
Token Perturbation Guidance for Diffusion Models

Anonymous Author(s)

Affiliation

Address

email

References

- 1 [1] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- 2
- 3 [2] Donghoon Ahn, Hyounghwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee
- 4 Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention
- 5 guidance. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.
- 6 [3] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of
- 7 attention. *arXiv preprint arXiv:2408.00760*, 2024.
- 8 [4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and
- 9 Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv*
- 10 *preprint arXiv:2307.01952*, 2023.

A Orthogonal token perturbation matrix designs

TPG directly applies structured perturbations to the intermediate token embeddings within the denoiser during inference. Specifically, consider the intermediate hidden-state activations of the denoiser at a given layer, represented as a tensor $\mathbf{H} \in \mathbb{R}^{B \times N \times C}$, where B denotes the batch size, N the number of tokens, and C the dimension of each token’s feature vector. At each denoiser block k and diffusion timestep i , we apply the perturbation by multiplying the token-embedding with an orthogonal (or approximately orthogonal) matrix $\mathbf{P}_{k,i} \in \mathbb{R}^{N \times N}$ along the token dimension:

$$\mathbf{H}' = \mathbf{P}\mathbf{H}.$$

The primary goal of these perturbations is to preserve global information flow while disrupting local correlations that may lead to overfitting or artifacts. To achieve this, we investigated four distinct perturbation methods for $\mathbf{P}_{k,i}$:

- **Token Shuffling.** Represented by a permutation matrix $\mathbf{S}_{k,i} \in \mathbb{R}^{N \times N}$, where k denotes the denoiser’s block index and i the timestep. The permutation matrix rearranges the tokens by selecting exactly one token from each position and assigning it to a new position; mathematically, this means that each row and column contains exactly one entry of "1", while all other entries are zeros. It simply changes the order of tokens without altering their magnitude or norm, satisfying:

$$\mathbf{S}_{k,i}^\top \mathbf{S}_{k,i} = \mathbf{I}.$$

- **Random Sign Flipping.** This perturbation method is defined using a diagonal matrix $\mathbf{D}_{k,i} \in \mathbb{R}^{N \times N}$, whose diagonal entries d_j are drawn independently and identically from $\{+1, -1\}$. Each token’s embedding is thus flipped in sign independently. By construction, the matrix $\mathbf{D}_{k,i}$ ensures that the ℓ_2 -norm of every token embedding is preserved, by satisfying the orthogonality condition:

$$\mathbf{D}_{k,i}^\top \mathbf{D}_{k,i} = \mathbf{I}.$$

- **Walsh–Hadamard Transform (WHT).** The WHT uses a normalized Hadamard matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$, whose entries are $\pm 1/\sqrt{N}$ and which satisfies $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_N$. When N is a power of two (i.e. $N = 2^m$), we compute the transform of an $N \times C$ token matrix \mathbf{X} in $m = \log_2(N)$ iterative stages. We begin with $\mathbf{W}^{(0)} = \mathbf{X}$, and for each stage $s = 1, \dots, m$, update

$$\begin{aligned} \mathbf{W}_{j,:}^{(s)} &= \mathbf{W}_{j,:}^{(s-1)} + \mathbf{W}_{j+2^{s-1},:}^{(s-1)}, \\ \mathbf{W}_{j+2^{s-1},:}^{(s)} &= \mathbf{W}_{j,:}^{(s-1)} - \mathbf{W}_{j+2^{s-1},:}^{(s-1)}, \end{aligned} \quad j = 1, 3, \dots, N - 2^{s-1} + 1.$$

After m stages, the result $\mathbf{W}^{(m)}$ equals $\mathbf{W}\mathbf{X}$. This structured, deterministic mixing redistributes each token’s information uniformly across all others while preserving every token’s ℓ_2 -norm.

- **Haar-Random Orthogonal Perturbation.** At each denoiser block k and diffusion timestep i , we generate a dense orthogonal matrix $\mathbf{Q}_{k,i} \in \mathbb{R}^{N \times N}$ by first sampling $\mathbf{A} \sim \mathcal{N}(0, 1)^{N \times N}$ and then computing its QR decomposition $\mathbf{A} = \mathbf{Q}\mathbf{R}$. We set $\mathbf{Q}_{k,i} = \mathbf{Q}$. Since the entries of \mathbf{A} are i.i.d. Gaussian, the orthogonal factor \mathbf{Q} is distributed uniformly (with respect to the Haar measure) over the orthogonal group $\mathcal{O}(N)$, and by construction

$$\mathbf{Q}_{k,i}^\top \mathbf{Q}_{k,i} = \mathbf{I}.$$

This yields an isotropic rotation in the N -dimensional token-index space, mixing all token positions globally without changing their ℓ_2 -norm.

Each method preserves the overall norm and energy of the embeddings but changes their local structure in uniquely effective ways. Empirically, these operations break up small-scale noise patterns that the denoiser might overfit, while still carrying global structure for high-quality sample generation. Among these methods, token shuffling typically provides the best overall performance: it is straightforward to implement, has minimal computational overhead, and consistently achieves significant improvements in both diversity and fidelity of generated samples.

53 B Societal impact

54 Generative modeling, particularly in the domains of images and videos, holds immense potential
 55 for misuse, raising important ethical concerns. While advancements in sample quality, such as
 56 those achieved through our method, can make generated content more realistic and convincing, this
 57 heightened believability can unfortunately facilitate the spread of disinformation. Such misuse may
 58 have far-reaching negative effects on society, including the amplification of existing stereotypes
 59 and the inadvertent reinforcement of harmful biases. Although our improvements do not introduce
 60 entirely new uses for the technology, they may nonetheless increase the risk of these unintended
 61 consequences. It is therefore crucial to remain vigilant and consider the broader societal impacts that
 62 enhancements in generative modeling capabilities might entail.

63 C Additional qualitative results

64 In this section, we provide additional qualitative results to showcase the effectiveness and adaptability
 65 of our Token Perturbation Guidance (TPG) method across different generation tasks and to compare
 66 its performance with other existing methods.



Figure 1: Visualization of the denoising process over time for different guidance strategies: CFG [1], PAG [2], SEG [3], and our proposed TPG. Each row shows generated images at various denoising timesteps, from $t = 981$ (left) to $t = 1$ (right). The text prompt used is "A red stop sign underneath green street signs".

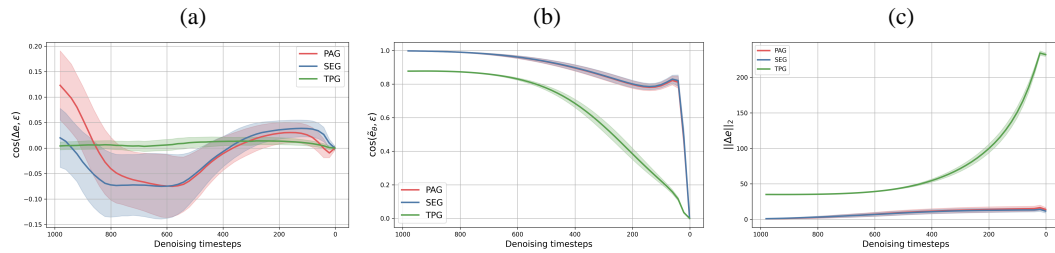


Figure 2: **Analysis of guidance behavior across denoising steps in unconditional setting.** (a) Cosine similarity between the added guidance term $\Delta\epsilon$ and the true noise ϵ . SEG [3] and PAG [2] exhibit negative alignment at intermediate steps, while TPG maintains near-zero cosine, indicating orthogonality to the noise. (b) Cosine similarity between the full guided score $\tilde{\epsilon}_\theta$ and ϵ . SEG [3] and PAG [2] maintain strong alignment throughout, while TPG decays more rapidly. (c) ℓ_2 norm of the guidance term $\Delta\epsilon$.

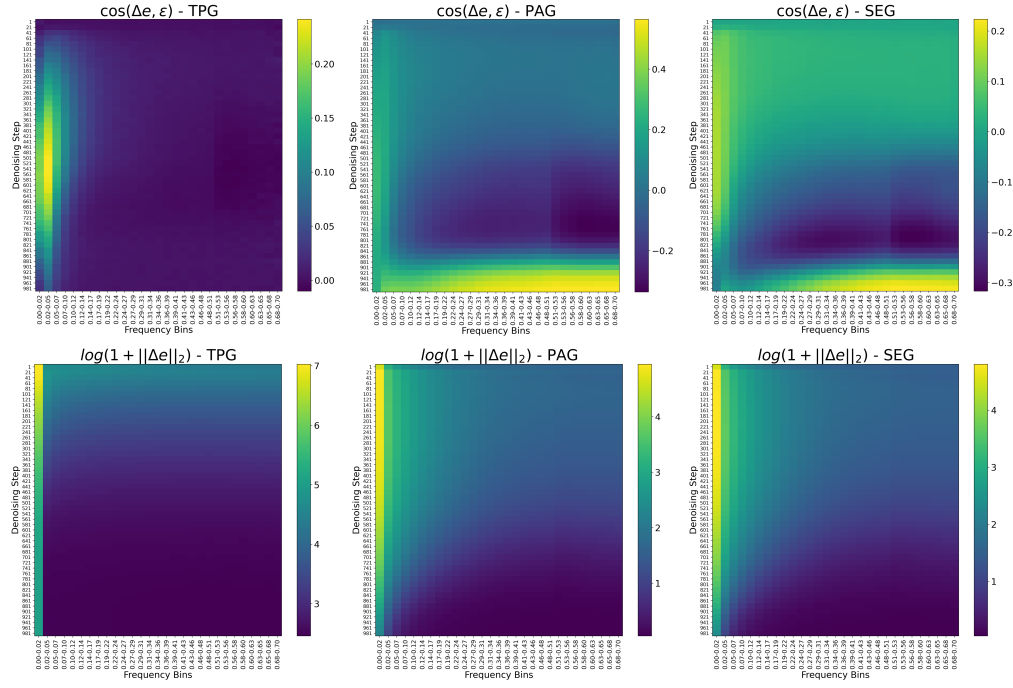


Figure 3: **Frequency-step analysis of guidance residuals in unconditional setting.** Each heat-map plots either the cosine similarity between the residual Δe and the ground-truth noise ϵ (top row) or the ℓ_2 -norm of the guidance term (bottom row) as a function of frequency bin (horizontal axis) and denoising step (vertical axis; 1000 \rightarrow 1). **Top:** For TPG, the guidance term stays almost orthogonal to the noise across all frequencies, with a mild positive bump in the very lowest bands. SEG [3], in contrast, transitions from weakly positive alignment in early steps to a pronounced negative band centred at medium frequencies. **Bottom:** TPG concentrates most of its energy in the first frequency bin and injects markedly larger magnitudes than SEG [3], whose energy remains two orders of magnitude smaller throughout denoising.

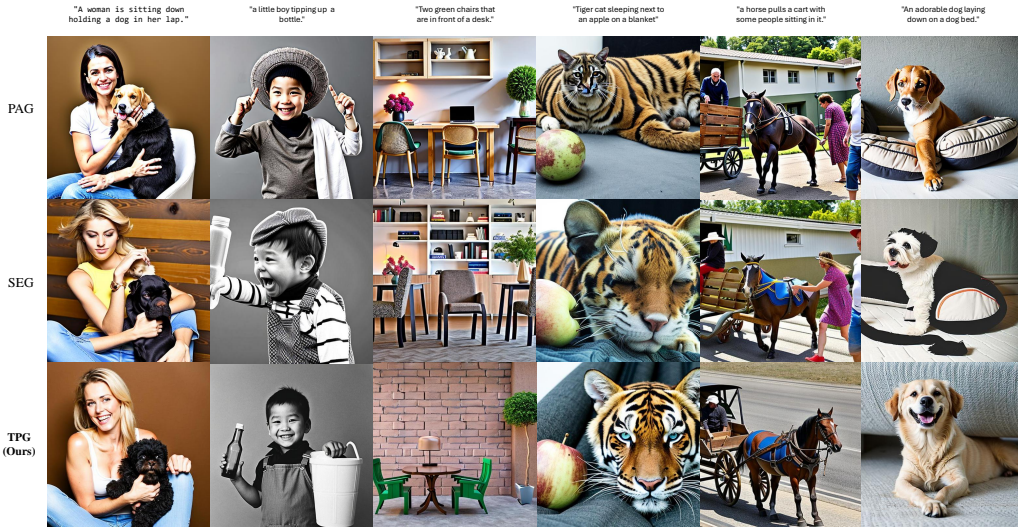


Figure 4: Qualitative comparison of conditional generations produced by PAG [2], SEG [3], and our TPG.



Figure 5: Qualitative comparison of unconditional generations produced by PAG [2], SEG [3], and our TPG.



Figure 6: Qualitative comparison of face images generated by SEG [3] and by our TPG under both conditional and unconditional settings. SEG [3] clearly produces unrealistic patterns in the generated faces.

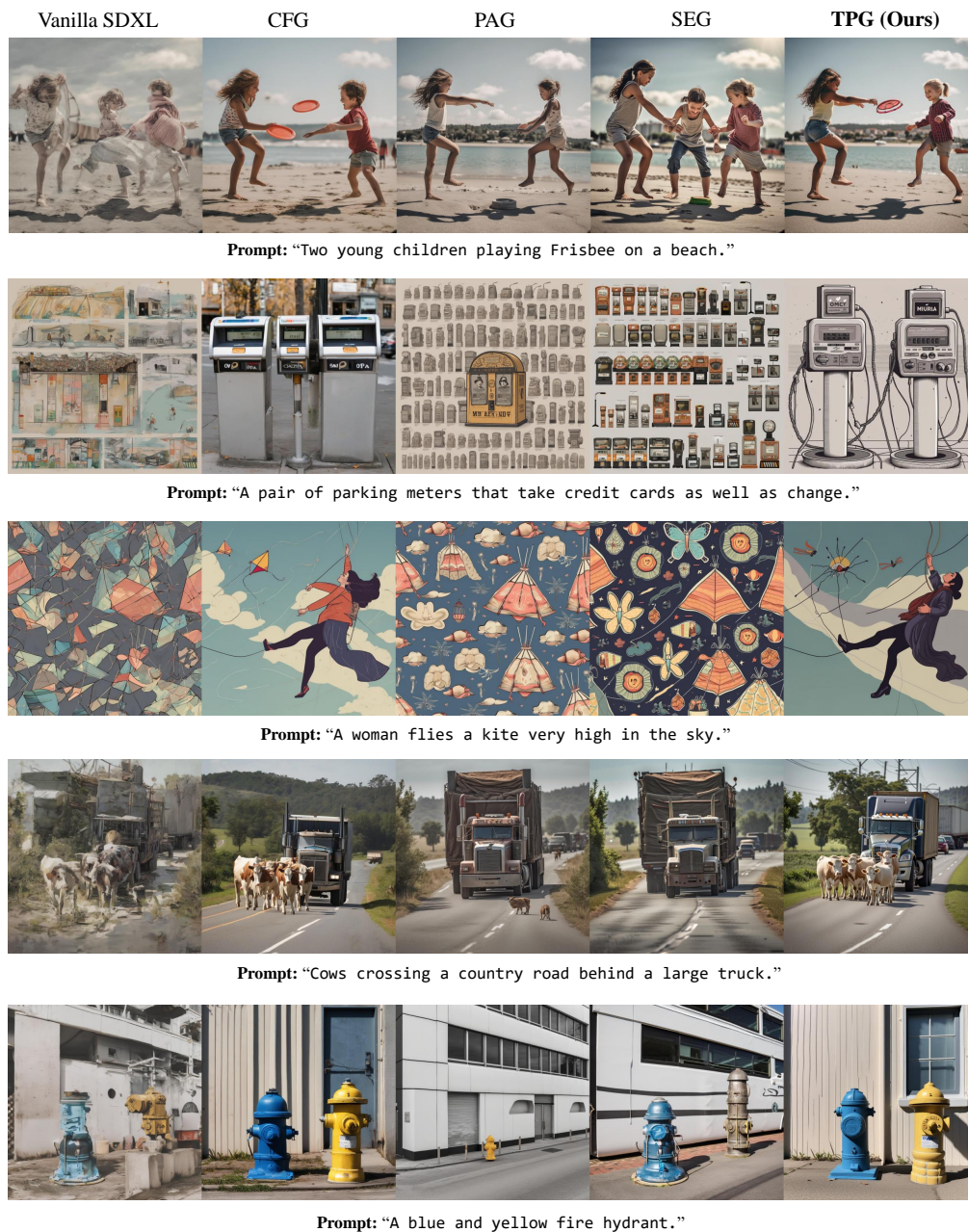


Figure 7: Qualitative comparison of conditional generations produced by Vanilla SDXL [4], CFG [1], PAG [2], SEG [3], and our TPG.

