DICE: Distilling Classifier-Free Guidance into Text Embeddings

Zhenyu Zhou[1,2], **Defang Chen**[3†], **Can Wang**[1,2], **Chun Chen**[1,2], **Siwei Lyu**[3]

[1]Zhejiang University, State Key Laboratory of Blockchain and Data Security
[2]HangZhou High-Tech Zong (Binjiang) Institute of Blockchain and Data Security
[3]University at Buffalo, State University of New York

Abstract

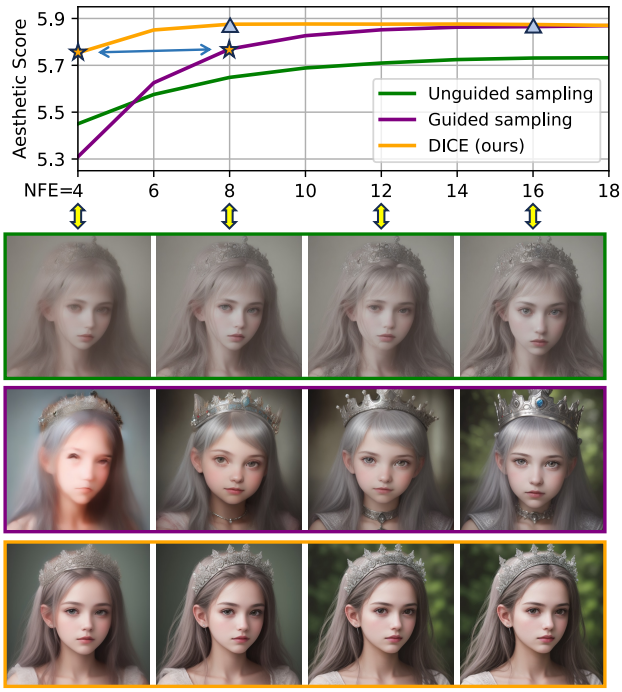
Text-to-image diffusion models are capable of generating high-quality images, but suboptimal pre-trained text representations often result in these images failing to align closely with the given text prompts. Classifier-free guidance (CFG) is a popular and effective technique for improving text-image alignment in the generative process. However, CFG introduces significant computational overhead. In this paper, we present **DI**stilling **C**FG by sharpening text **E**mbeddings (DICE) that replaces CFG in the sampling process with half the computational complexity while maintaining similar generation quality. DICE distills a CFG-based text-to-image diffusion model into a CFG-free version by refining text embeddings to replicate CFG-based directions. In this way, we avoid the computational drawbacks of CFG, enabling high-quality, well-aligned image generation at a fast sampling speed. Furthermore, examining the enhancement pattern, we identify the underlying mechanism of DICE that sharpens specific components of text embeddings to preserve semantic information while enhancing fine-grained details. Extensive experiments on multiple Stable Diffusion v1.5 variants, SDXL, and PixArt-α demonstrate the effectiveness of our method.

Code — https://github.com/zju-pi/dice
Extended version — https://arxiv.org/abs/2502.03726

Introduction

Diffusion-based generative models (Sohl-Dickstein et al. 2015; Song and Ermon 2019; Ho, Jain, and Abbeel 2020) have recently achieved remarkable advances, driven by continuously refined theoretical frameworks (Song et al. 2021; Karras et al. 2022; Chen et al. 2024b; Kingma and Gao 2024) and fast evolution of model architectures (Peebles and Xie 2023; Bao et al. 2023). Their impressive generation ability brings text-to-image generation to unprecedented levels (Rombach et al. 2022; Saharia et al. 2022; Podell et al. 2024; Esser et al. 2024), and enables a multitude of new conditional generation tasks (Croitoru et al. 2023).

In text-to-image generation (Nichol et al. 2022; Rombach et al. 2022; Saharia et al. 2022), diffusion models use text embeddings produced by pre-trained encoders such as CLIP (Radford et al. 2021) and T5 (Raffel et al. 2020). These embeddings are fixed-dimensional vectors that encapsulate the semantic content of text prompts. However, they are not specifically optimized for image generation (Li et al. 2024). Moreover, images often encompass more detailed information than text prompts can convey, making precise text-image semantic alignment challenging (Schrodi et al. 2024). Consequently, as illustrated in Figure 1, sampling with text-to-image models in their original conditional form—hereafter referred to as *unguided sampling*—often produces blurry and semantically inaccurate outputs (Meng



"Photo portrait of a girl in a silver crown."

Figure 1: Comparison of text-to-image generation: unguided sampling, guided sampling, and DICE. *Top*: Average aesthetic score (Schuhmann et al. 2022) over $5,000$ images plotted against the number of function evaluations (NFE). *Bottom*: An example of image synthesis using different methods at NFE = 4, 8, 12, and 16.

[†]Correspondence to: Defang Chen <defangch@buffalo.edu>

et al. 2023; Karras et al. 2024). To address the limited semantic signals provided by text embeddings, *guided sampling* techniques (Dhariwal and Nichol 2021; Ho and Salimans 2022) have been introduced to steer samples toward a more concentrated distribution. *Classifier-Free Guidance* (CFG) (Ho and Salimans 2022) is a widely adopted technique for guided sampling. It directs the generative process at each sampling step by extrapolating the direction between the conditional prediction and an unconditional prediction, with the guidance strength modulated by a hyperparameter known as the guidance scale. CFG enhances both image quality and text-image alignment, making it a popular choice in practice. However, an important drawback of CFG is that it requires an additional model evaluation at each step, thereby increasing the sampling overhead (Ho and Salimans 2022). Moreover, since CFG deviates from the sampling path of a normal diffusion model, it complicates the understanding of sampling dynamics (Karras et al. 2024; Zheng and Lan 2024; Bradley and Nakkiran 2024).

To mitigate the increased sampling overhead, prior research distilled CFG into a single model evaluation per sampling step (Meng et al. 2023; Hsiao et al. 2024). While these methods can effectively reduce the computational cost of CFG, they typically incur significant training overhead due to the large number of trainable parameters required and suffer from practical issues. For instance, on the Stable Diffusion v1.5 model (Rombach et al. 2022), Guided Distillation (GD) (Meng et al. 2023) fine-tunes the whole model involving 859M trainable parameters and the fine-tuned model cannot be applied to new scenarios. Plug-and-Play Distillation (PnP) (Hsiao et al. 2024) trains an auxiliary model with 361M parameters but requires multiple operations during inference, reducing the ratio of acceleration.

In this paper, we introduce **DI**stilling **C**FG by sharpening text **E**mbeddings (DICE) as an alternative approach for achieving high-quality image generation with unguided sampling. Specifically, we refine the model's input condition, *i.e.*, text embeddings, under CFG-based supervision by training a lightweight sharpener that operates only once independently of the primary text-to-image model with only 2M model parameters (Figure 2). With sharpened embeddings, our enhanced unguided sampling achieves image quality on par with guided sampling while maintaining computational efficiency. By inspecting the underlying mechanism, we reveal that DICE identifies a universal enhancement pattern: the semantically irrelevant components of the text embedding are primarily amplified, preserving essential semantic information while enriching fine-grained details in the generated images. Extensive experiments across various text-to-image models, encompassing different model capacities, image styles, and network architectures, validate the effectiveness of our method in diverse scenarios.

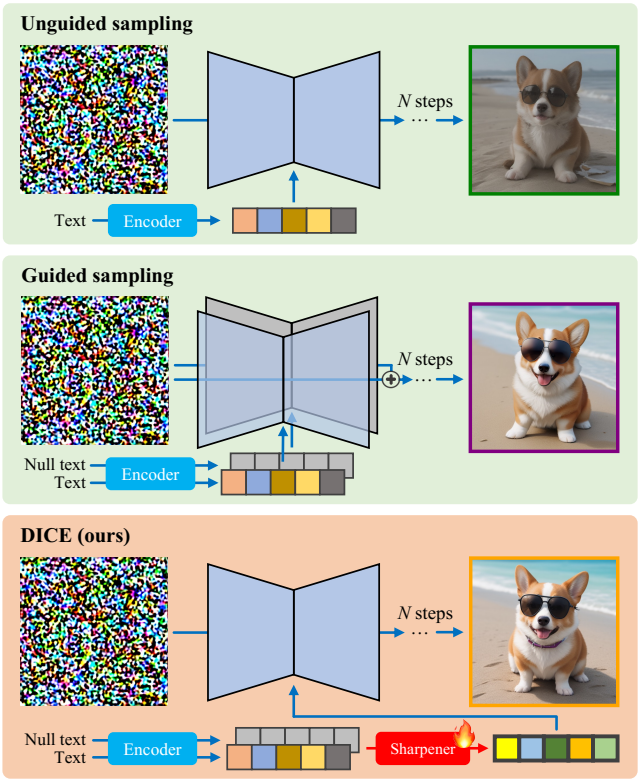


Figure 2: Overview of DICE sampling and comparison with traditional unguided and guided sampling. With sharpened text embeddings, DICE achieves high-quality image generation comparable to guided sampling while maintaining the same computational overhead as unguided sampling.

Preliminaries

Diffusion Models

Given a data sample $\mathbf{x}_0 \in \mathbb{R}^d$ from the implicit target data distribution p_0 (in this case, the distribution of all natural images), the forward process in diffusion models gradually adds white Gaussian noise to the sample, following a stochastic differential equation (SDE) (Song et al. 2021): $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$, where $t \in [0, T]$, $\mathbf{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d, g(\cdot) : \mathbb{R} \to \mathbb{R}$ are drift and diffusion coefficients and $\mathbf{w}_t \in \mathbb{R}^d$ is the Wiener process (Oksendal 2013). The backward process in diffusion models achieves the data reconstruction through a reverse-time SDE, $\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g^2(t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t$, which shares the same marginal distributions $\{p_t\}_{t=0}^T$ with the forward process (Song et al. 2021). This reverse-time SDE has a *probability flow* ordinary differential equation (PF-ODE) counterpart (Song et al. 2021; Chen et al. 2024b), $\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)]\mathrm{d}t$. Following the parametrization in EDM (Karras et al. 2022), where $\mathbf{f}(\mathbf{x}_t, t) = \mathbf{0}$ and $g(t) = \sqrt{2t}$, we simplify the PF-ODE into

$$\mathrm{d}\mathbf{x}_t = -t\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\mathrm{d}t. \tag{1}$$

The analytically intractable $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is known as the *score function* (Hyvärinen 2005; Lyu 2009), which is typically estimated by either a score-prediction model $\mathbf{s}_\theta(\mathbf{x}_t)$, or a noise-prediction model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$, i.e.,

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \approx \mathbf{s}_\theta(\mathbf{x}_t) = -\frac{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)}{t}. \tag{2}$$

For simplicity, unless otherwise specified, we will drop the time dependence of the model subsequently to re-

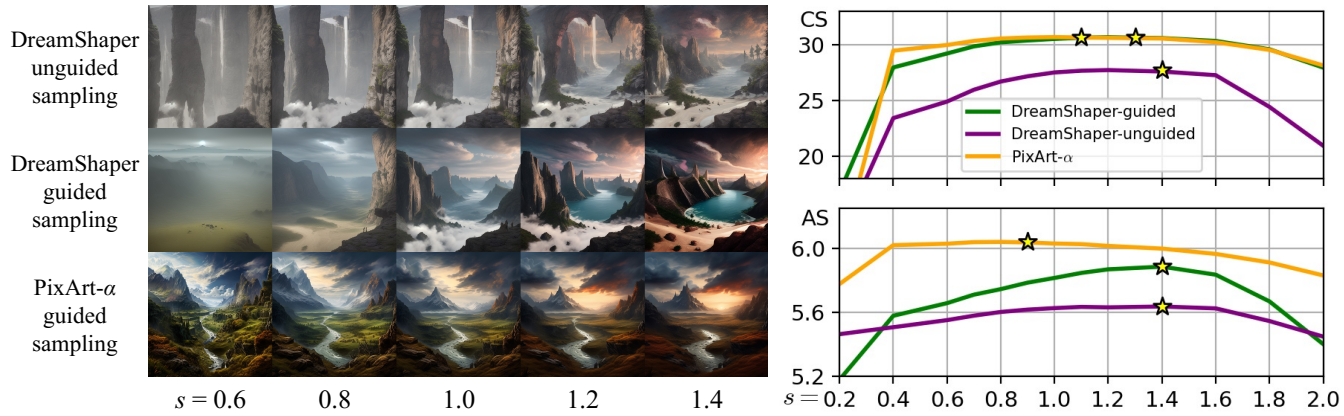


Figure 3: Text-image alignment with scaled text embeddings. Images are generated by DreamShaper (Lykon 2023), a popular variant of Stable Diffusion v1.5 (Rombach et al. 2022) with a CLIP text encoder (Radford et al. 2021), and PixArt-α (Chen et al. 2024c) with a T5-XXL text encoder (Raffel et al. 2020). *Left*: Text embeddings are scaled by a factor s and images are generated via unguided and guided sampling. *Right*: A grid search is conducted to identify the optimal scaling factor with respect to the CLIP score (CS) and Aesthetic score (AS). An optimal scaling factor improves the sample quality but varies across model. Meanwhile, naive scaling is insufficient to improve unguided sampling to the image quality achieved by guided sampling, which necessitates exploring the embedding space for a fine-grained dynamic scaling. Prompt: "*An epic landscape*".

duce notational clutter. The training objective of diffusion models is a weighted minimization of a regression loss (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Kingma and Gao 2024). For distillation tasks in which a student model $\boldsymbol{\epsilon}_\theta$ is supervised by a fixed teacher model $\tilde{\boldsymbol{\epsilon}}_{\tilde{\theta}}$, the training objective is defined as $\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,T),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\lambda(t)\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) - \tilde{\boldsymbol{\epsilon}}_{\tilde{\theta}}(\mathbf{x}_t)\|\right]$, where $\lambda(t)$ is a weighting function, $\mathbf{x}_t = \mathbf{x}_0 + t\boldsymbol{\epsilon}$, and $\mathbf{x}_0 \sim p_0$ follows the forward transition kernel $p_{0t}\left(\mathbf{x}_t|\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \mathbf{x}_0, t^2\mathbf{I}\right)$.

In text-to-image generation, the diffusion model receives embeddings of a text prompt $\mathbf{c} \in \mathbb{R}^{K\times d_e}$ encoded by a pre-trained text encoder to predict the score function conditioned on the text prompt, where K denotes the token number and d_e is the context dimension of each token. Starting from a random Gaussian noise $\mathbf{x}_T$ with a manually designed time schedule, sampling from diffusion models is to numerically solve $\mathrm{d}\mathbf{x}_t = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c})\mathrm{d}t$ through, for example, an Euler discretization (Song, Meng, and Ermon 2021),

$$\mathbf{x}_s = \mathbf{x}_t + (s-t)\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}), \tag{3}$$

where $0 \le s < t \le T$. Advanced numerical solvers using higher-order derivatives can also be employed to achieve accelerated sampling of diffusion models (Zhang and Chen 2023; Zhou et al. 2024).

Classifier-free Guidance

The standard class-conditional sampling for text-to-image generation with Equation 3 usually produces blurry, distorted, and semantically inaccurate images (Meng et al. 2023; Karras et al. 2024). In practice, classifier-free guidance (CFG) (Ho and Salimans 2022) is widely used to trade sample fidelity with diversity, allowing the model to achieve low-temperature sampling without the need for an auxiliary classifier-based guidance (Dhariwal and Nichol 2021). This technique modifies the model output by another model evaluation conditioned on a fixed null text embedding $\mathbf{c}_{\text{null}}$:

$$\boldsymbol{\epsilon}_\theta^{\omega,\mathbf{c}_{\text{null}}}(\mathbf{x}_t, \mathbf{c}) = \omega\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - (\omega-1)\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_{\text{null}}), \tag{4}$$

$$\mathbf{x}_s = \mathbf{x}_t + (s-t)\,\boldsymbol{\epsilon}_\theta^{\omega,\mathbf{c}_{\text{null}}}(\mathbf{x}_t, \mathbf{c}), \tag{5}$$

where $\omega \ge 1$ is known as the *guidance scale*, with $\omega = 1$ corresponding to unguided sampling, and $\omega > 1$ to guided sampling. Despite the ability to perform high-quality generation, CFG requires one more model evaluation in each guided sampling step, highly increasing the inference costs.

Method

Sharpening Text Embeddings by Scaling

Text-to-image diffusion models are trained on large datasets of text-image pairs (Rombach et al. 2022; Podell et al. 2024; Esser et al. 2024). In this process, text prompts are first encoded into embeddings using pre-trained text encoders (Radford et al. 2021; Raffel et al. 2020) and then integrated into the model inference via cross-attention modules. However, these models often struggle to generate images that closely align with the input prompts when using unguided sampling.

We hypothesize that this misalignment stems from two primary factors. First, current text encoders are not specifically designed for image generation. CLIP models align text and images in the embedding space via contrastive learning (Radford et al. 2021), while T5 models are fine-tuned on large-scale natural language processing tasks (Raffel et al. 2020). Neither is optimized to provide text embeddings tailored for high-quality image generation. Second, there is an inherent information imbalance between text and images. Images encapsulate rich details such as layout, texture, and fine-grained elements, whereas manually annotated captions typically describe only the main concepts (Radford

Algorithm 1: DICE Training

Input: dataset $\mathcal{D}$, guidance scale ω, maximum timestamp T, text-to-image model $\boldsymbol{\epsilon}_\theta(\cdot,\cdot)$, null text embedding $\mathbf{c}_{\text{null}}$, learning rate η
Initialize: sharpener $r_\phi(\cdot,\cdot)$
while not converged **do**
  Sample image-embedding pairs $(\mathbf{x}_0, \mathbf{c}) \sim D$
  Sample a timestamp $t \sim \mathcal{U}(0, T)$
  Forward diffusion process $\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_0, t^2 I)$
  $\boldsymbol{\epsilon}_\theta^{\omega,\mathbf{c}_{\text{null}}}(\mathbf{x}_t, \mathbf{c}) = \omega\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - (\omega-1)\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_{\text{null}})$
  $\mathbf{c}_\phi = \mathbf{c} + \alpha r_\phi(\mathbf{c}, \mathbf{c}_{\text{null}})$
  $\mathcal{L}(\phi) = \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_\phi) - \boldsymbol{\epsilon}_\theta^{\omega,\mathbf{c}_{\text{null}}}(\mathbf{x}_t, \mathbf{c})\|$
  $\phi \leftarrow \phi - \eta\nabla_\phi\mathcal{L}(\phi)$
end while

et al. 2021; Schuhmann et al. 2022). This disparity leads to a well-known modality gap between text and image domains (Liang et al. 2022; Schrodi et al. 2024), particularly when the text prompt length is limited. This may result in subpar sample quality in unguided sampling.

Instead of relying on CFG in the sampling process with double computational overhead, we improve the text-image alignment by sharpening the text embeddings. We begin by verifying the existence of such embeddings using the most straightforward approach: scaling. In Figure 3, we scale the text embeddings $\mathbf{c}$ input to the text-to-image models by a factor s and apply the scaled embeddings to both unguided (Equation 3) and guided (Equation 5) sampling. Sharpened text embeddings yield enriched image details and improved image contrast, but the optimal scaling factor varies across text-to-image models and text prompts. As illustrated in Figure 3, scaling factors of 0.6 and 1.4 can both enhance image details. Naive scaling alone is insufficient for improving unguided sampling to the level of image quality achieved by guided sampling. However, our pilot experiment demonstrates that while text-to-image models are trained on pretraiend text embeddings, they can generalize to a broader embedding space, making optimal sharpened text embeddings worth exploring. To learn the patterns of sharpened text embeddings that can more effectively improve text-image alignment, we propose training a lightweight neural network to dynamically scale the text embeddings.

DICE

We present **DI**stilling **C**FG by sharpening text **E**mbeddings (DICE) which enhances unguided sampling by aligning its sampling trajectory with the CFG trajectory. As such, DICE cuts the computational cost of CFG in half as it calls the denoising model only once per sampling step, while keeping the high generation quality of CFG. Specifically, given a text embedding $\mathbf{c}$ encoded by the text encoder, we train a lightweight sharpener $r_\phi(\cdot,\cdot) : \mathbb{R}^{(K\times d_e)\times(K\times d_e)} \to \mathbb{R}^{K\times d_e}$ with the trainable parameters ϕ, to sharpen the original text embedding, i.e.,

$$\mathbf{c}_\phi = \mathbf{c} + \alpha r_\phi(\mathbf{c}, \mathbf{c}_{\text{null}}), \tag{6}$$

where α is a hyperparameter controlling the sharpening strength. Similar to Equation 3, the unguided sampling becomes $\mathbf{x}_s = \mathbf{x}_t + (s-t)\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_\phi)$. We obtain the sharpened text embedding using CFG-based supervision while keeping the original text-to-image model frozen. Given image-embedding pairs $(\mathbf{x}_0, \mathbf{c})$, the training loss for the sharpener is formulated in a distillation manner as:

$$\mathbb{E}_{t\sim\mathcal{U}(0,T),\mathbf{x}_t\sim\mathcal{N}(\mathbf{x}_0,t^2\mathbf{I})}\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_\phi) - \boldsymbol{\epsilon}_\theta^{\omega,\mathbf{c}_{\text{null}}}(\mathbf{x}_t, \mathbf{c})\|, \tag{7}$$

where the trainable parameter is ϕ, and θ remains fixed. The training procedure is described in Algorithm 1. As shown in Figures 1 and 2, with the sharpened text embedding $\mathbf{c}_\phi$, DICE achieves high-quality image generation comparable to guided sampling while requiring only half the computation.

In text-to-image generation, descriptive text prompts are typically termed as positive prompts. However, images generated solely from positive prompts may not meet the desired quality standards. To address these issues, negative prompts are employed for image editing and quality enhancement. Previous works that distill CFG omit the entry for negative prompts, limiting practical applicability. In DICE, we can integrate the embedding of negative text prompts $\mathbf{c}_n$ into the sharpener, which is achieved by $\mathbf{c}_\phi = \mathbf{c} + \alpha r_\phi(\mathbf{c}, \mathbf{c}_n) - \beta(\mathbf{c}_n - \mathbf{c}_{null})$ where β is a hyperparamter controlling the strength of the introduced semantic shift. During training, negative prompts are randomly sampled from open-source datasets, and the training process remains consistent with Algorithm 1, except that negative text embeddings replace all the original null text embeddings. This strategy is especially effective for Stable Diffusion v1.5 variants (Rombach et al. 2022) and we consider it as an optional choice to endow DICE sharpener with better robustness to the semantic shift (see the Appendix).

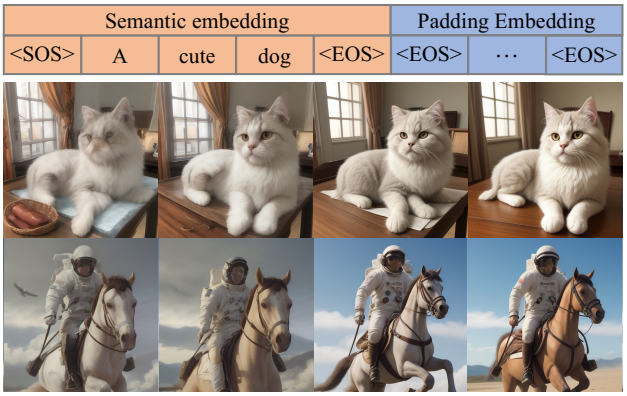


Figure 4: *Top*: a text embedding consists of a semantic and padding embedding. *Bottom*: replacing the original text embedding with the sharpened semantic and padding embedding. The latter one largely improves the sample quality.

Inspecting the sharpened Text Embedding

Compared to existing works that distill CFG (Meng et al. 2023; Hsiao et al. 2024), decoupling the sharpener from the text-to-image model allows us to gain a deeper understanding of the proposed method by focusing on analyzing the

Model	NFE	# Param	FID (↓)	CS (↑)	AS (↑)	HPS v2.1 (↑)				DrawBench (↑)
						Anime	Concept	Painting	Photo	
SD15 ($\omega = 5$)	40	-	22.04	30.22	5.36	24.29	23.16	22.88	24.62	23.83
SD15 ($\omega = 1$)	20	-	32.80	21.99	5.03	17.79	17.69	17.40	19.41	18.43
Scaling ($s = 1.2$) $*$	20	-	32.54	22.89	5.13	18.11	17.94	17.73	19.57	18.80
GD† (Meng et al. 2023)	20	859M	23.54	28.02	5.30	21.84	20.58	20.19	23.48	21.99
PnP† (Hsiao et al. 2024)	≈ 28	361M	26.57	27.72	**5.39**	**23.17**	**21.72**	**22.03**	24.17	23.12
DICE (ours)	20	2M	**22.22**	**28.54**	5.28	22.78	20.67	20.71	**24.96**	**23.32**
DreamShaper ($\omega = 5$)	40	-	30.35	30.50	5.87	30.20	28.92	28.85	27.62	26.84
DreamShaper ($\omega = 1$)	20	-	24.17	27.22	5.74	24.42	24.44	24.61	23.56	22.05
Scaling ($s = 1.3$) $*$	20	-	**24.05**	27.74	5.73	24.63	24.44	24.47	23.66	22.19
GD† (Meng et al. 2023)	20	859M	32.53	28.48	5.86	28.34	27.50	27.59	26.40	25.27
PnP† (Hsiao et al. 2024)	≈ 28	361M	35.57	28.46	**5.87**	**29.53**	**28.72**	**28.80**	**27.35**	**26.04**
DICE (ours)	20	2M	30.36	**29.03**	**5.87**	29.17	28.44	28.49	27.27	25.77
SDXL ($\omega = 5$)	40	-	23.95	32.10	5.60	29.67	28.19	28.19	26.51	26.03
SDXL ($\omega = 1$)	20	-	61.19	21.92	5.59	19.64	19.23	19.92	18.74	17.65
Scaling ($s = 1.5$) $*$	20	-	59.14	23.50	5.60	20.33	20.03	20.51	19.07	17.94
GD† (Meng et al. 2023)	20	2.6B	28.88	**30.84**	5.57	28.83	27.65	28.11	26.39	25.43
PnP† (Hsiao et al. 2024)	≈ 30	1.3B	32.52	30.31	**5.76**	**29.29**	27.59	**28.15**	26.44	25.35
DICE (ours)	20	3M	**28.01**	30.63	5.68	29.06	**27.72**	28.10	**26.48**	**25.44**
Pixart-α ($\omega = 5$)	40	-	38.39	30.67	6.03	31.43	29.97	29.60	28.97	27.95
Pixart-α ($\omega = 1$)	20	-	41.74	25.30	**6.11**	26.29	25.73	25.90	23.63	23.23
Scaling ($s = 1.2$) $*$	20	-	41.89	25.79	6.10	26.26	25.60	25.60	23.73	23.25
GD† (Meng et al. 2023)	20	611M	42.77	28.52	6.06	28.94	27.09	27.62	26.68	26.04
PnP† (Hsiao et al. 2024)	≈ 30	295M	40.06	**29.55**	5.99	29.29	28.24	27.96	26.55	25.55
DICE (ours)	20	5M	**39.80**	29.51	6.01	**30.10**	**28.59**	**28.69**	**27.91**	**26.60**

Table 1: Comparison of quantitative results. Images are generated with the same random seeds by the 20-step DPM-Solver++ (Lu et al. 2022). $*$: Naive scaling using searched optimal scaling factor. †: Our reimplementmentation of Guided Distillation (GD) (Meng et al. 2023) and Plug-and-Play Distillation (PnP) (Hsiao et al. 2024). PnP trains a ControlNet (Zhang, Rao, and Agrawala 2023) which introduces near half of the parameters of the base models and thus leads to larger NFE.

sharpened text embeddings for inference. Next, we investigate the underlying mechanisms of our method and demonstrate how the sharpened text embeddings influence sample quality and sampling dynamics through both quantitative and qualitative evidence.

The text embedding used for text-to-image generation consists of a $<SOS>$ token (start of sentence), some semantic tokens and the remaining padded $<EOS>$ tokens (end of sentence). As shown by previous works, e.g., (Yu et al. 2024), based on the position of the first $<EOS>$ token, a text embedding can be divided into a semantic embedding that contains most semantic information and a padding embedding that encodes more about the image details. In Figure 4, we replace the original embedding with sharpened semantic and padding embeddings. To replace the padding embedding, we recognize the index of the first $<EOS>$ token and then replace the embedding after this token with a sharpened one. It is observed that sharpened padding embeddings largely improve the image quality. Moreover, we compute the cosine similarity between $1,000$ paired original and sharpened semantic embeddings, obtaining a mean value of 0.75 and a standard deviation of 0.05, while for padding embeddings, they are 0.23 and 0.02. This indicates that padding embeddings are more significantly modified compared to semantic ones. Combining both qualitative and quantitative results, we conclude that DICE mainly emphasizes sharpening the padding embedding while maintaining the original semantic embedding, leading to consistent semantic information but significantly improved image details.

Experiments

Text-to-Image Generation

DICE's sharpener consists of two fully-connected layers and an attention block. The number of trainable parameters is less than 1% of the text-to-image model, leading to a negligible increase in computational overhead. The sharpening strength $\alpha = 1$ and guidance scale $\omega = 5$ are fixed during training. Our experiments are conducted on state-of-the-art text-to-image generation models, namely, Stable Diffusion v1.5 (SD15) (Rombach et al. 2022), Stable Diffusion XL (SDXL) (Podell et al. 2024), Pixart-α (Chen et al. 2024c) and a series of SD15-based open source variants, including DreamShaper[1], AbsoluteReality[2], Anime Pastel Dream[3], DreamShaper PixelArt[4], and 3D Animation Diffusion[5]. We use MS-COCO 2017 (Lin et al. 2014) for training

[1]https://huggingface.co/Lykon/DreamShaper

[2]https://huggingface.co/digiplay/AbsoluteReality_v1.8.1

[3]https://huggingface.co/Lykon/AnimePastelDream

[4]https://civitai.com/models/129879/dreamshaper-pixelart

[5]https://civitai.com/models/118086?modelVersionId=128046

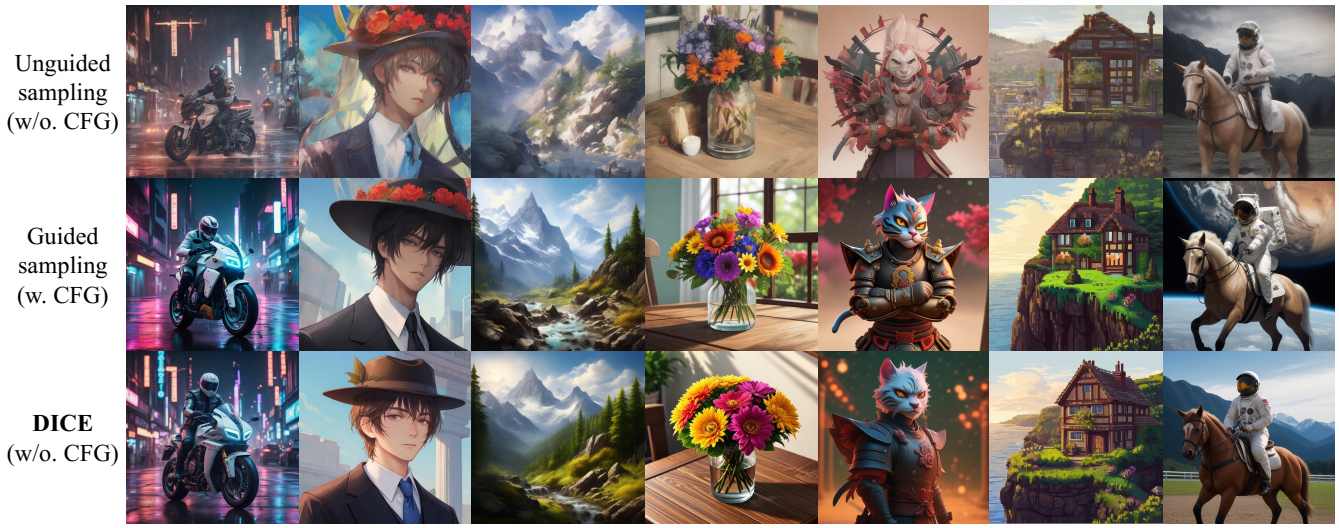


Figure 5: Qualitative results with different model capacities, image styles and network architectures. Images are generated by 20-step DPM-Solver++ (Lu et al. 2022) on 7 text-to-image models including multiple SD15 variants (Rombach et al. 2022), SDXL (Podell et al. 2024) and Pixart-α (Chen et al. 2024c). The used prompts are provided in the Appendix.

and evaluation. More details about training, evaluation and pre-trained models are included in the Appendix.

We evaluate DICE on text-to-image models with varying capacities, ranging from 0.6B to 2.6B parameters, across architectures such as U-Net (Ronneberger, Fischer, and Brox 2015) and DiT (Peebles and Xie 2023), and across diverse image styles including dreamlike, realistic, 3-D, pixel art, and anime style. The sample quality are measured by the Fréchet Inception Distance (FID) (Heusel et al. 2017), CLIP Score (CS) (Radford et al. 2021), Aesthetic Score (AS) (Schuhmann et al. 2022), HPS v2.1 (Wu et al. 2023) and DrawBench (Saharia et al. 2022). Quantitative results are presented in Table 1. Our enhanced unguided sampling achieves sample quality comparable to that of guided sampling and largely outperforms the original unguided sampling, as illustrated in Figure 5. Moreover, with only text embedding modified, DICE achieves performance comparable to existing method (Meng et al. 2023; Hsiao et al. 2024) with largely reduced trainable hyperparameters and without increasing inference costs.

Discussion and Ablation Study

Sharpening strength α. In practical applications, CFG offers flexibility in controlling image quality by adjusting the guidance scale. Although DICE maintains a fixed guidance scale during training, it allows for this flexibility via the sharpening strength α. This capability stems from the underlying mechanism of DICE, which emphasizes on enhancing image details while preserving semantic information. Figure 6 presents a comprehensive evaluation, demonstrating that the sharpening strength α serves a role akin to that of the guidance scale ω.

Generalization. As the sharpener operates independently of the text-to-image model, we investigate the feasibility of applying a well-trained sharpener to unseen text-to-image models and text prompts. In Figure 7, we separately train three sharpeners (sharpener i, $i = 1, 2, 3$) on three distinct text-to-image models, i.e., DreamShaper (model 1), DreamShaper PixelArt (model 2), and Anime Pastel Dream (model 3). Subsequently, we plug each sharpener into all models for unguided image generation. The results show that the sharpeners exhibit strong generalization capabilities across diverse domains, consistently and significantly improving the original unguided sampling. In Figure 8, we further investigate the generalization ability of DICE on unseen prompts outside the training dataset. We test the performance of DICE on unusual and long prompts and find DICE closely mimics the behavior of guided sampling and generalizes well to challenging text prompts.

Related Works

CFG distillation. Previous works have proposed distilling CFG-based text-to-image models. Guided distillation (Meng et al. 2023) incorporates the guidance scale as a new model input through fine-tuning. Plug-and-Play (Hsiao et al. 2024) trains an auxiliary guided model attached to the U-Net decoder, which is transferable to new domains. A recent work NoiseRefine (Ahn et al. 2024) proposes to refine the initially sampled Gaussian noise to enhance unguided sampling. However, the distillation loss requires samples generated by both unguided and guided sampling, introducing extensive computational overhead during training. In contrast, our method solely modifies the text conditioning without altering the generative process of diffusion models and retains fast training speed. A more detailed comparison is provided in the Appendix.

Reward-based methods. The text encoder plays a crucial role in text-to-image generation. Several studies aim to improve guided sampling by fine-tuning the text encoder through reinforcement learning (Chen et al. 2024a) and re-

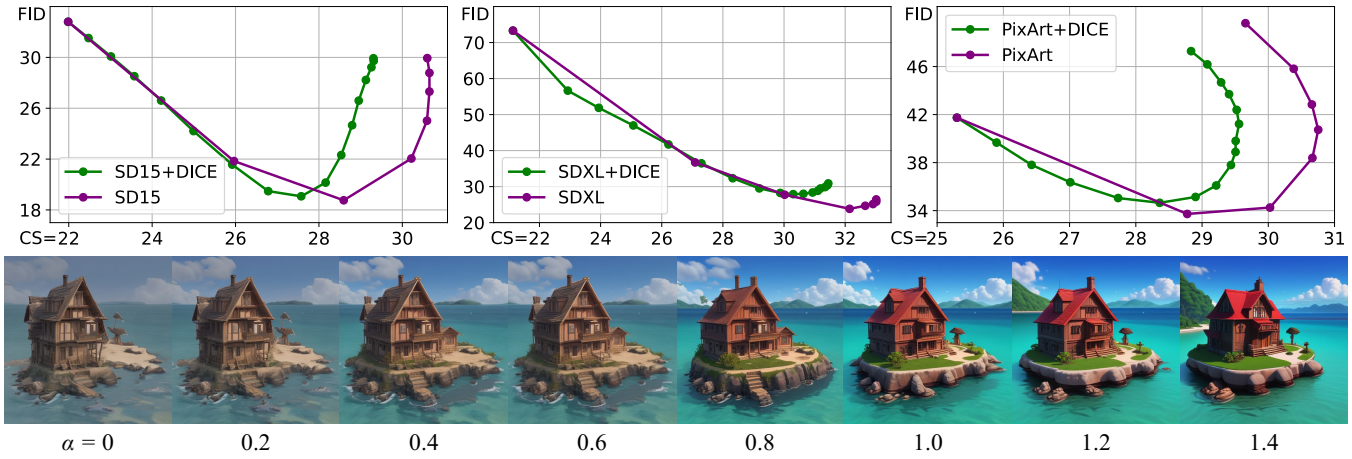


Figure 6: *Top*: FID-CS curves over guidance scale (for guided sampling) and sharpening strength α (for DICE) on different text-to-image models. The sharpening strength acts like the guidance scale. Guidance scales: $\{1, 1.5, 2.5, 5, 7.5, 10, 12.5, 15\}$. Sharpening strengths: $\{0, 0.1, 0.2, \cdots, 1.6\}$. *Bottom*: as α increases, the sharpener can maintain the semantic information while improving the sample quality.

Figure 7: Generalization across different models. The original unguided sampling results are provided for comparison. Each sharpener is independently trained on DreamShaper (model 1), DreamShaper PixelArt (model 2), or Anime Pastel Dream (model 3), and applied to other models.

Figure 8: Generalization of DICE to unusual and long text prompts. DICE closely mimics the behavior of guided sampling and generalizes well to unseen text prompts. The unusual prompts are "*A blue apple*" and "*A cubic watermelon*". The detailed long prompts are provided in the Appendix.

ward propagation (Li et al. 2024). Our method differs in two key aspects. First, it is specifically designed to improve unguided sampling without relying on CFG. Second, it is trained under CFG-based supervision and does not require human feedback or any reward models. We include further discussion in the Appendix.

Conclusion

Classifier-Free Guidance (CFG) is a prevalent technique in text-to-image generation, enhancing image quality but introducing increased sampling overhead. In this work, we introduce DICE, which fortifies text embeddings by training an sharpener under CFG-based supervision, achieving efficient and effective unguided text-to-image generation. We reveal that DICE enhances fine-grained image details through a universal enhancement pattern without compromising essential semantic information. Extensive experiments across various model capacities, image styles, and architectures demonstrate the effectiveness of our method. Our approach also exhibits strong generalization capability on unseen text-to-image models and challenging text prompts.

Limitations. Similar to existing methods that distill the CFG, the performance of DICE has not yet converged to the level of guided sampling. To overcome this limitation, future work will focus on enhancing our method beyond guided sampling by mitigating the information loss caused by distillation. Exploring ways to improve unguided sampling without CFG-based supervision is also a promising direction.

Acknowledgments

Zhenyu Zhou and Can Wang are supported by the National Natural Science Foundation of China (No. 62476244), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study, China (Grant No: SN-ZJU-SIAS-001) and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

Ahn, D.; Kang, J.; Lee, S.; Min, J.; Kim, M.; Jang, W.; Cho, H.; Paul, S.; Kim, S.; Cha, E.; et al. 2024. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*.

Bao, F.; Nie, S.; Xue, K.; Cao, Y.; Li, C.; Su, H.; and Zhu, J. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22669–22679.

Bradley, A.; and Nakkiran, P. 2024. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*.

Chen, C.; Wang, A.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024a. Enhancing diffusion models with text-encoder reinforcement learning. In *European Conference on Computer Vision*, 182–198. Springer.

Chen, D.; Zhou, Z.; Wang, C.; Shen, C.; and Lyu, S. 2024b. On the Trajectory Regularity of ODE-based Diffusion Sampling. In *International Conference on Machine Learning*, 7905–7934. PMLR.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2024c. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*.

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

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. In *Advances in Neural Information Processing Systems*, 6626–6637.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.

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

Hsiao, Y.-T.; Khodadadeh, S.; Duarte, K.; Lin, W.-A.; Qu, H.; Kwon, M.; and Kalarot, R. 2024. Plug-and-Play Diffusion Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13743–13752.

Hyvärinen, A. 2005. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6: 695–709.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*.

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*.

Kingma, D.; and Gao, R. 2024. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36.

Li, Y.; Liu, X.; Kag, A.; Hu, J.; Idelbayev, Y.; Sagar, D.; Wang, Y.; Tulyakov, S.; and Ren, J. 2024. Textcraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7985–7995.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

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.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Lykon. 2023. DreamShaper. https://huggingface.co/Lykon/DreamShaper.

Lyu, S. 2009. Interpretation and Generalization of Score Matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 359–366.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14297–14306.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Oksendal, B. 2013. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

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.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

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. In *Advances in Neural Information Processing Systems*, 36479–36494.

Schrodi, S.; Hoffmann, D. T.; Argus, M.; Fischer, V.; and Brox, T. 2024. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Representation Learning. *arXiv preprint arXiv:2404.07983*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*.

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*.

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*.

Yu, H.; Luo, H.; Wang, F.; and Zhao, F. 2024. Uncovering the Text Embedding in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2404.01154*.

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

Zhang, Q.; and Chen, Y. 2023. Fast Sampling of Diffusion Models with Exponential Integrator. In *International Conference on Learning Representations*.

Zheng, C.; and Lan, Y. 2024. Characteristic Guidance: Nonlinear Correction for Diffusion Model at Large Guidance Scale. In *Forty-first International Conference on Machine Learning*.

Zhou, Z.; Chen, D.; Wang, C.; and Chen, C. 2024. Fast ODE-based Sampling for Diffusion Models in Around 5 Steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.