DIVERSE TEXT-TO-IMAGE GENERATION VIA CONTRASTIVE NOISE OPTIMIZATION

Anonymous authors
Paper under double-blind review

(a) DDIM (b) CADS (c) Ours

Figure 1: **Example results from our diverse image generation approach.** Three distinct prompts are used: (top) “A person skiing on a very snowy slope”, (middle) “A cow sits in a truck with hay barrels in it”, and (bottom) “A man sitting on a couch next to a dog”. Standard DDIM (a) exhibits pronounced mode collapse, producing repetitive images and often failing to capture complex compositional details. CADS (Sadat et al., 2024) (b) improves diversity but still yields limited variation and occasional prompt misalignment. Our method (c) delivers markedly greater diversity and fidelity, generating a wide range of images that remain strongly aligned with the input text.

ABSTRACT

Text-to-image (T2I) diffusion models have demonstrated impressive performance in generating high-fidelity images, largely enabled by text-guided inference. However, this advantage often comes with a critical drawback: limited diversity, as outputs tend to collapse into similar modes under strong text guidance. Existing approaches typically optimize intermediate latents or text conditions during inference, but these methods deliver only modest gains or remain sensitive to hyperparameter tuning. In this work, we introduce Contrastive Noise Optimization, a simple yet effective method that addresses the diversity issue from a distinct perspective. Unlike prior techniques that adapt intermediate latents, our approach shapes the initial noise to promote diverse outputs. Specifically, we develop a contrastive loss defined in the Tweedie data space and optimize a batch of noise latents. Our contrastive optimization repels instances within the batch to maximize diversity while keeping them anchored to a reference sample to preserve fidelity. We further provide theoretical insights into the mechanism of this preprocessing to substantiate its effectiveness. Extensive experiments across multiple T2I backbones demonstrate that our approach achieves a superior quality-diversity Pareto frontier while remaining robust to hyperparameter choices.

1 INTRODUCTION

In recent years, diffusion models (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022) have emerged as the leading paradigm for text-to-image (T2I) generation. A key driver of their success is the use of text-guided inference, which steers the generation process to produce images that are not

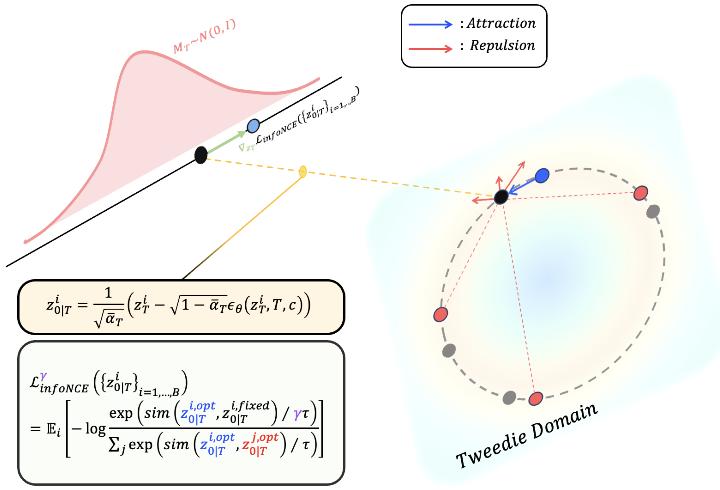


Figure 2: **Conceptual overview of contrastive noise optimization.** Our method enhances generation diversity by optimizing the initial latent vectors, $\mathbf{z}_T$, prior to the DDIM sampling process. We employ an InfoNCE loss that operates on a batch of noise vectors. This loss function pushes the optimizing sample (blue dot) away from all other negative samples in the batch to maximize separation. To preserve semantic fidelity, this repulsion is counterbalanced by an attraction force that pulls the anchor towards its original, non-optimized version (the positive pair), which acts as a fixed reference point. The attraction coefficient γ regulates this anchoring force, stabilizing the fidelity-diversity trade-off. This pre-processing step effectively diversifies the final image outputs without fine-tuning or altering the foundational diffusion sampler.

only high-fidelity but also closely aligned with a given prompt. To maximize this alignment and enhance image quality, practitioners often employ strong guidance mechanisms, with techniques like Classifier-Free Guidance (CFG) (Ho & Salimans, 2021) becoming a standard practice. However, this pursuit of high fidelity comes at a significant cost: a pronounced lack of diversity. Under strong textual guidance, the model's outputs often collapse into a few dominant modes, failing to capture the rich variety of interpretations a text prompt can have. This fidelity-diversity trade-off (Dhariwal & Nichol, 2021) remains a critical bottleneck, severely restricting the creative potential of T2I models.

To address this challenge, a common line of work has focused on interventions during the iterative denoising process. These approaches typically optimize intermediate latents (Corso et al., 2024; Kirchhof et al., 2025) or manipulate text embeddings (Sadat et al., 2024; Um & Ye, 2025b) to enforce separation between samples, while other strategies rely on multi-agent systems or complex fine-tuning schedules (Ghosh et al., 2017). While these methods have shown promise, they often provide only modest gains in diversity and introduce significant practical challenges. Their limitations include high implementation complexity, potential instability, and a heavy reliance on fragile hyperparameter tuning, which hinders their robustness and scalability.

In this work, we tackle the problem at its fundamental source by shaping the initial noise distribution. We introduce **Contrastive Noise Optimization**, a simple yet powerful pre-processing framework that optimizes a batch of initial noise latents to be inherently distinct before the diffusion process begins. Our main contribution is a contrastive objective defined in the data space of denoised predictions via Tweedie's formula (Robbins, 1992). Specifically, the objective integrates two complementary forces: a **repulsion** term that pushes Tweedie predictions from different noises apart, maximizing semantic distance to promote diversity; and an **attraction** term that anchors each optimized noise to its unoptimized counterpart, preserving semantic fidelity and preventing distributional shift.

We highlight that unlike existing approaches that impose diversity loss in the latent space (*e.g.*, the space of $\mathbf{z}_T$), our method applies the loss in the denoised Tweedie space, which provides the best estimate of clean generated data, thereby maximizing effectiveness. See Figure 2 for an illustration. To provide theoretical intuitions, we extend the traditional mutual information bound of the InfoNCE loss to incorporate both positive and negative pairs, thereby offering a unified explanation of how our contrastive objective simultaneously promotes diversity and preserves semantic fidelity. Comprehensive experiments confirm that this simple preprocessing substantially enhances the ability of modern T2I models to generate visually distinct images with minimal loss in quality or text–image alignment. We highlight that our method requires neither heavy computation nor extensive hyperparameter tuning, unlike existing approaches (Sadat et al., 2024; Um & Ye, 2025b).

Our key contributions are summarized as follows:

- We introduce **Contrastive Noise Optimization**, a simple yet powerful pre-processing framework that optimizes initial noise latents to be inherently distinct in the denoised Tweedie space before the diffusion process begins, tackling the diversity issue at its fundamental source.
- We provide theoretical insights by extending the mutual information bound of the InfoNCE loss to incorporate both positive and negative pairs, offering a unified explanation of how our contrastive objective simultaneously promotes diversity and preserves semantic fidelity.
- We show through comprehensive experiments that this lightweight pre-processing significantly boosts visual diversity with minimal quality or text-image alignment loss, while remaining robust to hyperparameters and broadly applicable across modern T2I backbones.

2 Related Work

Improving the diversity of diffusion models has recently attracted much attention, mainly due to their increasing use in critical applications such as text-to-image generation (Rombach et al., 2022). One prominent effort is CADS (Sadat et al., 2024), which enhances sample diversity by gradually annealing noise perturbations on conditional embeddings. Although effective, their approach is sensitive to the noise annealing schedule and requires laborious hyperparameter searches to see the diversity gain. A fundamentally different strategy is seen in Particle Guidance (PG) (Corso et al., 2024). The idea is to repel intermediate latent samples that share the same condition, thereby encouraging the final generated samples to exhibit distinct features. While it does not require difficult parameter searches like CADS, it often provides limited diversity gain (Kirchhof et al., 2025).s This approach shares a similar spirit as PG (Corso et al., 2024) and incorporates diversity-improving guidance for repelling intermediate latent instances during inference, yet in a sparse manner, *i.e.*, not at every inference timestep. A key distinction from ours is that its diversity optimization (by injecting guidance) is performed over inference time, which may be more expensive compared to ours that focuses on the initial latent space.

A related yet different task is to generate *minority* samples – low-density instances in the data manifold (Sehwag et al., 2022; Um & Ye, 2023; 2025b). Pioneer works in this area are offered by Sehwag et al. (2022); Um & Ye (2023), which share a similar idea of incorporating classifier guidance (Dhariwal & Nichol, 2021) to push intermediate samples toward low-density regions. The reliance on external classifiers was addressed in Um & Ye (2024; 2025b); Um et al. (2025), offering self-contained approaches for producing minority samples with diffusion models. While relevant, the task of generating low-density minority samples is distinct from improving diversity and does not guarantee distinct outputs. Notably, MinorityPrompt (Um & Ye, 2025b) considers text-to-image generation and provides a prompt optimization framework that can also be used to enhance the diversity of generated samples. However, it requires optimizing the diversity-improving prompt during inference, which imposes substantial computational overhead (Um & Ye, 2025b).

Initial noise optimization in diffusion models has been explored in various contexts (Guo et al., 2024; Ahn et al., 2024). One instance is InitNO (Guo et al., 2024), where the idea is to optimize the initial noise latent to promote improved prompt alignment in text-to-image generation. A distinction with respect to to ours is that their focus is on enhancing text adherence, unlike ours. Another notable work was done by Ahn et al. (2024), who aim to characterize the influence of classifier-free guidance (Ho & Salimans, 2021) through a properly optimized latent noise, enabled by an additional neural network that maps to the optimal noise. While interesting, their focus is inherently distinct from ours. To the best of our knowledge, our framework is the first to incorporate the idea of noise optimization for addressing the diversity challenge of diffusion models.

3 PRELIMINARIES

3.1 LATENT DIFFUSION MODELS

Latent Diffusion Models (LDMs) (Rombach et al., 2022) improve upon traditional Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) by performing the diffusion process in a computationally efficient, lower-dimensional latent space. LDMs first use a pre-trained autoencoder to map a high-resolution image $\mathbf{x}_0$ into a compressed latent representation, $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. The diffusion process is then applied directly to these latent vectors.

The forward process is a Markov chain that gradually adds Gaussian noise to an initial latent vector $\mathbf{z}_0$ over a series of T discrete timesteps. At each step t, the transition is defined as:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ is a fixed variance schedule that controls the noise level at each step. A key property of this process is that the marginal distribution at any arbitrary step t can be expressed in a closed form conditioned only on the initial latent $\mathbf{z}_0$:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where we define $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$. As t increases towards T, the signal term $\sqrt{\bar{\alpha}_t}$ approaches zero, and the variance $1 - \bar{\alpha}_t$ approaches 1. This ensures that the noised latent $\mathbf{z}_T$ reliably converges to an isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, regardless of the initial latent vector $\mathbf{z}_0$. Once the reverse process generates a clean latent, the decoder $\mathcal{D}$ is used to map it back to the pixel space.

3.2 REVERSE PROCESS AND DENOISING VIA TWEEDIE'S FORMULA

The generative process is achieved by reversing the forward process, conditioned on external information such as a text embedding $\mathbf{c}$ for Text-to-Image (T2I) synthesis. This involves learning a model $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})$ that approximates the true posterior. In DDPM (Ho et al., 2020), this conditional reverse process is parameterized as a Gaussian whose mean is learned by a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) = \mathcal{N}\left(\mathbf{z}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\right), \sigma_t^2\mathbf{I}\right). \tag{3}$$

The core of this process is the network $\boldsymbol{\epsilon}_\theta$, which is trained to predict the noise component from the noisy latent vector $\mathbf{z}_t$ based on the condition $\mathbf{c}$. The key insight is that this trained network can be used to directly estimate the original clean latent $\mathbf{z}_0$ at any timestep t. This denoised estimate $\hat{\mathbf{z}}_0$ is implemented via Tweedie's formula (Chung et al., 2025; Um & Ye, 2025a), which for our specific noise model takes the form:

$$\hat{\mathbf{z}}_0(\mathbf{z}_t, t, \mathbf{c}) := \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\right). \tag{4}$$

This equation forms the foundation of the iterative denoising process in many conditional diffusion models, allowing for the generation of latent vectors that align with the given context $\mathbf{c}$.

3.3 INFORMATION NOISE-CONTRASTIVE ESTIMATION (INFONCE)

Information Noise-Contrastive Estimation (InfoNCE) (van den Oord et al., 2019) is a fundamental objective for self-supervised representation learning (Chen et al., 2020; He et al., 2020). It aims to construct an embedding space that maximizes mutual information (Cover, 1999) between representations of positive (similar) pairs while minimizing it for negative (dissimilar) pairs. The learning process can be viewed as a classification task in which, for a given anchor sample, the model must correctly identify its positive counterpart from a set of negative samples.

Specifically, for an anchor embedding vector $\mathbf{z}_i$, its positive pair $\mathbf{z}_j$, and a set of $B-1$ negative samples $\{\mathbf{z}_k\}_{k=1,k\neq j}^B$, the InfoNCE loss is formulated as

$$\mathcal{L}_{\text{InfoNCE}} := \mathbb{E}_i\left[-\log\left(\frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^B \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right)\right], \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes a similarity measure (e.g., cosine similarity) between two representation vectors, and τ is a temperature parameter controlling the sharpness of the distribution. Intuitively, the loss encourages the anchor to be close to its positive pair while pushing it away from all negative samples, thereby tightening intra-class similarity and enlarging inter-class separation in the embedding space.

4 PROPOSED METHOD

4.1 OPTIMIZING INITIAL NOISE WITH CONTRASTIVE LOSS BETWEEN TWEEDIES

While text-to-image diffusion models generate high-quality images for a given prompt c, the standard approach of sampling a random initial noise $\mathbf{z}_T$ often yields outputs that collapse into similar modes. This tendency is a primary obstacle to achieving a diverse set of generations. To address this issue, we propose a novel approach that enhances diversity by optimizing the initial noise $\mathbf{z}_T$ itself, prior to commencing the standard DDIM sampling process. This procedure is detailed in **Algorithm 1** in Appendix B.1.

The algorithm proceeds as follows. First, we sample a batch of initial latent codes $\mathcal{B} = \{\mathbf{z}_T^i\}_{i=1}^B$ from a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. Immediately after, we compute the initial target latents, $\{\mathbf{z}_{0|T}^i\}_{i=1}^B$, by applying the denoising estimator defined in Equation Eq. (4) to this initial noise at timestep T. Each resulting $\mathbf{z}_{0|T}^i$ is therefore the model's one-step prediction of the clean latent $\mathbf{z}_0$ from the noise $\mathbf{z}_T^i$. These pre-computed latents then serve as fixed anchors, each defining a unique identity for its respective sample throughout the optimization.

Before computing the loss, we employ a practical optimization to enhance efficiency. The high dimensionality of the latents (B, C, S, S) makes the pairwise similarity calculation computationally intensive. We found experimentally that applying an adaptive average pooling operation to downsample the latents to a sufficiently smaller spatial resolution (B, C, w, w), where $w < S$, did not compromise performance (Section 5.1). This step substantially reduces memory usage and accelerates the similarity matrix computation, making the optimization process more tractable. Downsampled latents $\{\mathbf{z}_{0|T}^{i,\text{fixed}}\}_{i=1}^B$, $\{\mathbf{z}_{0|T}^{i,\text{opt}}\}_{i=1}^B$ are then normalized, and the noise $\{\mathbf{z}_T^i\}_{i=1}^B$ is updated using a contrastive loss $\mathcal{L}_{\text{CNO}}$:

$$\mathcal{L}_{\text{CNO}} \coloneqq \mathbb{E}_{i\sim\mathcal{B}}\left[-\log\frac{\exp(\mathrm{sim}(\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}})/\tau)}{\sum_{j\in\mathcal{B}}\exp(\mathrm{sim}(\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{j,\text{opt}})/\tau)}\right]. \tag{6}$$

This loss function is designed to achieve two objectives simultaneously.

Attraction (Numerator). It encourages the current latent $\mathbf{z}_{0|T}^{i,\text{opt}}$ to remain similar to its corresponding initial target latent $\mathbf{z}_{0|T}^{i,\text{fixed}}$. This ensures that each sample maintains coherence with its initial concept and does not drift away during optimization.

Repulsion (Denominator). It pushes the current latent $\mathbf{z}_{0|T}^{i,\text{opt}}$ to be dissimilar from all other current latents $\{\mathbf{z}_{0|T}^{j,\text{opt}}\}_{j=1,j\neq i}^B$ in the batch. This directly promotes diversity by forcing the latent representations to disperse within the batch.

By iteratively updating $\mathbf{z}_T$ with the gradient of this loss ($\nabla_{\mathbf{z}_T}\mathcal{L}_{\text{CNO}}$), we guide the initial noise vectors to positions in the latent space that are predisposed to generating a diverse set of images. Once the optimization is complete, this well-distributed batch of noise $\{\mathbf{z}_T^i\}_{i=1}^B$ is fed into a standard, pre-trained DDIM denoiser to produce the final images. Consequently, our method effectively enhances output diversity through a simple pre-processing stage that modulates the starting point of the generation, all without requiring any modifications to the pre-trained diffusion model itself.

Stop-gradient for computational efficiency. Optimizing the loss in Eq. (6) requires backpropagation through diffusion models, which can incur substantial computational overhead. To mitigate this, we apply a `stopgrad` operator (Chen & He, 2021) on the model path used in computing the Tweedie's estimate. As also demonstrated in Ahn et al. (2024), this simple strategy yields significant savings in training cost with only marginal impact on performance (see Table 4).

4.2 GAMMA EFFECT: REGULATED ATTRACTION FOR STABLE IMAGE DIVERSIFICATION

In our proposed algorithm, the InfoNCE loss for a single sample within a batch of size B consists of one attraction term (to itself) and $B-1$ repulsion terms (from all other samples in the batch). When the batch size B is large, the cumulative repulsion force can become excessively strong. This risks pushing the optimized noise out of the intended distribution, potentially leading to the generation of less plausible or out-of-distribution images.

To mitigate this issue and achieve a more stable optimization, we introduce a coefficient, γ, to dynamically regulate the attraction force. This is done by dividing the similarity term in the numerator of the loss function by γ. The modified InfoNCE loss is as follows:

$$\mathcal{L}_{\text{CNO}}^{\gamma} \coloneqq \mathbb{E}_{i\sim\mathcal{B}}\left[-\log\frac{\exp(\text{sim}(\mathbf{z}_{0|T}^{i,\text{opt}},\mathbf{z}_{0|T}^{i,\text{fixed}})/\gamma\tau)}{\sum_{j\in\mathcal{B}}\exp(\text{sim}(\mathbf{z}_{0|T}^{i,\text{opt}},\mathbf{z}_{0|T}^{j,\text{opt}})/\tau)}\right]. \tag{7}$$

Empirically, we found that γ in our framework behaves similarly to a Gaussian regularizer (Guo et al., 2024), which penalizes large deviations from the Gaussian prior. A detailed analysis is provided in Section C.2.

Desirable value for γ. The desirable value for γ is derived by creating a balance between the regulated attraction force and the cumulative repulsion forces. We achieve this by equating the maximum value of the attraction term (numerator) with the sum of the maximum values of the $B-1$ repulsion terms. Assuming the maximum similarity score is 1, this balance can be expressed as:

$$\exp(1/\gamma\tau) = (B-1)\exp(1/\tau). \tag{8}$$

Solving for γ gives us the following relationship:

$$\gamma = (\tau ln(B-1)+1)^{-1}. \tag{9}$$

For instance, in our common experimental setting where $\tau = 0.1$ and $B = 5$, the calculated γ is approximately 0.88, which is very close to the fixed value of $\gamma = 1.0$ we have consistently used. For a fixed $\tau = 0.1$, the optimal γ changes moderately with batch size B:

$$B = 13 \rightarrow \gamma \approx 0.8 \quad B = 73 \rightarrow \gamma \approx 0.7 \quad B = 775 \rightarrow \gamma \approx 0.6$$

This shows that as the batch size B grows larger, γ is not highly sensitive. Therefore, using a single, appropriately chosen fixed value for γ can also yield stable results without significant performance degradation.

4.3 THEORETICAL INTUITIONS

We provide mathematical insights into our contrastive framework by establishing its connection to mutual information. We begin with the classical view of InfoNCE as a variational lower bound on mutual information, as shown by van den Oord et al. (2019). Specifically, the InfoNCE loss in Eq. (5) satisfies

$$\mathcal{L}_{\text{InfoNCE}} \geq \log B - I(Z; Z_{\text{pos}}), \tag{10}$$

where $I(X;Y)$ is the mutual information between random variables X and Y. This inequality implies that minimizing $\mathcal{L}_{\text{InfoNCE}}$ indirectly maximizes $I(Z;Z_{\text{pos}})$, encouraging the learned embedding space to cluster positive pairs. However, this classical relationship does not clarify how negative pairs shape the embedding space—an aspect that is critical in our framework, where negative samples drive diversity.

To capture this effect, we augment the traditional bound to incorporate mutual information with respect to negative pairs. The following proposition formalizes this result.

Proposition 1. *The InfoNCE loss in Eq. (5) satisfies*

$$\mathcal{L}_{\textit{InfoNCE}} \;\geq\; -I(Z; Z_{\text{pos}}) + I(Z; Z_{\text{neg}}) + \log(B-1), \tag{11}$$

where B denotes the batch size, and $I(X;Y)$ is the mutual information between random variables X and Y:

$$I(X;Y) \coloneqq \mathbb{E}_{p(X,Y)}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right] = \mathbb{E}_{p(X,Y)}\left[\log\frac{p(X\mid Y)}{p(X)}\right].$$

Model	Method	Prec ↑	Rec ↑	Den ↑	Cov ↑	CLIP ↑	Pick ↑	IR ↑	MSS ↓	Vendi ↑
SD1.5	DDIM	0.7018	0.6706	0.6033	0.7382	**31.5863**	**21.5081**	**0.2222**	0.1657	4.6949
	PG	0.6940	0.7024	0.5975	0.7446	31.3222	21.2086	0.1712	0.1426	4.7630
	CADS	0.6866	**0.7240**	0.5686	0.7292	31.4863	21.2938	0.1137	0.1330	4.7805
	DiversityPrompt	0.6878	0.7006	0.5839	0.7416	31.5457	21.3510	0.1332	0.1393	4.7599
	Ours	**0.7308**	0.6926	**0.6528**	**0.7728**	31.4525	21.3779	0.1284	**0.1317**	**4.7855**
SDXL	DDIM	**0.6858**	0.6538	**0.5713**	0.7368	31.8788	**22.4761**	**0.7302**	0.2169	2.8377
	PG	0.5820	**0.7088**	0.3855	0.5606	31.5679	22.1631	0.6950	0.2050	2.8545
	CADS	0.6486	0.6796	0.5262	0.7108	**31.9424**	22.2078	0.6162	0.1765	2.8864
	Ours	0.6720	0.6992	0.5553	**0.7568**	31.8129	22.3859	0.7273	**0.1623**	**2.9019**
SD3	DDIM	0.7184	**0.5828**	0.6472	0.6770	31.7783	**22.5763**	1.0301	0.3028	4.2205
	PG	**0.7782**	0.3900	**0.8110**	**0.7370**	**32.0463**	22.3500	**1.0357**	0.3066	4.2097
	CADS	0.6984	0.5752	0.6110	0.6682	31.6974	22.4987	1.0233	0.2960	4.2487
	Ours	0.7100	0.5806	0.6573	0.6938	31.7713	22.5647	1.0233	**0.2909**	**4.2644**

Table 1: **Quantitative results of zero-shot diverse samplers.** Our proposed method is benchmarked against the standard DDIM sampler and state-of-the-art diversity-enhancing techniques: PG (Corso et al., 2024) and CADS (Sadat et al., 2024). DiversityPrompt refers to the prompt-optimization-based diversity approach developed in Um & Ye (2025b). The evaluation demonstrates that our approach consistently achieves superior performance in diversity metrics, including MSS(↓) and Vendi Score(↑), across Stable Diffusion 1.5, XL, and 3. Notably, it enhances diversity while effectively preserving image quality and prompt fidelity, successfully navigating the fidelity-diversity trade-off by optimizing the initial latent space.

The proof is provided in Section A.1. This result shows that the InfoNCE loss is inherently linked to negative samples as well as positive ones: minimizing the loss decreases the mutual information with negatives while increasing that with positives.

Extension with Gamma. We further analyze our modified loss in Eq. (7), which introduces a coefficient γ to control the relative strength of positive pairs.

Proposition 2. *For the loss function defined in Eq. (7), the following inequality holds:*

$$\mathcal{L}^{\gamma}_{InfoNCE} \geq -\frac{1}{\gamma} I(Z; Z_{pos}) + I(Z; Z_{neg}) + \log(B - 1). \tag{12}$$

The proof is given in Section A.2. This proposition indicates that γ scales the positive mutual information term, serving as a control knob to modulate the influence of positive pairs in our contrastive objective. We provide empirical results to demonstrate the impact of γ in the appendix; see Section C.1.

5 Experiments

Implementation details. Our experiments are conducted on three distinct pre-trained text-to-image diffusion frameworks: Stable Diffusion v1.5 (SD1.5), SDXL, and SD3. We compare our method with state-of-the-art zero-shot diversity samplers, including Condition-Annealed Diffusion Sampler (CADS) (Sadat et al., 2024) and Particle Guidance (PG) (Corso et al., 2024), as well as the prompt-optimization-based diversity method of Um & Ye (2025b), referred to as *DiversityPrompt*. All evaluations use text prompts randomly sampled from the MS-COCO (Lin et al., 2014) validation set. For each prompt, we generate 3–5 images, yielding a total of roughly 6–10 K samples.

Evaluation metrics. The goal of our research is to enhance the Pareto frontier between image quality and diversity of generated images while maintaining a high degree of relevance to the text prompt. To quantitatively assess this, we used the following key metrics: **CLIPScore, PickScore, Image-Reward** for evaluating image quality, and **Vendi Score, Mean Pairwise Similarity (MSS)** for diversity. Details for those metrics appear in Section B.2.

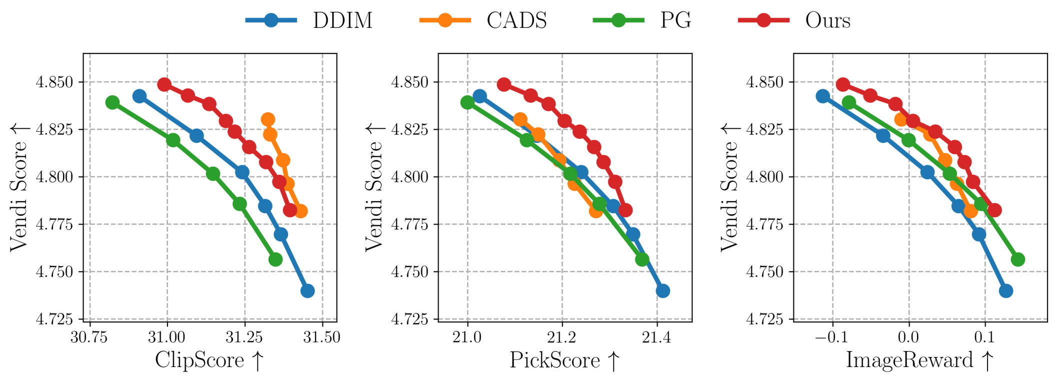


Figure 3: Pareto curves of diverse sampling methods between Vendi Score and text-to-image alignment metrics. For our methods, we use $N_{opt} = 5, \gamma = 1.0, w = 8, \tau = 0.1$ in common.

5.1 Results

Comparison with existing zero-shot diversity samplers. The results in Table 1 demonstrate the effectiveness of our approach. Our method achieves high performance on the key diversity metrics, Vendi Score and MSS, consistently outperforming all baselines across the different foundation models. While performance on Density and Coverage is highly competitive, our approach's strong results on Vendi Score and MSS prove its robust, model-agnostic ability to mitigate mode collapse.

Crucially, these substantial gains in diversity do not compromise generation quality. Our method maintains strong prompt fidelity, evidenced by competitive CLIP scores, and sustains a competitive or superior Pick-Score compared to CADS across all Stable Diffusion models. This indicates our outputs are not only more varied but also aesthetically preferable. This quantitative strength is mirrored in our qualitative results (see Figure 5), where our model shows particular strength on complex compositional prompts that cause competitor methods to fail. Where gains in diversity often come at the cost of quality, our method achieves both, delivering outputs that are not only more varied but also consistently high in fidelity.

Figure 3 illustrates the quality-diversity trade-off by plotting the Pareto frontiers for our method and key baselines. The plots reveal a clear and compelling advantage for our approach, which establishes a dominant frontier across metrics trained on large-scale human preferences. This is most evident in the **PickScore** and **Image-Reward** charts, where our method is strictly superior to competitors like CADS, indicating our outputs are more aesthetically pleasing for any given level of diversity.

While CADS may achieve a marginally higher peak CLIPScore, this metric is known to favor rigid semantic alignment rather than creative or aesthetically superior interpretations. In contrast, our method's dominance across both PickScore and Image-Reward demonstrates a more intelligent trade-off. It prioritizes what a human user would find visually appealing and contextually appropriate over a mechanical, word-for-word adherence to the prompt. This quantitative strength is mirrored in our qualitative results (see Appendix C.4), where our model uniquely succeeds on complex compositional prompts that cause competitors to fail.

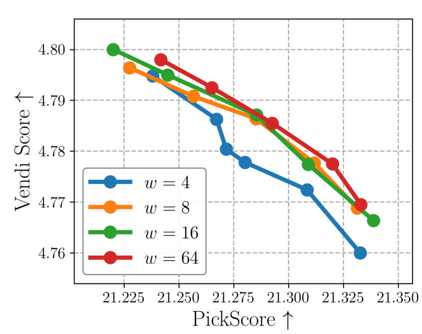
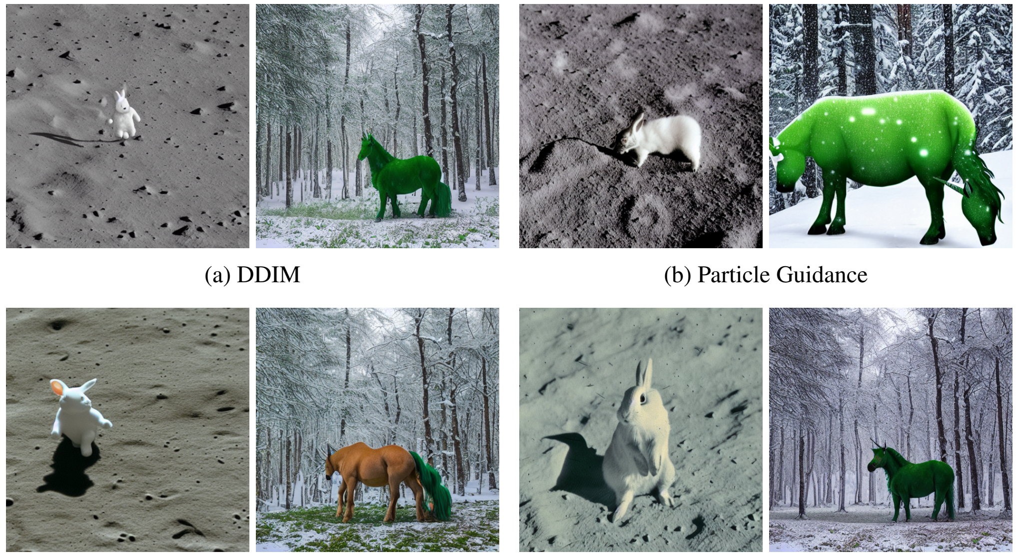


Figure 4: Ablation on the window size w. The Pareto frontier of PickScore vs. Vendi Score.

Effect of window size w. We conduct an ablation study to investigate the effect of the downsampling window size, w, applied to the Tweedie latent shape $\mathbf{z}^i_{0|T}$ (Algorithm 1, Line 10). This step is crucial for capturing the global structure of the initial noise prediction while reducing computational cost. We experiment with



(a) DDIM (b) Particle Guidance

(c) CADS (d) Ours

Figure 5: **Qualitative comparison with pre-existing zero-shot diverse generative methods** For the prompt *"A white rabbit on the moon."*(left) and *"A green unicorn in a snowy forest"*(right), we compare our method (d) with baseline approaches. Our method successfully generates high-fidelity images that are strongly aligned with the text prompts. In contrast, the other methods exhibit various failures.

$w \in \{4, 8, 16\}$ and compare these against the baseline of $w = 64$, which effectively uses the full-resolution latent shape.

The results are illustrated in the PickScore-Vendi Score Pareto frontier in Figure 4. As shown, an aggressive downsampling with $w = 4$ leads to a noticeable performance degradation, failing to match the frontier established by larger window sizes. In contrast, moderate downsampling with $w = 8$ **and** $w = 16$ **achieves highly competitive performance compared to the** $w = 64$ **baseline**.

This suggests that moderate downsampling successfully preserves the essential structural information for diversification while benefiting from increased computational efficiency. Excessive downsampling ($w = 4$), however, appears to discard critical details necessary for the optimization process. Based on these findings, we select $w = 16$ for our main experiments, as it provides the best trade-off between performance and efficiency.

6 Conclusion

We introduced **Contrastive Noise Optimization**, a simple yet effective pre-processing method to address mode collapse in text-to-image (T2I) diffusion models. By applying a contrastive loss directly to the initial noise vectors for a given text prompt, our approach ensures diverse starting points for generation, eliminating the need for the complex sampling guidance or laborious hyperparameter tuning required by prior work. Our method sets a new state-of-the-art on the quality-diversity Pareto frontier, outperforming strong baselines on key diversity metrics without compromising prompt fidelity or image quality.

For future work, while our method focuses on the initial noise $\mathbf{z}_T$, we believe that applying a similar optimization strategy to intermediate latents $\mathbf{z}_t$ (where $t < T$) could be a promising avenue for further enhancing diversity for a single prompt. Some studies have explored optimizing these intermediate latents to generate images with high fidelity to complex textual conditions (Wallace et al. 2023 ; Ding et al. 2024). The effectiveness of such an approach may depend on the model and the degree to which its Tweedie prediction is already structured to reflect the semantic content of the input prompt at intermediate timesteps. This direction may warrant deeper investigation on our approach.

REFERENCES

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=E77uvbOTtp`.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=KqbCvIFBY7`.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html`.

Sandra Zhang Ding, Jiafeng Mao, and Kiyoharu Aizawa. Training-free sketch-guided diffusion with latent optimization. *CoRR*, abs/2409.00313, 2024. doi: 10.48550/ARXIV.2409.00313. URL `https://doi.org/10.48550/arXiv.2409.00313`.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Trans. Mach. Learn. Res.*, 2023, 2023. URL `https://openreview.net/forum?id=g97OHbQyk1`.

Arnab Ghosh, Viveka Kulharia, Vinay P. Namboodiri, Philip H. S. Torr, and Puneet Kumar Dokania. Multi-agent diverse generative adversarial networks. *CoRR*, abs/1704.02906, 2017. URL `http://arxiv.org/abs/1704.02906`.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9380–9389. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00896. URL `https://doi.org/10.1109/CVPR52733.2024.00896`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595. URL `https://doi.org/10.18653/v1/2021.emnlp-main.595`.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and marco cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=XAckVo0iNj.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14512–14522. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01413. URL https://doi.org/10.1109/CVPR52688.2022.01413.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in statistics*, pp. 388–394. Springer, 1992.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022.

Jonathon Shlens. Notes on kullback-leibler divergence and likelihood. *CoRR*, abs/1404.2000, 2014. URL http://arxiv.org/abs/1404.2000.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Soobin Um and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. *arXiv preprint arXiv:2301.12334*, 2023.

Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *European Conference on Computer Vision*, pp. 414–430. Springer, 2024.

Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20926–20936. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.01949. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Um_Minority-Focused_Text-to-Image_Generation_via_Prompt_Optimization_CVPR_2025_paper.html`.

Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20926–20936, 2025b.

Soobin Um, Beomsu Kim, and Jong Chul Ye. Boost-and-skip: A simple guidance-free diffusion for minority generation. *arXiv preprint arXiv:2502.06516*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL `https://arxiv.org/abs/1807.03748`.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 7246–7256. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00669. URL `https://doi.org/10.1109/ICCV51070.2023.00669`.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html`.

A PROOFS

A.1 PROOF OF PROPOSITION 1

Proposition 1. *The InfoNCE loss in Eq. (5) satisfies*

$$\mathcal{L}_{\textit{InfoNCE}} \;\geq\; -I(Z; Z_{\text{pos}}) + I(Z; Z_{\text{neg}}) + \log(B-1), \tag{13}$$

where B denotes the batch size, and $I(X;Y)$ is the mutual information between random variables X and Y:

$$I(X;Y) := \mathbb{E}_{p(X,Y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] = \mathbb{E}_{p(X,Y)}\left[\log \frac{p(X \mid Y)}{p(X)}\right].$$

Proof. In general, we can formulate the infoNCE loss as Eq. (14):

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{p(\mathbf{z},\mathbf{z}_{\text{pos}},\mathbf{z}_{\text{neg}})}\left[\log \frac{f(\mathbf{z},\mathbf{z}_{\text{pos}})}{f(\mathbf{z},\mathbf{z}_{\text{pos}}) + \sum_{i=1}^{B-1} f(\mathbf{z},\mathbf{z}_{\text{neg}}^{(i)})}\right]. \tag{14}$$

Let us simply notate $\mathbf{z}_{\text{pos}}$, $\mathbf{z}_{\text{neg}}$ as $\mathbf{z}_p$, $\mathbf{z}_n$, respectively. This can be split into two terms:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{p(\mathbf{z},\mathbf{z}_p)}\left[\log f(\mathbf{z},\mathbf{z}_p)\right] + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_p,\mathbf{z}_n)}\left[\log\left\{f(\mathbf{z},\mathbf{z}_p) + \sum_{i=1}^{B-1} f(\mathbf{z},\mathbf{z}_n^{(i)})\right\}\right]. \tag{15}$$

We can easily assume that $f(\mathbf{z},\mathbf{z}_p) \geq 0$ by leveraging an exponential function in convention. Then, Eq. (15) has a lower bound by omitting the positive pair similarity from second term:

$$\mathcal{L}_{\text{InfoNCE}} \geq -\mathbb{E}_{p(\mathbf{z},\mathbf{z}_p)}\left[\log f(\mathbf{z},\mathbf{z}_p)\right] + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log \sum_{i=1}^{B-1} f(\mathbf{z},\mathbf{z}_n^{(i)})\right]. \tag{16}$$

According to van den Oord et al. (2019), function $f(\mathbf{z},\mathbf{z}')$ estimates the probability density ratio $\frac{p(\mathbf{z} \mid \mathbf{z}')}{p(\mathbf{z})}$ related to mutual information maximization. This formulation of f induces the following Eq. (17):

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}} &\geq -\mathbb{E}_{p(\mathbf{z},\mathbf{z}_p)}\left[\log \frac{p(\mathbf{z} \mid \mathbf{z}_p)}{p(\mathbf{z})}\right] + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log \sum_{i=1}^{B-1} \frac{p(\mathbf{z} \mid \mathbf{z}_n^{(i)})}{p(\mathbf{z})}\right] \\ &= -I(Z;Z_p) + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log \sum_{i=1}^{B-1} p(\mathbf{z} \mid \mathbf{z}_n^{(i)})\right] - \mathbb{E}_{p(\mathbf{z})}\left[\log p(\mathbf{z})\right].\end{aligned} \tag{17}$$

Using the property of logarithm and Jensen's Inequality, then

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}} &\geq -I(Z;Z_p) + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log \sum_{i=1}^{B-1} \frac{p(\mathbf{z} \mid \mathbf{z}_n^{(i)})}{B-1}\right] + \log(B-1) - \mathbb{E}_{p(\mathbf{z})}\left[\log p(\mathbf{z})\right] \\ &\geq -I(Z;Z_p) + \frac{1}{B-1}\sum_{i=1}^{B-1} \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log p(\mathbf{z} \mid \mathbf{z}_n^{(i)})\right] + \log(B-1) - \mathbb{E}_{p(\mathbf{z})}\left[\log p(\mathbf{z})\right].\end{aligned} \tag{18}$$

Note that negative sample $\mathbf{z}_n$s are sampled in same distribution. According to Law of large numbers, we can approximate $\frac{1}{B-1}\sum_{i=1}^{B-1} \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log p(\mathbf{z} \mid \mathbf{z}_n^{(i)})\right] \approx \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log p(\mathbf{z} \mid \mathbf{z}_n)\right]$.

Therefore, the last inequality Eq. (19) holds.

$$\mathcal{L}_{\text{InfoNCE}} \geq -I(Z;Z_p) + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)}\left[\log p(\mathbf{z} \mid \mathbf{z}_n)\right] + \log(B-1) - \mathbb{E}_{p(\mathbf{z})}\left[\log p(\mathbf{z})\right] \tag{19}$$

$$= -I(Z;Z_p) + I(Z;Z_n) + \log(B-1). \tag{20}$$

□

A.2 Proof of Proposition 2

Proposition 2. *For the loss function defined in Eq. (7), the following inequality holds:*

$$\mathcal{L}^{\gamma}_{\textit{InfoNCE}} \geq -\frac{1}{\gamma} I(Z; Z_{\textit{pos}}) + I(Z; Z_{\textit{neg}}) + log(B-1). \tag{21}$$

Proof. Compared to Equation Eq. (14), similarity function $f(z, z')$ is replaced with $f_\gamma(z, z') = \exp(\frac{sim(z,z')}{\gamma}) = \{f(z, z')\}^{\frac{1}{\gamma}}$. Therefore, Equation Eq. (16) can be rewritten as:

$$\mathcal{L}^{\gamma}_{\text{InfoNCE}} \geq -\mathbb{E}_{p(\mathbf{z},\mathbf{z}_p)} \left[\log\{f(\mathbf{z}, \mathbf{z}_p)\}^{\frac{1}{\gamma}} \right] + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)} \left[\log \sum_{i=1}^{B-1} f(\mathbf{z}, \mathbf{z}_n^{(i)}) \right] \tag{22}$$

$$= -\frac{1}{\gamma} \mathbb{E}_{p(\mathbf{z},\mathbf{z}_p)} \left[\log f(\mathbf{z}, \mathbf{z}_p) \right] + \mathbb{E}_{p(\mathbf{z},\mathbf{z}_n)} \left[\log \sum_{i=1}^{B-1} f(\mathbf{z}, \mathbf{z}_n^{(i)}) \right]. \tag{23}$$

Following similar derivations with Appendix A.1, we can simply show that $\mathcal{L}^{\gamma}_{\text{InfoNCE}} \geq -\frac{1}{\gamma} I(Z; Z_p) + I(Z; Z_n) + \log(B-1)$. □

B Implementation Details

B.1 Pseudocode

Detailed algorithm for our sampling method is provided in Algorithm 1.

B.2 Evaluation metrics

- **Image Quality and Prompt Alignment:** To measure the quality and textual relevance of the generated images, we employ a suite of widely-recognized automated metrics.
  - **CLIPScore:** This metric evaluates the semantic consistency between a generated image and its corresponding text prompt by calculating the cosine similarity of their embeddings from a pre-trained CLIP model (Hessel et al., 2021).
  - **PickScore:** We use PickScore (Kirstain et al., 2023), a reward model trained on large-scale human preferences, to assess the overall aesthetic quality and prompt alignment of the images.
  - **Image-Reward:** As a complementary metric, Image-Reward (Xu et al., 2023) is another human-preference-based reward model that provides scores reflecting the general quality of the generated content.
- **Diversity:** To evaluate the intra-prompt diversity of the generated images, we utilize two distinct metrics that capture different aspects of variation.
  - **Vendi Score:** The Vendi Score (Friedman & Dieng, 2023) measures the diversity of a set of samples by analyzing the eigenvalue distribution of their similarity matrix. It provides a holistic assessment of both the variety and balance of the generated images.
  - **Mean Pairwise Similarity (MSS):** This metric directly quantifies the average similarity between all unique pairs of images generated for a single prompt. We first extract image features using the self-supervised descriptor for image copy detection (SSCD) model (Pizzi et al., 2022). Then, we compute the pairwise cosine similarity matrix of these features and calculate the mean of its off-diagonal elements. A lower MSS value indicates higher diversity, as images in the set are, on average, less similar to one another.

B.3 Hyperparameter settings

For our main experiments, we use a set of 2K prompts, with each prompt generating a batch of B images. The number of inference steps was set to 50 for Stable Diffusion 1.5 and XL, and 28 for Stable Diffusion 3. The batch size (B) was set to 5 for SD1.5 and SD3, and 3 for SDXL. The specific hyperparameters for our proposed method are detailed in Table 2. As shown in the table, most settings are shared across different T2I backbones, highlighting the robustness of our approach to hyperparameter choices.

Table 2: Model-specific hyperparameters for our proposed method.

Hyperparameter	**Stable Diffusion 1.5**	**Stable Diffusion XL**	**Stable Diffusion 3**
CFG Scale	6.0	6.0	7.0
Optimization Steps (N_{opt})	3	3	3
Gamma (γ)	1.0	1.0	1.0
Window Size (w)	16	16	32
Learning Rate (η)	0.01	0.01	0.001

C ADDITIONAL ANALYSES AND RESULTS

C.1 GAMMA EFFECT: STABILIZING OPTIMIZATION PROCESS

To validate the stability of our proposed method, we conduct an ablation study on the hyperparameter γ to analyze its impact on output variability. We set γ to values of $\{1.0, 0.9, 0.8, 0.7\}$. To ensure that our findings are not contingent on a specific learning rate, we vary the learning rate η within the range of $[0.01, 0.02]$. For each setting of γ, we generate 5 images per prompt, collecting a total of 5K images using SD1.5 model. We then compute evaluation metrics and calculate their sample variance to quantify the statistical variability of the outputs.

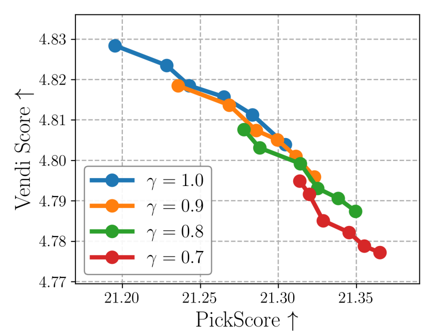


Figure 6: Impact of γ. The Pareto frontier of PickScore vs. Vendi Score.

The results of this experiment are summarized in Table 3, where we calculate the sample variance of those metrics in $\eta \in [0.01, 0.02]$. We observe a clear **saturation effect**: as γ is decreased from 1.0, the variance of the evaluation metrics stabilizes. Specifically, the most significant change in variance occurs when γ is reduced from 1.0 to 0.9. Further decreasing γ to 0.8 and 0.7 yields diminishing changes in variance, indicating that the metrics enter a stable regime. For instance, s^2_{VS} exhibits a steady downward trend as γ decreases, while the other metrics maintain a relatively consistent level of variance for $\gamma \leq 0.9$.

C.2 LEVERAGING KL DIVERGENCE FOR NOISE REGULARIZATION

To further analyze the stability of our method, we investigate the effect of an explicit regularization term. This can be achieved by penalizing the deviation of the optimized noise batch $\{\mathbf{z}_T^i\}_{i=1}^B$ from the standard Gaussian prior, $\mathcal{N}(0, \mathbf{I})$, using a Kullback-Leibler (KL) divergence term (Shlens, 2014).

Table 3: **Gamma Effect.** Subscripts VS, CS, PS, IR mean Vendi Score, CLIPScore, PickScore, Image-Reward, respectively.

γ	Sample variance ($\times 10^{-4}$)			
	s^2_{VS}	s^2_{CS}	s^2_{PS}	s^2_{IR}
1.0	0.076	5.49	1.54	0.60
0.9	0.068	2.47	1.00	0.61
0.8	0.061	3.89	0.78	0.34
0.7	0.050	1.61	0.41	0.11

For a single noise tensor $\mathbf{z}_T$, we treat all of its constituent elements as a single population of data points to estimate an underlying distribution. First, we compute the sample mean ($\hat{\mu}$) and sample variance ($\hat{\sigma}^2$) across all $D = C \times H \times W$ elements within the tensor:

$$\hat{\mu} = \frac{1}{D} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \mathbf{z}_T[c, h, w], \quad \hat{\sigma}^2 = \frac{1}{D-1} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} (\mathbf{z}_T[c, h, w] - \hat{\mu})^2.$$

Here, $\mathbf{z}_T[c, h, w]$ represents the pixel value allocated in c-th channel and (h, w)-position of the latent tensor $\mathbf{z}_T$. These statistics define an estimated univariate Gaussian distribution, $P = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, that characterizes the single noise tensor. We then measure the divergence of this distribution from the standard normal prior, $Q = \mathcal{N}(0, 1)$. The KL divergence for these univariate Gaussian distributions is:

$$D_{KL}(\mathcal{N}(\hat{\mu}, \hat{\sigma}^2) \| \mathcal{N}(0, 1)) = \log \frac{1}{\hat{\sigma}} + \frac{\hat{\sigma}^2 + \hat{\mu}^2}{2} - \frac{1}{2}.$$

By minimizing this KL penalty, we enforce a constraint that encourages the internal statistics of the optimized noise tensor to remain close to those of a standard normal distribution. Integrated algorithm is shown in Algorithm 2.

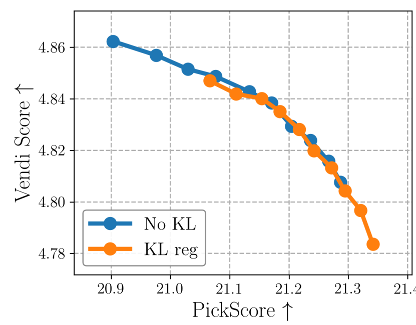


Figure 7: Ablation study on applying Kullback-Leibler divergence. Weight for KL divergence is set as $\lambda = 1000$.

As shown in Figure 7, incorporating this KL penalty shifts the quality-diversity Pareto frontier to the lower-right, indicating a trade-off towards higher textual fidelity at the cost of lower diversity. Interestingly, we observe an analogous phenomenon in our analysis of the attraction coefficient γ. As detailed in Section C.1, lowering the value of γ similarly shifts the frontier to the lower-right and stabilizes performance; see Figure 6 for therein. This parallel suggests that the γ coefficient in our contrastive loss implicitly functions as a regularizer, controlling the diversity-fidelity trade-off in a manner similar to an explicit KL divergence penalty.

C.3 Computational Analysis

To evaluate the practical efficiency and computational overhead of our proposed method, we conducted a comparative analysis against several baseline and state-of-the-art techniques. All experiments in this section were performed using the **Stable Diffusion v1.5** model.

Our evaluation focuses on the trade-off between computational cost and performance. We generated a total of 5K samples for each method to measure the average time per batch, along with key performance indicators for quality (PickScore) and diversity (VendiScore). The results, summarized in Table 4, provide a clear overview of each method's performance profile.

Table 4: Computational cost and performance comparison on Stable Diffusion v1.5.

Method	**Time (sec / batch)** ↓	**PickScore** ↑	**VendiScore** ↑
DDIM	11.131	21.2398	4.8024
Particle Guidance	11.164	21.2164	4.8016
CADS	11.167	21.2254	4.7964
DiversityPrompt	18.703	21.3067	4.7599
Ours (w/o `stopgrad`)	12.853	21.3125	4.8010
Ours (with `stopgrad`)	11.866	21.3044	4.8039

As presented in Table 4, our approach demonstrates a highly compelling efficiency-performance profile. With an optimization step of $N_{opt} = 3$, our method incurs a modest computational overhead of approximately 5% relative to the standard DDIM sampler.

Despite this, our approach is notably faster and achieves superior metric scores compared to MinorityPrompt. It also remains significantly more efficient than computationally intensive methods like Particle Guidance. Crucially, this slight increase in latency is a highly acceptable trade-off. Our method achieves a unique point on the Pareto frontier of efficiency, quality, and diversity. The combination of high PickScore and VendiScore delivered by our approach represents a state-of-the-art balance unmatched by any other method at any computational cost. This result underscores the practical value of our method, offering a solution that is both powerful and efficient for real-world applications.

C.4 Additional generated samples

Figure 8 represents that our method shows high image quality and textual fidelity compared to DDIM and CADS. We use SDXL in common, and hyperparameters of our method are equivalent to Table 2.

D Use of Large Language Models

We used Large Language Models (LLMs) to aid in the verbal refinement and polishing of our paper. This usage was only limited to improving readability and fixing some grammar errors. The core research, including the formulation of our method, experimental design, and analysis of results, was conducted solely by the authors.

Algorithm 1 Diverse T2I Generation via Contrastive Noise Optimization

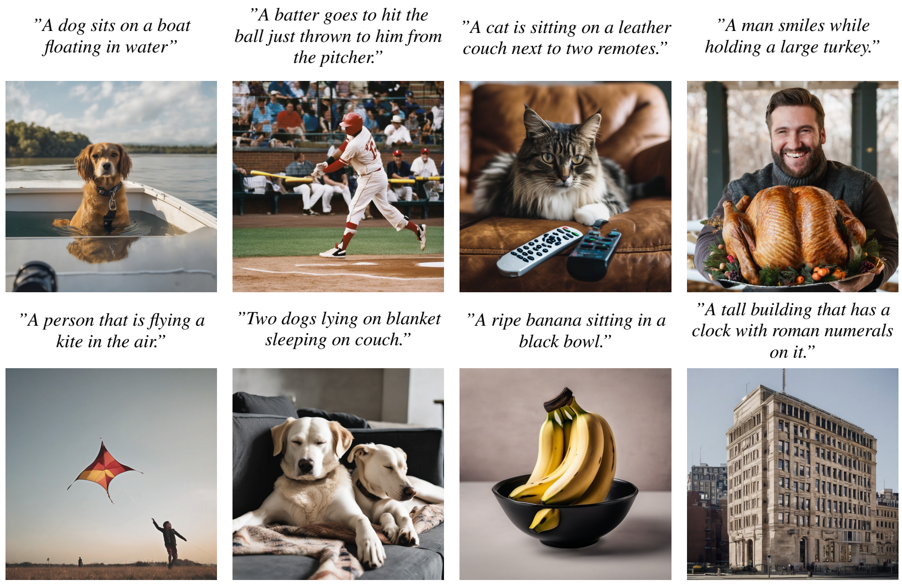
Inputs: Text embedding c, batch size B, optimization steps N_{opt}, learning rate η, temperature τ, attraction coefficient γ, window size w
Outputs: $\{\mathbf{x}_0^b\}_{b=1}^B$, a batch of diverse images.

Models: Diffusion model ϵ_θ, DDIM sampler $\mathcal{D}_{\text{DDIM}}$
Initialize a batch of trainable noise vectors $\mathbf{Z}_T = \{\mathbf{z}_T^i\}_{i=1}^B \sim \mathcal{N}(0, \mathbf{I})$
Let $\mathbf{Z}_T^{\text{fixed}} \leftarrow \mathbf{Z}_T$
for $n = 1$ **to** N_{opt} **do**
  for $i = 1$ **to** B **do**
    $\mathbf{z}_{0|T}^{i,\text{opt}} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_T}}(\mathbf{z}_T^{i,\text{opt}} - \sqrt{1-\bar{\alpha}_T}\texttt{stopgrad}\{\epsilon_\theta(\mathbf{z}_T^{i,\text{opt}}, T, c)\})$
    if $n = 1$ **then**
      $\mathbf{z}_{0|T}^{i,\text{fixed}} \leftarrow \mathbf{z}_{0|T}^{i,\text{opt}}$
    end if
    $\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}} \leftarrow \text{DownSample}(\mathbf{z}_{0|T}^{i,\text{opt}}; w),\ \ \text{DownSample}(\mathbf{z}_{0|T}^{i,\text{fixed}}; w)$
    $\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}} \leftarrow \text{Normalize}(\mathbf{z}_{0|T}^{i,\text{opt}}),\ \ \text{Normalize}(\mathbf{z}_{0|T}^{i,\text{fixed}})$
  end for
  $\mathcal{L}_{\text{CNO}}^{\gamma} := \mathbb{E}_{i\sim\mathcal{B}}\left[-\log \frac{\exp(\text{sim}(\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}})/\gamma\tau)}{\sum_{j\in\mathcal{B}} \exp(\text{sim}(\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{j,\text{opt}})/\tau)}\right]$
  $\mathbf{Z}_T \leftarrow \mathbf{Z}_T - \eta \cdot \nabla_{\mathbf{Z}_T}\mathcal{L}_{\text{InfoNCE}}^{\gamma}$
end for
$\{\mathbf{z}_0^b\}_{b=1}^B \leftarrow \mathcal{D}_{\text{DDIM}}(\mathbf{Z}_T, c)$
$\{\mathbf{x}_0^b\}_{b=1}^B \leftarrow \text{Decode}(\{\mathbf{z}_0^b\}_{b=1}^B)$
return $\{\mathbf{x}_0^b\}_{b=1}^B$

Algorithm 2 Contrastive Noise Optimization with KL Regularization

Inputs: Text embedding c, batch size B, optimization steps N_{opt}, learning rate η, temperature τ, attraction coefficient γ, window size w, KL divergence weight λ
Outputs: $\{\mathbf{x}_0^b\}_{b=1}^B$, a batch of diverse images.

Models: Diffusion model ϵ_θ, DDIM sampler $\mathcal{D}_{\text{DDIM}}$
Let D be the number of pixels in a single noise tensor (channel $\times$ height $\times$ width)
Initialize a batch of trainable noise vectors $\mathbf{Z}_T = \{\mathbf{z}_T^i\}_{i=1}^B \sim \mathcal{N}(0, \mathbf{I})$
Let $\mathbf{Z}_T^{\text{fixed}} \leftarrow \mathbf{Z}_T$
for $n = 1$ **to** N_{opt} **do**
  for $i = 1$ **to** B **do**
    $\mathbf{z}_{0|T}^{i,\text{opt}} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_T}}(\mathbf{z}_T^{i,\text{opt}} - \sqrt{1-\bar{\alpha}_T}\texttt{stopgrad}\{\epsilon_\theta(\mathbf{z}_T^{i,\text{opt}}, T, c)\})$
    if $n = 1$ **then**
      $\mathbf{z}_{0|T}^{i,\text{fixed}} \leftarrow \mathbf{z}_{0|T}^{i,\text{opt}}$
    end if
    $\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}} \leftarrow \text{DownSample}(\mathbf{z}_{0|T}^{i,\text{opt}}; w),\ \ \text{DownSample}(\mathbf{z}_{0|T,}^{i,\text{fixed}}; w)$
    $\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}} \leftarrow \text{Normalize}(\mathbf{z}_{0|T}^{i,\text{opt}}),\ \ \text{Normalize}(\mathbf{z}_{0|T}^{i,\text{fixed}})$
  end for
  $\mathcal{L}_{\text{CNO}}^{\gamma} := \mathbb{E}_{i\sim\mathcal{B}}\left[-\log \frac{\exp(\text{sim}(\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{i,\text{fixed}})/\gamma\tau)}{\sum_{j\in\mathcal{B}} \exp(\text{sim}(\mathbf{z}_{0|T}^{i,\text{opt}}, \mathbf{z}_{0|T}^{j,\text{opt}})/\tau)}\right]$
  $\hat{\mu}_i \leftarrow \frac{1}{D}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}\mathbf{z}_T[c,h,w] \quad$ for $i = 1, \dots, B$
  $\hat{\sigma}_i^2 \leftarrow \frac{1}{D-1}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}(\mathbf{z}_T[c,h,w] - \hat{\mu})^2 \quad$ for $i = 1, \dots, B$
  $\mathcal{L}_{\text{KL}} := \frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{1}{\hat{\sigma}_i} + \frac{\hat{\sigma}_i^2 + \hat{\mu}_i^2}{2} - \frac{1}{2}\right]$
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CNO}}^{\gamma} + \lambda\mathcal{L}_{\text{KL}}$
  $\mathbf{Z}_T \leftarrow \mathbf{Z}_T - \eta \cdot \nabla_{\mathbf{Z}_T}\mathcal{L}_{\text{total}}$
end for
$\{\mathbf{z}_0^b\}_{b=1}^B \leftarrow \mathcal{D}_{\text{DDIM}}(\mathbf{Z}_T, c)$
$\{\mathbf{x}_0^b\}_{b=1}^B \leftarrow \text{Decode}(\{\mathbf{z}_0^b\}_{b=1}^B)$
return $\{\mathbf{x}_0^b\}_{b=1}^B$



(a) DDIM

(b) CADS

(c) Ours

Figure 8: Qualitative comparison of images generated from the same set of text prompts by (a) DDIM, (b) CADS, and (c) our proposed method. Images with the same position in individual grids share the same prompt and seed.