Toward the Frontiers of Reliable Diffusion Sampling via Adversarial Sinkhorn Attention Guidance

Kwanyoung Kim

Samsung Research
k_0.kim@samsung.com

Abstract

Diffusion models have demonstrated strong generative performance when using guidance methods such as classifier-free guidance (CFG), which enhance output quality by modifying the sampling trajectory. These methods typically improve a target output by intentionally degrading another, often the unconditional output, using heuristic perturbation functions such as identity mixing or blurred conditions. However, these approaches lack a principled foundation and rely on manually designed distortions. In this work, we propose **A**dversarial **S**inkhorn **A**ttention **G**uidance (**ASAG**), a novel method that reinterprets attention scores in diffusion models through the lens of optimal transport and intentionally disrupt the transport cost via Sinkhorn algorithm. Instead of naively corrupting the attention mechanism, ASAG injects an adversarial cost within self-attention layers to reduce pixel-wise similarity between queries and keys. This deliberate degradation weakens misleading attention alignments and leads to improved conditional and unconditional sample quality. ASAG shows consistent improvements in text-to-image diffusion, and enhances controllability and fidelity in downstream applications such as IP-Adapter and ControlNet. The method is lightweight, plug-and-play, and improves reliability without requiring any model retraining.

Introduction

Diffusion models have led to substantial progress in the field of image and video generation (Rombach et al. 2022; Esser et al. 2024a; Ruiz et al. 2023; Chen et al. 2024b; Xie et al. 2024; Blattmann et al. 2023; Chen et al. 2024a). Despite their effectiveness, direct or naïve sampling often results in subpar output quality. A widely adopted solution is Classifier-Free Guidance (CFG) (Ho and Salimans 2022), which enhances class-conditional generation by computing the difference between the score functions of conditional and unconditional models and applying a weighted adjustment. Although CFG improves sample fidelity, it introduces additional training and can lead to degraded outputs when the guidance scale is too large.

Inspired by CFG, a number of guidance-based sampling techniques have emerged (Hong et al. 2023; Karras et al. 2024; Ahn et al. 2025; Hong 2024; Chung et al. 2024; Sadat et al. 2024; Li et al. 2024). A common strategy involves generating "weakened outputs" to serve as auxiliary signals that guide the primary model toward better sampling. While such approaches have shown empirical success, they come with inherent limitations. For example, AutoGuidance (AG) (Karras et al. 2024) relies on a poorly trained model, which is often unstable and difficult to optimize. To avoid retraining, attention based alternatives have been proposed. Perturbed Attention Guidance (Ahn et al. 2025) distorts attention maps using identity masking, and Smooth Energy Guidance (SEG) (Hong 2024) applies Gaussian blur to attention weights.

Although these methods aim to weaken model outputs for improved guidance, they rely on naive heuristic functions and lack clear theoretical justification as to why such perturbations consistently enhance sample quality. This theoretical gap naturally raises a fundamental question: *what constitutes an optimal perturbation for weakening outputs in a theoretically grounded manner?*

To address this question, we first revisit the intrinsic connection between attention mechanisms and optimal transport (OT) theory. One of the key contributions of this paper is the discovery and reinterpretation of classical results from optimal transport, highlighting how attention mechanisms can be framed within the OT framework. Prior studies, such as those by (Sander et al. 2022; Kim, Oh, and Ye 2024), have shown that attention computations can be improved by employing OT-inspired methods, specifically the Sinkhorn-Knopp algorithm (Cuturi 2013), to produce doubly stochastic attention maps. While these previous studies have largely utilized OT theory to boost attention performance in vision and language tasks (Chen et al. 2020; Liu et al. 2020; Chen et al. 2022), our approach deviates significantly from this established perspective.

Specifically, we propose a novel guidance framework, termed Adversarial Sinkhorn Attention Guidance (ASAG), which explicitly reinterprets self-attention scores—representing pixel-level similarities and interactions—through the lens of OT theory. In contrast to classical OT-based approaches that minimize transport cost to encourage semantic alignment between image embeddings, ASAG adopts an adversarial perspective by intentionally minimizing their interactions. This leads to an entropy-maximizing attention map, which corresponds to an optimally perturbed self-attention distribution.

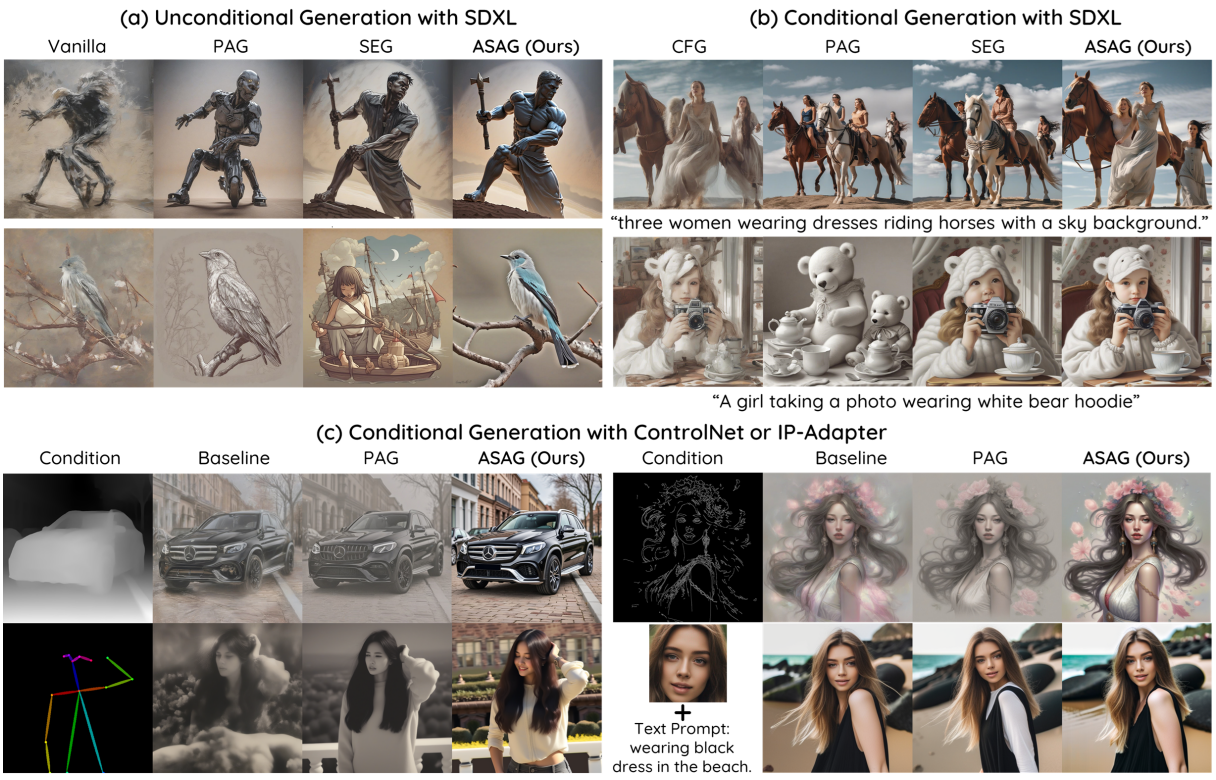


Figure 1: Qualitative comparison. (a) unconditional generation, (b) conditional generation with other guidance sampling methods, and (c) conditional generation using ControlNet and IP-Adapter. Our method, **ASAG**, significantly improves visual quality in both unconditional and conditional settings. It also remarkably enhances external frameworks like ControlNet and IP-Adapter. Crucially, ASAG requires no additional training, making it broadly compatible and readily deployable.

In doing so, we establish a theoretically grounded method for systematically disrupting similarity-based alignments in diffusion models. To the best of our knowledge, this is the first work to leverage OT theory to construct an adversarial attention perturbation, leading to significant improvements in the fidelity and controllability of diffusion-based image generation.

Building upon these theoretical foundations and insights, we demonstrate that our proposed method, ASAG, effectively improves generation quality in both unconditional and conditional diffusion sampling settings. Furthermore, extensive experiments show that when combined with existing frameworks such as ControlNet and IP-Adapter, ASAG consistently outperforms other guidance approaches by a significant margin. Our key contributions can be summarized:

- We propose ASAG, a novel and theoretically grounded diffusion guidance method that adversarially disrupts attention mechanisms by intentionally minimizing interaction between image embeddings.
- We provide an in-depth theoretical analysis, leveraging OT theory to establish a clear and rigorous foundation for our guidance method. To the best of our knowledge, our study is the first to reinterpret OT theory from an adversarial perspective for perturbing attention scores to improve diffusion generative performance.
- We empirically demonstrate that ASAG significantly enhances performance across general generative tasks, including unconditional and conditional image generation. Furthermore, we show the broad applicability and generalizability of ASAG by integrating it with widely-used generative frameworks such as ControlNet and IP-Adapter, achieving substantial improvements over existing guidance approaches.

Preliminary

Diffusion Models

Diffusion models (DM) (Ho, Jain, and Abbeel 2020; Song et al. 2021) are a class of generative models that produce samples by reversing a known forward diffusion process through estimation of the score function of a given data distribution. Specifically, given samples $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x})$, DMs define a forward noise-adding Markov process $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad t = 1, \ldots, T$, where the variance schedule $\{\beta_t\}_{t=1,\ldots,T}$ is predefined.

Consequently, the distribution at any intermediate timestep t can be explicitly expressed as $q(\mathbf{x}_t) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, with the cumulative coefficient defined as $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, where $\alpha_t = 1 - \beta_t$. As t approaches T, the distribution $q(\mathbf{x}_T)$ converges to an isotropic Gaussian distribution, $q(\mathbf{x}_T) \approx \mathcal{N}(0, \mathbf{I})$. To generate samples, diffu-

sion models parameterize the reverse diffusion process as, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$, where the model parameters θ can be optimized by denoising score matching (DSM) objective (Vincent 2011):

$$\min_\theta \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) - \boldsymbol{\epsilon}\|_2^2 \right], \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Once trained, sampling from a diffusion model proceeds by sequentially reversing the diffusion steps starting from an isotropic Gaussian noise sample. For instance, the Denoising Diffusion Implicit Model (DDIM) (Song, Meng, and Ermon 2021) generates samples by iteratively applying the update rule:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0(t) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t), \tag{2}$$

where the denoised estimate $\hat{\mathbf{x}}_0(t) = \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]$ is computed using Tweedie's formula (Efron 2011; Kim and Ye 2021):

$$\hat{\mathbf{x}}_0(t) = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{3}$$

The sampling step described above is repeated recursively from timestep T down to timestep 1.

Guidance Sampling in Diffusion Models

To enhance generation with arbitrary conditions (typically class or textual embeddings), various guidance-based sampling methods have been proposed (Dhariwal and Nichol 2021; Ho and Salimans 2022; Chung et al. 2024; Hong et al. 2023; Karras et al. 2024; Ahn et al. 2025; Hong 2024; Kim and Sim 2025). Let the conditional model be defined as $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$, which we simplify as $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c})$, and the unconditional counterpart as $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \varnothing)$.

Several studies (Ho and Salimans 2022; Hong et al. 2023; Ahn et al. 2025) have proposed generalized guidance frameworks using imaginary labels. We revisit this idea with an implicit discriminator $\mathcal{D}(\mathbf{x}_t) = \frac{p(\mathbf{y}|\mathbf{x}_t)}{p(\tilde{\mathbf{y}}|\mathbf{x}_t)}$, which distinguishes desirable from undesirable samples during the diffusion process. Here, $\mathbf{y}$ and $\tilde{\mathbf{y}}$ denote imaginary desirable and undesirable labels, respectively.

Similar to CFG, which uses an implicit classifier, implicit discriminator $\mathcal{D}$ encourages sampling along desirable trajectories while suppressing undesirable ones. By applying Bayes' rule and a mathematical transformation, we derive the following guided sampling objective:

$$\begin{aligned}\boldsymbol{\epsilon}'_\theta(\mathbf{x}_t, \mathbf{c}) &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - s\sigma_t\nabla_{\mathbf{x}_t}\left(\log p(\mathbf{x}_t|\mathbf{y}) - \log p(\mathbf{x}_t|\tilde{\mathbf{y}})\right) \\ &= \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) + s\left(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - \tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c})\right),\end{aligned} \tag{4}$$

where s is the guidance strength. Since the diffusion model learns to approximate the score function of the desirable distribution, the term $\sigma_t\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{y})$ can be replaced by $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c})$. The term $\tilde{\boldsymbol{\epsilon}}_\theta$ is a heuristically constructed weaker variant that simulates an undesirable score. For example, in CFG, the class condition is dropped, while in PAG, an identity condition is injected to produce undesirable outputs.

Although these approaches have shown empirical success and clearly demonstrate the necessity of modeling undesirable paths, the theoretical justification for how to construct such paths remains limited.

Optimal Transport and Sinkhorn-Knopp

In classical discrete optimal transport (OT), given a predefined cost matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ and two probability vectors $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Sigma_n = \{\boldsymbol{p} \in \mathbb{R}^n : \boldsymbol{p} \geq 0, \mathbf{1}^\top\boldsymbol{p} = 1\}$, the OT problem is formulated as:

$$d_{\mathbf{M}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P}\in\mathcal{U}(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{P}, \mathbf{M}\rangle, \tag{5}$$

where $\mathcal{U}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\mathbf{P} \in \mathbb{R}_+^{n\times n} | \mathbf{P}\mathbf{1}_n = \boldsymbol{\mu}, \mathbf{P}^\top\mathbf{1}_n = \boldsymbol{\nu}\}$ denotes the transport polytope consisting of non-negative matrices with prescribed row and column sums, and $\langle\cdot,\cdot\rangle$ indicates the Frobenius inner product. Directly solving the OT problem incurs a computational cost of $O(n^3 \log n)$, which is often prohibitively expensive.

To overcome this, the entropy-regularized OT formulation, solved via the Sinkhorn-Knopp (Sinkhorn) algorithm (Cuturi 2013), is defined as:

$$d^\lambda_{\mathbf{M}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P}\in\mathcal{U}(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{P}, \mathbf{M}\rangle - \frac{1}{\lambda}\langle \mathbf{P}, \log \mathbf{P}\rangle, \tag{6}$$

where $\lambda > 0$ is the regularization parameter, and the second term is the entropy regularizer. This problem is efficiently solved by iteratively updating scaling vectors. Specifically, at iteration $i \to \infty$, the optimal transport plan via Sinkhorn, $\texttt{Sinkhorn}(\lambda\mathbf{M}) := \mathbf{P}^*$, can be expressed as:

$$\mathbf{P}^* = \text{diag}(\boldsymbol{u}^i)\exp(-\lambda\mathbf{M})\text{diag}(\boldsymbol{v}^i), \tag{7}$$

where scaling vectors $\boldsymbol{u}^i$ and $\boldsymbol{v}^i$ are updated via:

$$\boldsymbol{u}^i = \frac{\boldsymbol{\mu}}{\exp(-\lambda\mathbf{M})\boldsymbol{v}^{i-1}}, \quad \boldsymbol{v}^i = \frac{\boldsymbol{\nu}}{\exp(-\lambda\mathbf{M})\boldsymbol{u}^i}, \tag{8}$$

with initialization $\boldsymbol{v}^0 = \mathbf{1}$. To improve numerical stability, we adopt the log-domain scaling version of Sinkhorn optimization (Schmitzer 2019).

Connecting Self-Attention with Sinkhorn

Given an input sequence $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ embedded in a d-dimensional space, the self-attention mechanism in Transformer models is defined as:

$$\text{SA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{9}$$

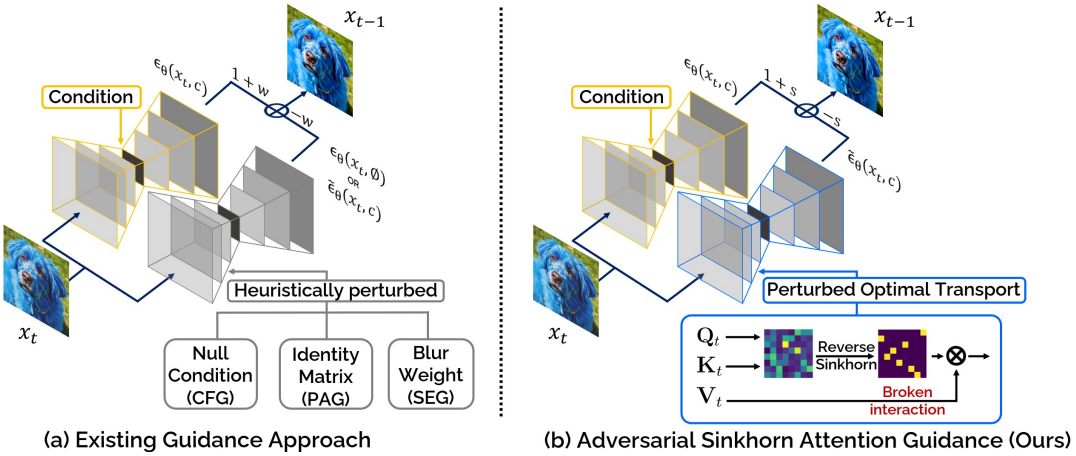
$$\text{where} \quad \mathbf{Q} = \phi_q(\mathbf{X}), \mathbf{K} = \phi_k(\mathbf{X}), \ \mathbf{V} = \phi_v(\mathbf{X}).$$

Here, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n\times d}$ denote the query, key, and value matrices obtained via learned linear projections, each corresponding $\phi_q(\cdot)$, $\phi_k(\cdot)$, and $\phi_v(\cdot)$

It has been shown that the first iteration of the Sinkhorn algorithm corresponds exactly to the $\texttt{SoftMax}$ (Sander et al. 2022), suggesting a natural generalization of attention via optimal transport. Specifically, by interpreting the attention score as a similarity matrix and defining the cost as $\mathbf{M} = (\mathbf{1} - \mathbf{Q}\mathbf{K}^\top)$, Sinkhorn-based attention (Sink-Attention) optimizes a doubly-stochastic plan that favors high similarity alignments. This leads to more structured and expressive attention maps compared to standard softmax-based attention, and empirically improves performance (Kim, Oh, and Ye 2024). The Sink-Attention can be formulated as:

$$\text{Sink-A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{Sinkhorn}\left(\lambda\mathbf{M}\right)\mathbf{V}, \tag{10}$$

where $\texttt{Sinkhorn}(\cdot)$ computes a doubly-stochastic matrix via Eq.(7), with regularization parameter λ set to $1/\sqrt{d}$.



(a) Existing Guidance Approach (b) Adversarial Sinkhorn Attention Guidance (Ours)

Figure 2: Conceptual comparison between ASAG and other guidance methods. Existing methods often use null conditions or heuristic self-attention perturbations, such as identity injection or Gaussian blur, to model undesirable paths. In contrast, ASAG defines a pixel-level attention cost and deliberately disrupts semantic interactions by minimizing this cost via Sinkhorn.

Methods

Self- and Sinkhorn Attention in Diffusion Models

Recent DMs extensively utilize attention mechanisms throughout their architecture. At each timestep t, the model processes features via self-attention, typically expressed as:

$$\mathrm{SA}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \texttt{SoftMax}\left(\frac{\mathbf{Q}_t\mathbf{K}_t^\top}{\sqrt{d}}\right)\mathbf{V}_t. \quad (11)$$

As introduced in Eq. 10, an alternative based on optimal transport is Sinkhorn attention, which replaces the `Softmax` with `Sinkhorn` operator which compute a transport plan via iterative optimization:

$$\text{Sink-A}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \texttt{Sinkhorn}(\lambda \mathbf{M}_t^{\uparrow})\mathbf{V}_t, \quad (12)$$

where $\lambda = 1/\sqrt{d}$, $\mathbf{M}_t^{\uparrow} = (\mathbf{1} - \mathbf{Q}_t\mathbf{K}_t^\top)$. Minimizing this cost means that maximize similarity between query and key matrices at t. While Sink-A can improve alignment, it is computationally expensive due to its iterative nature. When applied to all attention layers, the overhead becomes even more significant. However, our goal is not to enhance all attention mechanisms. Instead, within the guidance sampling framework, we aim to construct an effective undesirable path $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c})$ that simulates degraded attention behavior.

Adversarial Sinkhorn Attention Guidance

To construct the undesirable score function $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c})$, we propose Adversarial Sinkhorn Attention Guidance (ASAG), which selectively applies Sink-A in reverse direction to degrade attention quality in specific layers as shown Fig 2. This design is inspired by prior works such as PAG and SEG, which perturb only a subset of attention maps.

In ASAG, we replace cost $\mathbf{M}_t^{\uparrow}$ in Eq. (12) with $\mathbf{M}_t^{\downarrow} = (\mathbf{Q}_t\mathbf{K}_t^\top)$ for direction minimizing similarity between query and key matrices:

$$\mathrm{ASA}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \texttt{Sinkhorn}\left(\lambda \mathbf{M}_t^{\downarrow}\right)\mathbf{V}_t, \quad (13)$$

Algorithm 1: Adversarial Sinkhorn Attention

Input : query $\mathbf{Q}_t$, key $\mathbf{K}_t$, and value $\mathbf{V}_t$ matrices at timestep t, hyper-parameter $\lambda = 1/\sqrt{d}$, error threshold ϵ_{max}

Initialization: Attention cost $\mathbf{M}_t^{\downarrow} = (\mathbf{Q}_t\mathbf{K}_t^\top)$, $\Delta_{\boldsymbol{v}} = \infty$, $i = 1$, $\boldsymbol{v}^0 = 0$

Calculate Sinkhorn distance within inner loop ;

while $\Delta_{\boldsymbol{v}} < \epsilon_{max}$ **do**

  $\boldsymbol{u}^i = \log \boldsymbol{\mu} - \log\left[\sum \exp\left(-\lambda \mathbf{M}_t^{\downarrow} + \boldsymbol{v}^{i-1}\right)\right]$;

  $\boldsymbol{v}^i = \log \boldsymbol{\nu} - \log\left[\sum \exp\left(-\lambda (\mathbf{M}_t^{\downarrow})^\top + \boldsymbol{u}^i\right)\right]$;

  $\Delta_{\boldsymbol{v}} = \|\boldsymbol{v}^i - \boldsymbol{v}^{i-1}\|_1$;

  $i \leftarrow i + 1$;

$\mathbf{P}^* = \mathrm{diag}(\exp(\boldsymbol{u}^i)) \cdot \exp(-\lambda \mathbf{M}_t^{\downarrow}) \cdot \mathrm{diag}(\exp(\boldsymbol{v}^i))$

Output: $\mathrm{ASA}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \mathbf{P}^*\mathbf{V}_t$

ASA(Adversarial Sinkhorn Attention) minimizes interactions between noisy queries and keys, whereas Sink-A maximizes similarity via the OT objective, causing attention collapse and disrupted interactions (Algorithm 1). Because degradation does not require precise convergence, only a few Sinkhorn iterations suffice, incurring minimal overhead, and the resulting attention maps simulate an undesirable path with broken semantic coherence. Using ASA, we compute $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c})$ in Eq. (4), and provide pseudocode in Algorithm 2.

Theoretical Justification for ASAG. We formally justify ASA's perturbation as an entropy-maximizing transport plan that disrupts semantic alignment. This direction is not arbitrary but results from a constrained optimization aligned with a disruptive axis in score space.

Theorem 1 (Entropy-Maximizing Plan via Adversarial Sinkhorn)**.** *Let $\mathbf{Q}_t, \mathbf{K}_t \in \mathbb{R}^{n\times d}$ be the query and key matrices at diffusion timestep t. Define the adversarial cost matrix*

Algorithm 2: Diffusion Sampling with ASAG

Input: Diffusion model $\epsilon_\theta(\mathbf{x}_t)$ with self-attention module, $\tilde{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c})$ with ASA guidance scale s.

for t *in* $T, T-1, \cdots, 1$ **do**

$\quad \epsilon_\theta(\mathbf{x}_t, \mathbf{c}) \leftarrow \epsilon_\theta(\mathbf{x}_t, \mathbf{c}) + s\,(\epsilon_\theta(\mathbf{x}_t, \mathbf{c}) - \tilde{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}))$

$\quad \hat{\mathbf{x}}_0(t) \leftarrow (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, \mathbf{c}))/\sqrt{\bar{\alpha}_t}$

$\quad \mathbf{x}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0(t) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t, \mathbf{c})$

return: $\mathbf{x}_0$

as $\mathbf{M}_t^{\downarrow} = (\mathbf{Q}_t\mathbf{K}_t^\top)$. *The entropy-regularized OT problem is defined as* $d^{\lambda}_{\mathbf{M}_t^{\downarrow}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P}\in\mathcal{U}(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{P}, \mathbf{M}_t^{\downarrow}\rangle - \frac{1}{\lambda}\langle \mathbf{P}, \log \mathbf{P}\rangle$. *Then in the limit* $\lambda \to 0$ *(i.e.,* $1/\lambda \to \infty$*), the solution converges to the maximum-entropy plan:*

$$\lim_{\lambda\to 0} \mathbf{P}_t^* = \frac{1}{n^2}\mathbf{1}\mathbf{1}^\top.$$

Lemma 1. *Under uniform marginals* $\boldsymbol{\mu}$ *and* $\boldsymbol{\nu}$*, the coupling* $\mathbf{P}^* = \frac{1}{n^2}\,\mathbf{1}\,\mathbf{1}^\top$ *uniquely maximizes the Shannon entropy over the transport polytope* $\mathcal{U}(\boldsymbol{\mu}, \boldsymbol{\nu})$.

All proofs are deferred in supplement.

Remark 1. *By Theorem 1 and Lemma 1, the adversarial Sinkhorn plan converges to the maximum-entropy uniform coupling as* $\lambda \to 0$*, suppressing semantic preferences and yielding an increasingly unstructured attention map. This limiting case characterizes semantic alignment degradation, and ASA exploits this behavior to construct adversarial attention maps that systematically disrupt semantic correspondence.*

Corollary 1.1. *Let* $\delta_t := \epsilon_\theta(\mathbf{x}_t, \mathbf{c}) - \tilde{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c})$ *be the guidance energy between the original and ASA-induced score estimates in Eq. 4. Then* δ_t *defines a semantically grounded direction in score space, anchored to the original attention while deviating toward the entropy-maximizing adversarial trajectory. Unlike heuristic perturbations, it enables principled contrastive guidance that preserves structural semantics while deliberately weakening alignment.*

Practical Justification via Sinkhorn Approximation. While the uniform plan $\frac{1}{n^2}\mathbf{1}\mathbf{1}^\top$ represents the theoretical extreme of semantic disruption, applying it directly often leads to reduced generation diversity and unstable behavior. ASA instead adopts a Sinkhorn plan with a small but finite λ, which retains a doubly stochastic structure and enables controlled entropy increase while preserving attention stability. Unlike heuristic perturbations, this approach maintains the optimization structure of attention with a reversed objective, offering a principled and tunable guidance strategy grounded in theory and robust in practice.

Experiments

Setup. For fair comparison, we use SDXL(Podell et al. 2023) and SD3 (Esser et al. 2024b). We compare against PAG and SEG with a guidance scale of 3.0, following their official settings. SEG is not officially supported on SD3 and IP-Adapter, so it is omitted in those cases. For ASAG, we set the guidance scale $s = 1.5$ and use 25 sampling steps. Full implementation details are provided in the supplement. All experiments are run on a single NVIDIA H100 GPU.

Condition	Method	FID ↓	KID ↓	IS ↑
Unconditonal	Vanilla	122.07	0.086	7.052
	PAG	108.63	0.067	10.46
	SEG	95.43	0.062	10.35
	ASAG (Ours)	**92.01**	**0.059**	**10.54**
Condition	Method	FID ↓	CLIPScore ↑	IR ↑
Conditonal (SDXL)	CFG	28.15	25.21	0.415
	PAG	24.32	25.41	0.448
	SEG	26.80	25.39	0.431
	ASAG (Ours)	**23.30**	**25.85**	**0.459**
Condition	Method	FID ↓	CLIPScore ↑	IR ↑
Conditonal (SD3)	CFG	24.19	26.03	0.931
	PAG	23.31	26.14	0.956
	ASAG (Ours)	**22.87**	**26.33**	**0.978**

Table 1: Quantitative results of various guidance methods on the MS-COCO dataset with SDXL and SD3 in uncoditional and conditional generation.

Evaluation Metrics. We evaluate our method across multiple dimensions. For visual quality, we report the Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID), and Inception score (IR) using 30K randomly sampled prompts from the MS-COCO validation set (Lin et al. 2014). To assess text-image alignment and human preference, we use CLIPScore (Hessel et al. 2021), ImageReward (IR) (Xu et al. 2024), PickScore (Kirstain et al. 2023), and Human Preference Score (HPS v2.1) (Wu et al. 2023). Evaluations are conducted on prompts from both MS-COCO (Lin et al. 2014) and additional benchmarks including DrawBench (Saharia et al. 2022) and HPD (Wu et al. 2023). Further details are provided in the supplement.

Results on Diffusion Generation

To rigorously evaluate the effectiveness of our method, we generate 30K samples on the MSCOCO dataset using various guidance sampling techniques in both unconditional and conditional generation settings.

Unconditional Generation To isolate the effect of our method, we first evaluate ASAG in an unconditional generation setup. As shown in Tab. 1, ASAG consistently outperforms other guidance approaches across all evaluation metrics, demonstrating improved visual fidelity and greater diversity, as reflected by higher Inception Scores. Qualitatively, ASAG also produces outputs that are more visually appealing and better aligned with the original vanilla results, particularly in cases where other guidance methods generate

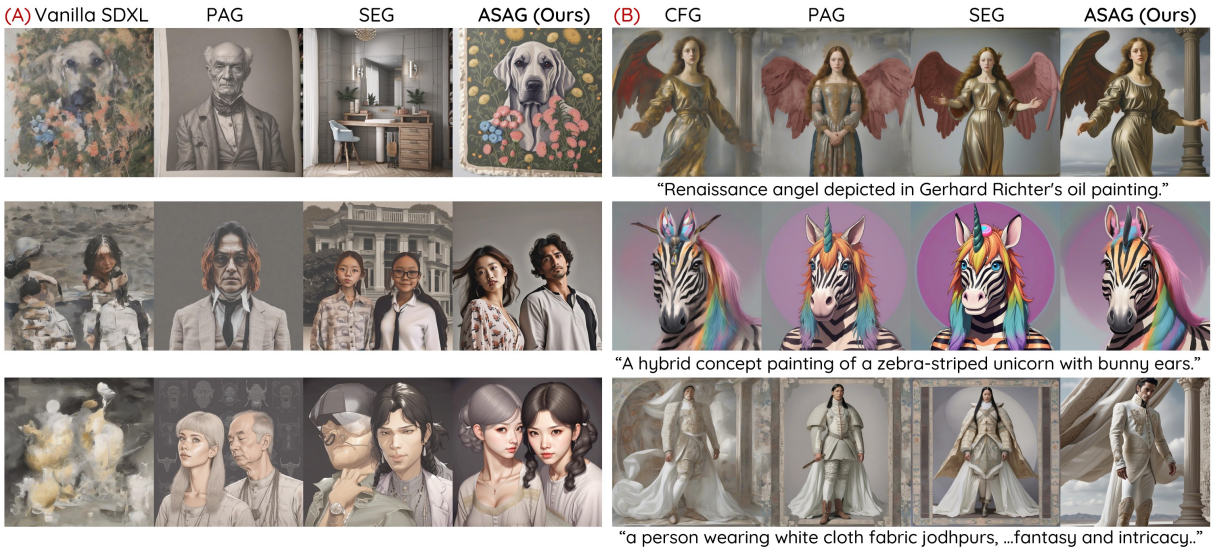


Figure 3: Comparison results on (A) unconditional and (B) conditional generation using Vanilla, CFG, PAG, SEG, and ASAG. While other guidance methods often alter the structure of the original outputs, ASAG achieves both higher visual quality and stronger consistency in structure and intent.

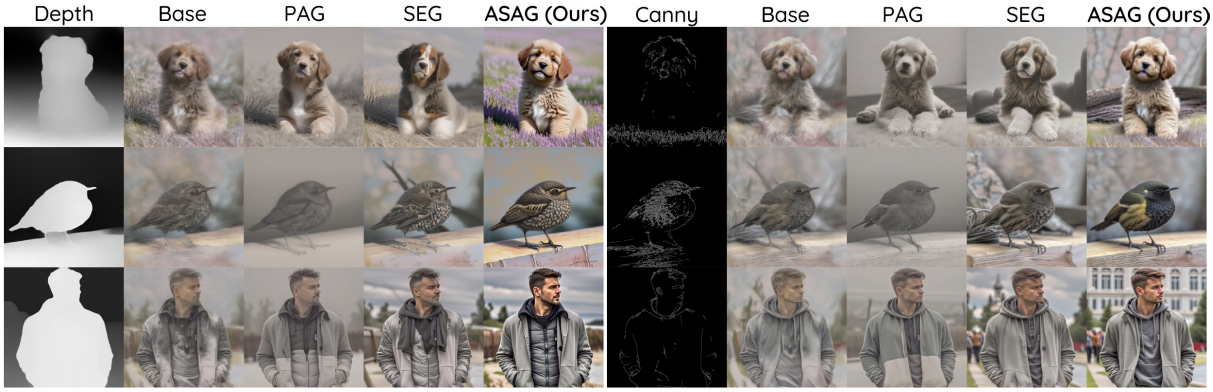


Figure 4: ControlNet examples with different guidance sampling methods. Left: Canny condition; Right: Depth condition. Our method, when integrated with ControlNet, substantially improves visual quality and preserves fine-grained image details.

Dataset	Method	CLIPScore ↑	PickScore ↑	IR ↑	HPSv2 ↑
Draw bench	CFG	25.61	21.70	0.196	26.81
	PAG	26.17	21.93	0.294	26.84
	SEG	26.03	21.78	0.290	27.06
	ASAG	**26.62**	**21.99**	**0.316**	**27.08**
HPD	CFG	27.88	21.97	0.565	26.63
	PAG	28.00	22.12	0.635	28.83
	SEG	28.15	21.96	0.622	28.73
	ASAG	**28.58**	**22.21**	**0.673**	**28.84**

Table 2: Quantitative comparison of human preference across datasets using various guidance methods with SDXL. For PAG, SEG, and ASAG, CFG guidance is used jointly.

high-quality images that nonetheless diverge significantly from the vanilla outputs (Fig. 3). These findings suggest that ASAG serves as a strong guidance strategy even in scenarios without any conditional information.

Conditional Generation We further evaluate our method in a conditional generation setup, where guidance sampling is combined with CFG. As shown in Tab. 1, while existing guidance methods benefit from CFG, ASAG achieves superior performance in both visual quality and text-image alignment. It is also compatible with the SD3 backbone and consistently outperforms other methods, demonstrating strong generalizability. To further validate its effectiveness, we evaluate on a human preference dataset (Tab. 2), where ASAG with CFG achieves state-of-the-art performance across all metrics. Qualitative results in Fig. 3 show that ASAG enhances visual quality while preserving the structural intent of the original CFG output, unlike other methods that often alter the generation semantics.

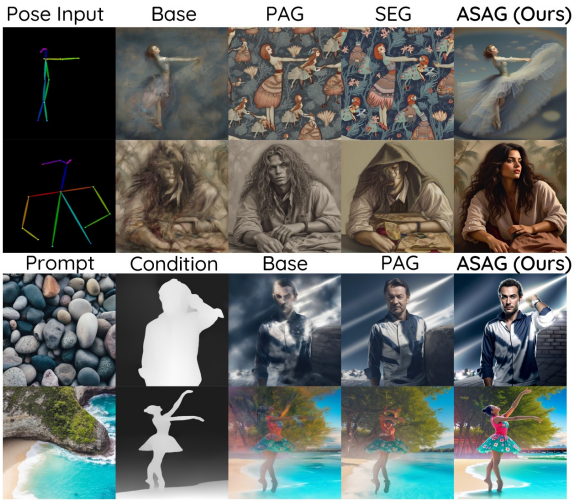


Figure 5: Comparison of guidance methods combined with ControlNet and IP-Adapter under pose and depth conditions.

Results on Downstream Tasks

To assess ASAG in downstream tasks, we evaluate ControlNet and IP-Adapter across various conditioning settings, keeping all configurations fixed except for the guidance method. As shown in Fig. 4, ASAG consistently outperforms baselines under Canny and depth conditions by better preserving structural fidelity. In challenging setups—pose with ControlNet and multimodal IP-Adapter (Fig. 5)—ASAG captures finer details and maintains stronger visual coherence where other methods degrade.

These gains are achieved without any additional training, highlighting the plug-and-play nature of ASAG. Its principled perturbation effectively steers pretrained models along favorable generation paths, yielding robust performance across diverse conditional settings. Additional qualitative results are provided in the supplement.

Ablation Study

Optimal Transport Cost. Instead of maximizing similarity with $\mathbf{M}_t = 1 - \mathbf{Q}_t\mathbf{K}_t^\top$, ASAG sets $\mathbf{M}_t = \mathbf{Q}_t\mathbf{K}_t^\top$ to explicitly minimize similarity and disrupt semantic alignment. As shown in Tab. 3, the similarity-maximizing variant also outperforms vanilla sampling, likely via CFG-like extrapolation, but incurs higher inference time due to the more demanding Sinkhorn optimization.

We also test the extreme case of the uniform plan $\frac{1}{n^2}\mathbf{1}\mathbf{1}^\top$, a theoretical upper bound on semantic disruption. Although it improves performance over the baseline, it reduces sample diversity. In contrast, our Sinkhorn-based formulation attains comparable semantic disruption with better diversity and stability, supporting its effectiveness as a practical and principled guidance strategy.

Computational Complexity. We further assess the efficiency of ASAG by measuring inference time and memory usage, as shown in Tab. 4. As described in Algorithm 1, the Sinkhorn process includes early stopping based on a convergence threshold. In practice, only 2 iterations are sufficient in most cases to produce stable transport plans. Remarkably, using just 2 iterations increases inference time by only $+0.35$ seconds while even improving generation quality, confirming that ASAG is not only efficient but also effective under minimal Sinkhorn updates.

Method	FID	KID	IS	Sinkhorn Iteration
Vanilla	122.07	0.086	7.052	-
$\mathbf{M}_t = 1 - \mathbf{Q}\mathbf{K}^\top$	111.53	0.078	9.085	≈ 10
$\mathbf{P}_t^* = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$	92.11	**0.058**	9.710	-
$\mathbf{M}_t = \mathbf{Q}\mathbf{K}^\top$	**92.01**	0.059	**10.54**	≈ 2

Table 3: Ablation study on transport cost for Sinkhorn.

Method	CFG	PAG	SEG	Ours
Inference Time (sec) ↓	1.198	1.280	1.513	1.551 (+0.35)
Memory (G) ↓	16.41	16.46	16.61	16.61 (+0.20)

Table 4: Comparison computation cost with various approaches. Inference time is measured per prompt.

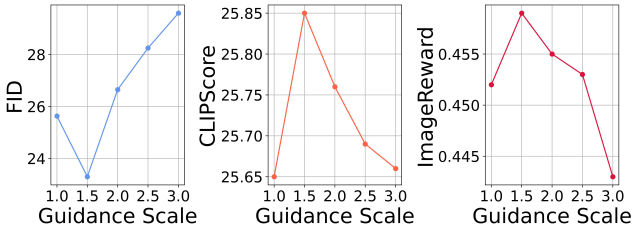


Figure 6: Analysis of guidance scale across various metrics.

Guidance Scale To determine the optimal guidance scale s in ASAG, we conduct experiments by varying s and measuring the impact on generation performance, as shown in Figure 6. We observe that performance improves with increasing s up to a certain point, but overly large values cause slightly degradation as observed in CFG cases. The peak performance is achieved at $s = 1.5$, which we adopt as the default configuration throughout our experiments.

Conclusion

In this work, we propose **Adversarial Sinkhorn Attention Guidance (ASAG)**, a guidance sampling method that replaces heuristic self-attention perturbations with a principled scheme: we define an attention cost over pixel embeddings and disrupt their interactions via the Sinkhorn algorithm, interpreted through optimal transport. ASAG is theoretically grounded, achieves state-of-the-art performance in both unconditional and conditional generation, and integrates seamlessly with frameworks such as ControlNet and IP-Adapter without additional training. Overall, ASAG provides a principled and practical guidance mechanism for diffusion models, with strong generalization across attention-based generative tasks. Our results suggest that ASAG can generalize to a wide range of generative tasks and pave the way for future research in attention-based guidance strategies.

References

Ahn, D.; Cho, H.; Min, J.; Jang, W.; Kim, J.; Kim, S.; Park, H. H.; Jin, K. H.; and Kim, S. 2025. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, 1–17. Springer.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. Prompt Learning with Optimal Transport for Vision-Language Models. *arXiv preprint arXiv:2210.01253*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024b. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, 74–91. Springer.

Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, 1542–1553. PMLR.

Chung, H.; Kim, J.; Park, G. Y.; Nam, H.; and Ye, J. C. 2024. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Efron, B. 2011. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024a. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024b. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hong, S. 2024. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *arXiv preprint arXiv:2408.00760*.

Hong, S.; Lee, G.; Jang, W.; and Kim, S. 2023. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7462–7471.

Karras, T.; Aittala, M.; Kynkäänniemi, T.; Lehtinen, J.; Aila, T.; and Laine, S. 2024. Guiding a Diffusion Model with a Bad Version of Itself. *arXiv preprint arXiv:2406.02507*.

Kim, K.; Oh, Y.; and Ye, J. C. 2024. OTSeg: Multi-Prompt Sinkhorn Attention for Zero-Shot Semantic Segmentation. In *European Conference on Computer Vision*, 200–217. Springer.

Kim, K.; and Sim, B. 2025. PLADIS: Pushing the Limits of Attention in Diffusion Models at Inference Time by Leveraging Sparsity. *arXiv preprint arXiv:2503.07677*.

Kim, K.; and Ye, J. C. 2021. Noise2Score: Tweedie's Approach to Self-supervised Image Denoising Without Clean Images. *Advances in Neural Information Processing Systems*, 34: 864–874.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, T.; Luo, W.; Chen, Z.; Ma, L.; and Qi, G.-J. 2024. Self-Guidance: Boosting Flow and Diffusion Generation on Their Own. *arXiv preprint arXiv:2412.05827*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, Y.; Zhu, L.; Yamada, M.; and Yang, Y. 2020. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4463–4472.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Sadat, S.; Kansy, M.; Hilliges, O.; and Weber, R. M. 2024. No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv preprint arXiv:2407.02687*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Sander, M. E.; Ablin, P.; Blondel, M.; and Peyré, G. 2022. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, 3515–3530. PMLR.

Schmitzer, B. 2019. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3): A1443–A1481.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Vincent, P. 2011. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7): 1661–1674.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Xie, E.; Chen, J.; Chen, J.; Cai, H.; Tang, H.; Lin, Y.; Zhang, Z.; Li, M.; Zhu, L.; Lu, Y.; et al. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.