented codebase and pre-trained configurations under an open-source license upon publication.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: For code organizing and writting polish.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.