Don't Settle at the Mode! Mitigating Diversity Collapse in Pretrained Flow Models via Feature Self-Guidance

Pradhaan S Bhat[1⋆], Rishubh Parihar[1⋆], Abhijnya Bhat[1,2], and R. Venkatesh Babu[1]

[1] Vision and AI Lab, Indian Institute of Science, Bangalore, India
[2] Stanford University, USA

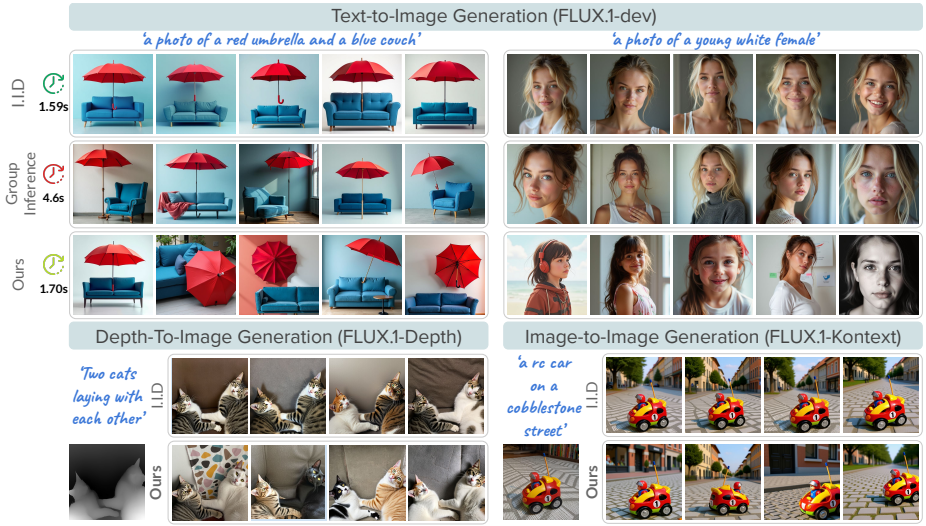


Fig. 1: We propose an efficient inference-time approach to enhance diversity in conditional flow models. Our plug-and-play module integrates seamlessly into various conditional flow models and consistently enhances diversity for text-to-image, depth-to-image and image-to-image models. Compared to recent Group Inference [39], which relies on reward-based sample rejection, our method achieves competitive diversity without external reward models while requiring only a fraction of their inference time.

Abstract. State-of-the-art flow models generate stunning images from text or image prompts. However, they suffer from diversity collapse when generating multiple samples under the same conditioning. Existing methods address this issue via either latent guidance, which has limited effectiveness, or sample selection, which relies on external reward models that incur significant inference-time overhead. In this work, we introduce an efficient, training-free self-guidance mechanism to mitigate diversity collapse without requiring additional reward models. Specifically, we disperse the internal features of the flow model during batch generation with **feature self-guidance**. Further, to keep the features close

⋆ Equal Contribution

to the manifold, we introduce a **manifold regularization** step that projects these dispersed features back onto the data manifold, ensuring diverse generation without sacrificing alignment with the input conditions. Our method integrates seamlessly as a plug-and-play module into pretrained flow models, adding only a marginal inference cost. Experiments demonstrate significant improvements in diversity while preserving fidelity across several conditional flow models, including multi-step and few-step text-to-image, depth-to-image, and reference image generation.

1 Introduction

While state-of-the-art flow-based generative models [9, 30, 31, 53] achieve exceptional generation quality, they often exhibit a lack of sample diversity when prompted with the same conditioning [17, 27, 39]. As illustrated in Fig. 1, widely used FLUX.1-dev [30] produces nearly redundant outputs for the prompt *'a portrait photo of a woman'*, capturing only a narrow subset of the potential distribution. Diverse generation is critical in practical applications, where a wide array of outputs is necessary to facilitate user choice and provide inspiration for iterative prompt engineering.

Existing methods to mitigate diversity collapse in pretrained flow models during inference fall into two categories. The first focuses on latent guidance [5, 25, 27, 29], which guides latents to generate varied samples. While computationally efficient, these approaches often yield marginal improvements in sample variety. The second category leverages external reward models to either select the most diverse samples from a candidate set [39] or optimize the initial noise for enhanced variety [17]. While these reward-driven strategies significantly enhance diversity, they incur substantial computational overhead due to the repeated inference of reward models (Fig. 1) over decoded latents in the pixel space. In this work, we introduce an efficient and effective method for recovering diversity in pretrained flow models without external models.

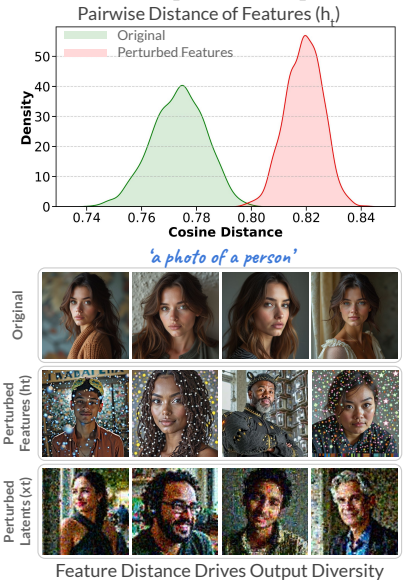


Fig. 2: Feature Distance vs. Output Diversity. Perturbing FLUX's internal DiT features h_t with Gaussian noise increases pairwise feature distances and sample diversity, while perturbing latents x_t corrupts images, indicating that feature dispersion drives flow model diversity.

Our key insight is that *diversity collapse can be mitigated by addressing the collapse of internal features of flow models*. To validate this hypothesis, we conduct a diagnostic experiment that intentionally expands the internal representation space of the FLUX.1-dev model.

Specifically, we introduce stochastic perturbations to the intermediate Multimodal Diffusion Transformer [9] (MMDiT) block features h_t by injecting Gaussian noise, defined as $\hat{h}_t = h_t + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. This setup allows us to observe how feature space shifts affect sample diversity. As illustrated in the pairwise distance histograms and results in Fig. 2, increasing feature-space variance significantly increases generated sample diversity. This reveals a clear correspondence: *the spatial dispersion of internal MMDiT features serves as a direct proxy for the diversity of final image outputs.* Interestingly, the simple alternate of latent perturbations (x_t) degrades image quality, causing loss of high-frequency and texture details and corrupting the image, whereas perturbing internal features largely preserves image structure and details, producing only mild overlaying artifacts. This highlights the robustness of the feature space as a target for diversity enhancement. However, unconstrained perturbation often drives features into low-density regions, manifesting as artifacts. To address this, we propose a principled feature dispersion that increases diversity while anchoring features to the manifold, preserving image fidelity and faithfulness.

To maximize the variance of internal MMDiT features without compromising semantic integrity, we introduce an efficient **feature self-guidance** mechanism. For a given batch of latents, we extract intermediate features h_t and apply a dispersion guidance that pushes the features apart. However, unconditional dispersion risks pushing features beyond the semantic boundary, risking the alignment with the input conditioning. To mitigate this, we propose a **manifold regularization** where the dispersed features are reprocessed through the same MMDiT block to project them back toward the valid conditional feature manifold. The projected features are then used as a regularizer for the raw dispersed features. Through rigorous analysis, we demonstrate that applying this guidance to a single MMDiT block over a selected timestep window is sufficient to maximize diversity while maintaining a minimal computational footprint.

Our training-free approach is designed for seamless integration into any pretrained flow model as a plug-and-play module. To demonstrate its broad applicability, we conduct extensive evaluations across a diverse set of conditional flow models: FLUX.1 for text-to-image generation, FLUX.1-Depth for depth-conditioned synthesis, FLUX-Kontext for reference image generation, and the step-distilled FLUX.2-Klein. Across all benchmarks, our method significantly outperforms existing diversity-enhancement baselines while preserving the native high-speed inference characteristic of the underlying flow models.

In summary, our key contributions are:

- The key discovery of a direct correlation between internal feature collapse in pretrained flow models and output diversity collapse.
- Feature Self-Guidance that effectively disperse of internal features of the flow model to enhance diversity while utilizing manifold regularization to anchor samples to the conditioning signal and preserve image fidelity.
- Extensive empirical validation across diverse tasks—including state-of-the-art text-to-image and conditional generation, demonstrating that our method consistently recovers diversity without sacrificing image quality.

2 Related Works

Guidance and Diversity: Standard Classifier-Free Guidance [20] can be used to control the diversity in pretrained diffusion models [19,45,46]. However, it has a poor faithfulness and diversity tradeoff, where highly diverse samples don't follow the input conditioning. Several guidance-based inference strategies have emerged to steer the model towards a better fidelity-diversity trade-off [4, 22]. Balancing-Act [38] guides the internal h-space features [28] of UNet based diffusion models to improve attribute diversity. Interval Guidance [29] balances these objectives by restricting the guidance to specific inference windows. Other approaches focus on the sampling trajectory itself: Particle Guidance [5] learns a joint-particle time-evolving potential to maximize diversity, while Shielded Diffusion [27] promotes diversity by sparsely repelling noisy latents at some inference steps. Alternatively, CNO [25] bypasses per-step guidance entirely by optimizing the initial latent space. Unlike methods operating in the latent space, our approach operates directly in the internal feature space of MMDiT [9] blocks. By shifting the intervention to the internal feature space, we provide an effective solution by mitigating collapse at its source.

Prompt Optimization for Diversity: Improving generative control by optimizing the input prompt embedding without training the flow/diffusion model is a well-established direction [1, 14, 15]. In similar lines, CADS [42] enhances diversity by perturbing the conditioning signal; however, this often compromises prompt adherence and leads to semantic drift. Other methods directly optimize [35, 49, 57, 58] for diversity or employ multimodal large language models to iteratively refine the prompt based on visual feedback [24, 61]. Because these methods necessitate test-time optimization, their practical application is constrained; moreover, optimization within the text space lacks the expressivity required to fully capture the complexity of diverse data distributions.

Representation Regularization for Generation: Recent works have leveraged self-supervised learning to introduce intermediate representation regularization during the training of Diffusion Transformers [43, 48, 51, 56, 59]. These methods, which primarily aim to enhance training efficiency, can be broadly categorized into two strategies: (1) aligning internal representations of the flow models with the pre-trained self-supervised models [43, 48, 56, 59], and (2) incorporating self-supervised learning as an auxiliary regularization task during flow model training [51, 60]. Our method takes inspiration from the latter set of methods; however, we modify the internal representations during inference via feature guidance without the need for training.

Inference-time Scaling. Inference-time scaling [44, 52] in diffusion and flow models improves sample quality by allocating more computational resources during inference, like having to reward the model or performing optimization. There are two primary directions in inference-time scaling: iterative optimization and discrete selection. Optimization-based methods refine the initial Gaussian noise for point-wise quality via quality rewards [10, 11, 47] along with set-level objec-

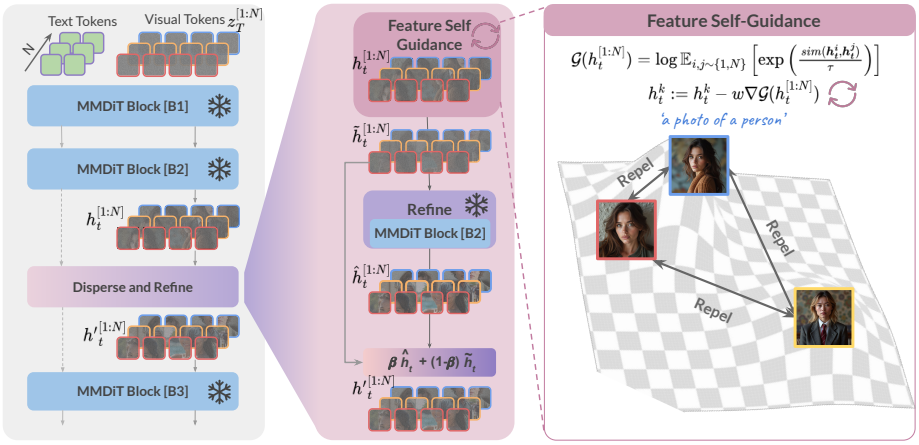


Fig. 3: Overview of Feature Self-Guidance. Our method integrates a "Disperse-and-Refine" module into the flow model backbone. **(a)** Dispersion Stage: Internal features are pushed apart using iterative self-guidance to expand sample variety. **(b)** Refinement Stage: These features are re-processed through the same MMDiT block to project them back onto the valid image manifold. Finally, we linearly interpolate the dispersed and projected features to ensure a high-fidelity and, diverse output.

tives to explicitly manage diversity [17] or integrate Determinantal Point Processes (DPP) into policy optimization loops [23]. On the other hand, selection-based approaches leverage reward models to identify optimal trajectories through verifier feedback. These methods scale performance either by filtering a large pool of sample candidates to identify 'Best-of-N' samples [36] or by sampling from a reward-tilted distribution [26]. This setting extends to set-wise diverse selection strategies via the use of Quadratic Integer Programming at intermediate denoising steps to balance quality and diversity within a generated group of samples [39]. However, a significant bottleneck shared by both types of approaches is their reliance on reward models, significantly improving the number of function evaluations during inference. In contrast, we demonstrate that high-quality, diverse samples can be produced efficiently by *dispersing* flow features on a constrained manifold, bypassing the need for additional reward supervision.

3 Method

3.1 Preliminaries

Flow Models. Flow models [33, 34] learn a continuous transformation from a Gaussian prior $p_1 \sim \mathcal{N}(0, I)$ to a target data distribution p_0 by modeling an Ordinary Differential Equation (ODE) trajectory. For training, these models define an interpolant x_t between a data point $x_0 \sim p_0(x)$ and a noise sample $x_T \sim p_1$ as $x_t = (1-t)x_0 + tx_T$ for $t \in \mathcal{U}[0, 1]$. A neural network v_θ is trained to predict the conditional velocity field $v_\theta(x_t, t, c)$ to match the true velocity

$(x_T - x_0)$. The training objective, known as the Conditional Flow Matching (CFM) loss, is defined as:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],x_0\sim p_0,x_T\sim\mathcal{N}(0,\mathbf{I})}\left[\|v_\theta(x_t,t,c) - (x_T - x_0)\|^2\right] \quad (1)$$

where c represents the conditioning signal such as text prompts.

Guidance. Analogous to classifier guidance in diffusion models [6], flow models can be adapted for inference-time conditioning by steering the predicted score $\nabla \log p_t(x_t)$ via a pretrained classifier $p(c|x_t)$ as follows:

$$\nabla \log p_t(x_t|c) = \nabla \log p_t(x_t) + \lambda \nabla_{x_t} p(c|x_t) \quad (2)$$

where c is the class label. In principle, this guidance can be applied using any differentiable energy function $\mathcal{G}(x_t, c, t)$, such as CLIP [40] for enhancing prompt-adherence, an object detector for layout-conditioned generation [3], or sketches for sketch-to-image synthesis [50]. Alternatively, the energy function can be computed directly from the model's internal features via self-guidance [8]. Because self-guidance bypasses external model evaluations, it only adds a marginal computational cost to the inference process.

3.2 Overview

We build on FLUX [30], a DiT architecture for text-to-image (T2I) generation which comprises of a series of multimodal DiT (MMDiT) blocks $B_1 \cdots B_{57}$ [9]. These blocks jointly process text and image tokens through self-attention and feed-forward layers facilitating rich information exchange between text and image tokens. Leveraging the rich information within the MMDiT blocks, we introduce a strategy to **disperse** the image features and mitigate mode collapse. To ensure these **dispersed** features remain on the data manifold, we incorporate a **refinement** step that regularizes the features, preventing them from drifting into low-probability regions. This dual **disperse-and-refine** mechanism enhances generative diversity without sacrificing fidelity. Detailed ablations and design choices follow in the subsequent sections. We present the overall algorithm in Algorithm 1.

3.3 Feature Self-Guidance for diversity

We build our method on the insight from Fig. 2, i.e., the diversity collapse in the generated samples is due to the collapse of the internal MMDiT features h_t. We design a principled method that effectively disperses the internal MMDiT features while preserving the image fidelity. We formulate a feature self-guidance method that iteratively disperses the batch sample features $h_t^{[1:N]}$ through guidance. We use pair-wise feature distance to formulate our self-guidance energy function $\mathcal{G}$ defined as:

$$\mathcal{G}(h_t^{[1:N]}) = \log \mathbb{E}_{i,j\sim\{1,N\}}\left[\exp\left(\frac{sim(\boldsymbol{h_t^i}, \boldsymbol{h_t^j})}{\tau}\right)\right] \quad (3)$$

where N is the number of samples in a batch and $sim(\mathbf{h}_t^i, \mathbf{h}_t^j)$ is the cosine similarity between the features. However, unlike existing guidance-based approaches that update the latents x_t with gradients of energy function $\nabla\mathcal{G}(h_t)$, we directly update the internal features h_t as below, enforcing feature diversity more explicitly by pushing the features apart:

$$h_t^k := h_t^k - w\nabla\mathcal{G}(h_t^{[1:N]}) \tag{4}$$

Notably, updating h_t is much faster than updating x_t with guidance, as it only modifies intermediate features without requiring backpropagation through previous transformer blocks. This efficient self-guidance enables diverse generations; however, it may reduce faithfulness to onditioning, as iterative dispersion may push features outside the feature manifold. To address this, we propose a manifold regularization step that keeps updated features in distribution (Sec. 3.4).

Which block is most effective for feature self-guidance?

Inspired by StableFlow [2], which shows that early MMDiT blocks capture most semantic information, we evaluate feature guidance across the initial blocks (Fig. 5a). Block B_2 performs best, while adding it to later blocks yields only marginal diversity gains. Furthermore, since image structure is primarily formed in the early denoising steps, we apply guidance only for $t \in [1.0, 0.8]$. Additional ablations on block and timestep selection are provided in the appendix.

3.4 Manifold regularization

As with our dispersive feature update, the features $\tilde{h}_t$ can potentially be pushed out-of-the feature manifold. We propose a manifold regularization to bring the dispersed features back to the feature manifold. We build on the findings from seminal work in mechanistic interpretability [7] that suggest the sequence of transformer blocks with residual connections behaves in a *read-write mechanism* where they read the features from the main stream (h_t) and add back a connection Δh_t to the stream (see Fig 4). We implement the regularizer by passing the updated features $\tilde{h}_t$ again through the same MMDiT block to project them into the output feature space, giving projected features $\hat{h}_t$. In our context, we pass the dispersed features again through the same block to make corrections to h_t with $\Delta\tilde{h}_t$. This allows the block, utilizing its pretrained internal priors, to project the perturbed features back onto the natural image manifold while preserving diversity. Next, we use the projected features and the dispersed features to obtain a final update:

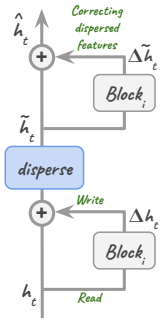
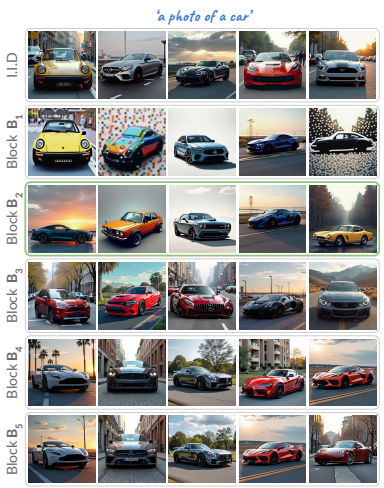


Fig. 4: Read-write mechanism of transformers: Our regularization method leverages the read-write mechanisms of transformer blocks to add corrections to dispersed features to project them to the natural image manifold.

$$h_t' := \beta\hat{h}_t + (1-\beta)\tilde{h}_t \tag{5}$$



(a) Block Ablation for feature guidance: Guidance at Block B_2 features is most effective, achieving high diversity while preserving fidelity. Guidance at subsequent blocks is not effective.

(b) Faithfulness-Diversity tradeoff: Lower β ensures higher diversity, but sacrifices faithfulness, while higher β regularizes features and improves prompt adherence with a marginal diversity loss. We use $\beta = 0.4$, which offers a good tradeoff between prompt adherence and diversity.

Fig. 5: Ablation over DiT Block for dispersion & mixing parameter β choices.

where, $\beta \in [0, 0.6]$. We pass the updated h'_t to the next MMDiT blocks for processing. Further, β acts as a tunable knob to steer the generation to achieve a desired faithfulness-diversity tradeoff as shown in Fig. 5b. The proposed method is a plug-and-play method that modifies outputs of only one MMDiT block and improves diversity while preserving the alignment with input conditioning.

Algorithm 1 Diverse Generation via Feature Self-Guidance

Inputs: Text embedding c, batch size N, optimization steps N_{opt}, learning rate w, temperature τ, start time T_a, end time T_b, blending rate β, Blocks $\mathbf{B}_1$ to $\mathbf{B}_{57}$
Outputs: $\{x_0^i\}_{i=1}^N$, a batch of diverse images.

```
def feature_self_guidance(c, N, N_opt, w, τ, T_a, T_b, β):
    z = torch.randn(N, ., ., .) # Initial Noise
    for t in reversed(range(T)):
        if T_a <= t <= T_b:
            h_t = B_2 ∘ B_1(z_t, t, c)
            for k in range(N_opt): → Self-guidance
                G = Eq. (3) (h_t, τ)
                h_t = h_t − w∇G → Update step
            ĥ_t = B_2(h_t, t, c) → Refine
            h'_t = β ĥ_t + (1 − β) h_t
            v_t = B_57 ∘ ··· B_3(h'_t, t, c)
        else:
            v_t = B_57 ∘ ··· B_1(z_t, t, c)
        z_t = z_t + Δt * v_t
    return Decode(z_0)
```

Fig. 6: Qualitative results for text-to-image generation: We demonstrate the advantages of our inference-time method over standard I.I.D. sampling, which often suffers from visual homogeneity in layout and color. In contrast, our approach generates significant diversity in viewpoints, compositions, and color palettes while maintaining strict prompt fidelity. This is seen in the ability of our method to adhere to prompts with numerical and positional constraints without sacrificing diversity.

4 Experiments

We evaluate our method across several conditional image generation flow models - text-to-image generation (FLUX.1-dev), depth-to-image generation (FLUX.1-Depth-dev), and foundational instruction-driven image editing model (FLUX.1-Kontext-dev) for personalization. Additionally, we also evaluate the step-distilled FLUX.2-Klein-4B model, UNet-based DDPM [19] model and other T2I models [53, 54] in the appendix. In the following sections, we discuss the baselines, implementation details, results, and ablations.

4.1 Experimental Setup

Baselines: We compare our method with several inference approaches designed for diverse text-to-image generation. **Particle Guidance** and **Interval Guidance** leverage diffusion guidance to steer samples toward greater diversity. **Shielded Diffusion** implements a repellency loss that pushes the samples away from a set of shielded samples. **Group Inference** leverages pretrained CLIP [40] and DINO [37] reward models to select a subset of the most diverse and faithful samples from a large set of candidates during the denoising process. We report results for 8 samples using 128 and 64 candidates for Group Inference. Furthermore, we

only compare with Group Inference on depth-to-image and personalized image generation tasks, as it can be easily adapted.

Dataset: We use task-specific evaluation datasets as discussed below:

- **Text-to-Image Generation**: We use the standard **GenEval** dataset [16], an object-focused benchmark to evaluate compositional properties of text-to-image models such as object co-occurrence, position, count, and color.
- **Depth-Conditioned Generation**: We use a 500 image subset of the **COCO** 2017 validation split [32] and extract their depth maps with DepthAnythingv2 [55] model for depth-to-image generation.
- **Personalized Image Generation**: We assess personalization results on the **Dreambooth** dataset [41] to measure our methods' ability to generate diverse samples for a given subject with the same prompt.

Evaluation Metrics: We evaluate our method across two major axes: **i)** *sample diversity* and **ii)** *prompt adherence*. For sample diversity, we use reference-free diversity metrics: the pairwise cosine distance between DINOv2 [37] features, image dissimilarity using DreamSim [13], mean similarity using MSS [42], VendiScore [12], which measures the effective number of unique elements in a set and for prompt adherence, we use CLIPScore [18, 40]. We report the mean and standard deviation of all the metrics across the entire dataset. We include additional metrics in the appendix.

Implementation Details: To ensure reproducibility, all methods are evaluated using a fixed seed of 42. We sample 8 initial noise latents from an isotropic Gaussian distribution, $\mathcal{N}(0, I)$, and generate images at a resolution of 512^2. For the base models, we adhere to default configurations - specifically employing a guidance scale of 3.5 and 28 inference steps. Our feature self-guidance is applied at the second MMDiT block during the initial denoising phase ($t \in [1, 0.8]$). This process involves 30 optimization steps with a learning rate of 0.5 and temperature of 0.5. Finally, the dispersed and refined latents are integrated using a blending scale of 0.4. We use the default hyperparameters provided by baselines wherever available; a full list of hyperparameters is available in the appendix.

4.2 Text-to-Image Generation

Qualitative Comparisons: We present qualitative results of our method in Fig.6. The samples generated by our method are significantly more diverse in terms of style, layout, and view-points as compared to I.I.D. baseline. We present a comparison with several baselines for text-to-image generations in Fig.7. Our method synthesizes diverse images and backgrounds from minimal textual input (e.g.*'a photo of a bear'*). Our approach excels in complex compositional examples, such as *'a photo of a white dog and a blue potted plant'* where it maintains distinct object identities with accurate attribute binding as it works on semantic latent features rather than spatial latents. In contrast, guidance-based methods such as Particle Guidance [5] and Interval Guidance [29] generate images that are visually similar to the I.I.D. baseline. Sparse intervention methods

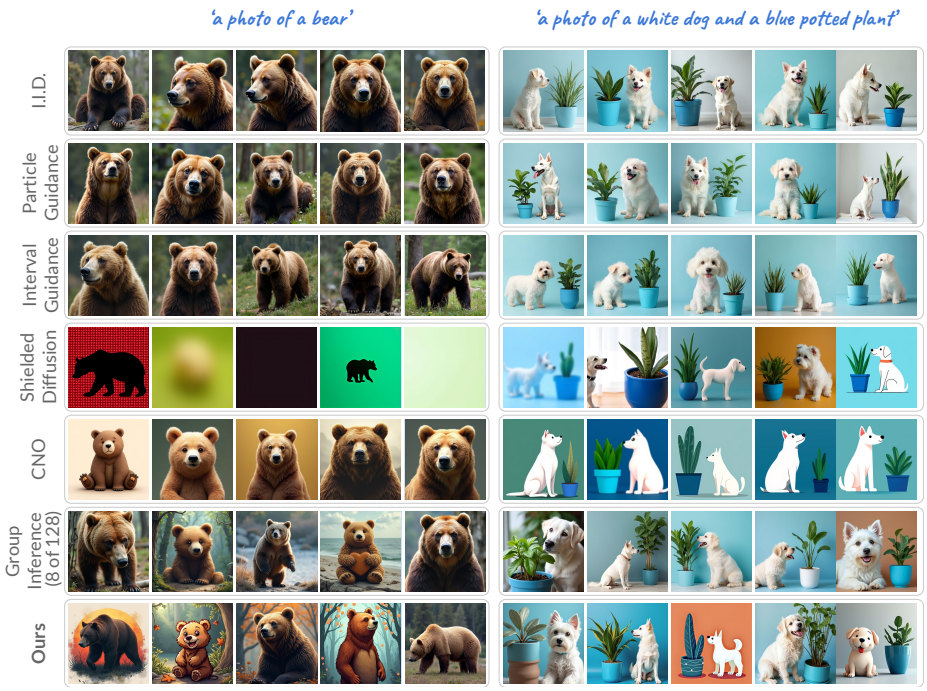


Fig. 7: Qualitative Comparison for text-to-image generation: We compare against works on diverse sampling in T2I Generation: *Particle Guidance* [5], *Interval Guidance* [29], *Shielded Diffusion* [27], *CNO* [25] and *Group Inference* [39]. Our method generates highly diverse samples in terms of pose, style, and appearance while adhering to the text prompt. The baseline methods either have limited diversity (Particle Guidance, Interval Guidance, CNO) or collapse the generation (Shielded Diffusion). Group-Inference achieves good diversity; however, it requires sampling from a large number of candidates, making it much slower than our method.

like Shielded Diffusion [27] apply sparse repulsion to approximate clean latents, thereby pushing samples off the manifold and leading to visual artifacts (bear example). CNO [25] utilizes initial noise optimization to explore stylistic variations, but remains strictly tethered to the diffusion prior, limiting its output to superficial stylistic diversity rather than meaningful structural variety. Group Inference [39] offers the most diversity among the baselines in terms of poses and backgrounds. However, it selects 8 images from a candidate pool of 64/128 images, making it inefficient for practical scenarios.

Quantitative Comparisons: We present our quantitative results in Tab.1. Analysis of the metrics reveals several limitations in existing state-of-the-art methods. Both Particle Guidance [5] and Interval Guidance [29] fail to surpass the I.I.D. baseline in generative diversity; furthermore, Particle Guidance is an inefficient paradigm due to its prohibitive memory requirements as seen in Fig.10 **b)**. Shielded Diffusion [27] although time efficient, is limited in the sample di-

Table 1: Quantitative results of text-to-image diverse samplers with FLUX.1-dev.

Method	Latency↓	DINO↑	VS↑	DreamSim↑	MSS↓	CLIPScore↑
IID	**1.59s**	$0.57_{\pm 0.10}$	$2.77_{\pm 1.06}$	$0.27_{\pm 0.10}$	$0.32_{\pm 0.10}$	$32.43_{\pm 3.71}$
Particle Guidance [5]	2.94s	$0.58_{\pm 0.11}$	$3.16_{\pm 1.17}$	$0.31_{\pm 0.12}$	$0.29_{\pm 0.11}$	$32.22_{\pm 3.90}$
Interval Guidance [29]	1.59s	$0.57_{\pm 0.10}$	$2.88_{\pm 1.11}$	$0.28_{\pm 0.10}$	$0.300_{\pm 0.10}$	$32.46_{\pm 3.75}$
Shielded Diffusion [27]	1.59s	$0.61_{\pm 0.10}$	$3.35_{\pm 1.22}$	$0.40_{\pm 0.13}$	$0.23_{\pm 0.08}$	$32.11_{\pm 3.89}$
CNO [25]	1.75s	$0.63_{\pm 0.08}$	$3.46_{\pm 1.10}$	$0.39_{\pm 0.10}$	$0.22_{\pm 0.07}$	$31.97_{\pm 3.77}$
GroupInference [39] (8 of 64)	2.86s	$0.67_{\pm 0.07}$	$3.37_{\pm 1.17}$	$0.35_{\pm 0.10}$	$0.21_{\pm 0.07}$	$\mathbf{32.49_{\pm 3.74}}$
GroupInference [39] (8 of 128)	4.55s	$\mathbf{0.70_{\pm 0.07}}$	$\mathbf{3.61_{\pm 1.26}}$	$0.37_{\pm 0.11}$	$0.19_{\pm 0.07}$	$32.37_{\pm 3.77}$
Ours	1.70s	$0.68_{\pm 0.07}$	$3.57_{\pm 1.17}$	$\mathbf{0.40_{\pm 0.09}}$	$\mathbf{0.19_{\pm 0.06}}$	$32.05_{\pm 3.64}$

versity and has inferior overall prompt adherence. Similarly, CNO [25] is limited in its diversity to stylistic variations and is unable to capture all the modes of the distribution. Finally, while Group Inference [39] obtains high quality and diversity, it is extremely slow due to the selection of a subset from a large pool, as reported in latency. Its reliance on additional reward models to score the large pool of samples increases inference time, making it less practical for real-time applications. In contrast, our method offers significantly enhanced diversity while preserving prompt adherence and comparable latency to I.I.D. generation, making it highly practical for real-world scenarios. These results underscore our method's ability to unlock structural and stylistic diversity without compromising the underlying diffusion prior's stability or usability. We compare with additional baselines and also evaluate on DPG-Bench [21] in the appendix.

4.3 Additional Image Generation Tasks

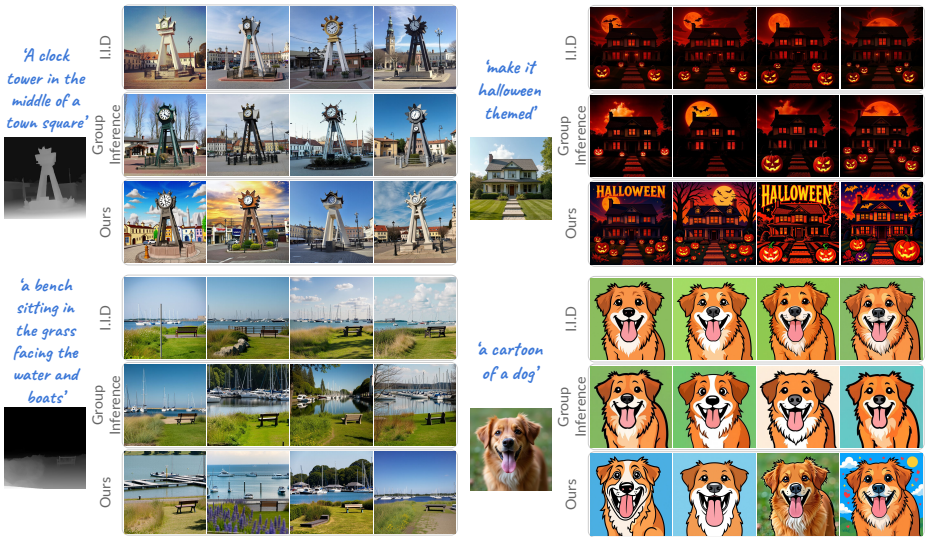
We show the generalization of our method by integrating it with other conditional flow models. We demonstrate results on depth-conditioned generation and reference-conditioned generation, where the same subject needs to be generated in diverse contexts in Fig.8 and Tab.2. Our method consistently produces diverse compositions and styles, and faithfully adheres to strong conditioning like depth maps (Fig.8a) and reference images (Fig.8b). In contrast, Group Inference has inferior diversity and has high inference latency. These results are further quantified in Tab.2, which demonstrates our method's superior diversity with marginal impact on overall prompt adherence.

Table 2: Quantitative results of conditional image generation: Our method significantly improves diversity over I.I.D. while preserving the prompt alignment.

Task	**Depth to Image**		**Personalized Image**	
Method↓	DINO	CLIPScore	DINO	CLIPScore
I.I.D.	$0.43_{\pm 0.08}$	$\mathbf{30.85_{\pm 3.12}}$	$0.45_{\pm 0.10}$	$31.93_{\pm 4.31}$
Group Inference	$0.47_{\pm 0.09}$	$30.63_{\pm 3.12}$	$0.49_{\pm 0.15}$	$\mathbf{32.58_{\pm 3.54}}$
Ours	$\mathbf{0.52_{\pm 0.09}}$	$30.49_{\pm 3.23}$	$\mathbf{0.57_{\pm 0.09}}$	$32.04_{\pm 4.11}$

4.4 Ablations

We conduct ablation studies to evaluate the key design choices, including the regularization parameter β and inference batch size. Further, we provide the ablation over the selection of MMDiT block and the specific denoising timestep interval for feature self-guidance in the appendix.



(a) Depth-to-Image Generation **(b)** Reference-Image Generation

Fig. 8: Additional Generation Tasks: Our method demonstrates the ability to generate diverse samples while maintaining adherence to the input conditioning, such as **depth maps** or **reference images**. For depth-conditioned generation, our method generates diverse backgrounds and textures. For reference conditioned generation, our method preserves the house/dog identity and generates diverse styles, whereas baseline methods have limited sample diversity.

Faithfulness-Diversity Tradeoff with β: The regularization parameter β allows for a trade-off between the faithfulness, i.e., prompt adherence, and diversity of samples. We illustrate this in Fig.5b and Fig.10 **a)**, where lower β has higher diversity but inferior prompt-adherence and vice-versa. This provides a smooth control to the user to adapt our method based on the final use-case. We evaluate this tradeoff across baselines in the appendix.

Scaling Batch Size: To evaluate the robustness of our approach for different batch sizes, we ablate the generation with batch size upto 128 samples. As illustrated in Fig.9, our method maintains high diversity and prompt alignment even at elevated scales, demonstrating superior scalability. In contrast, Group Inference [39] exhibits diminishing diversity as the batch size increases. As the fraction of selected samples

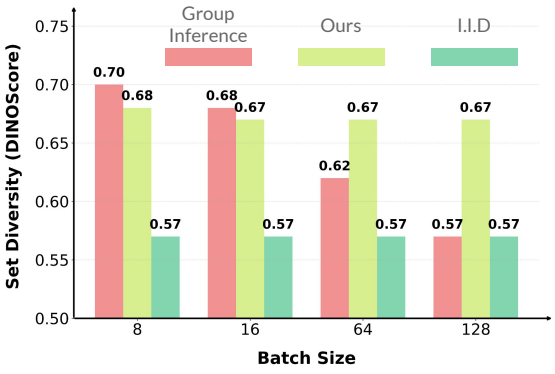


Fig. 9: Performance on scaling Batch Size

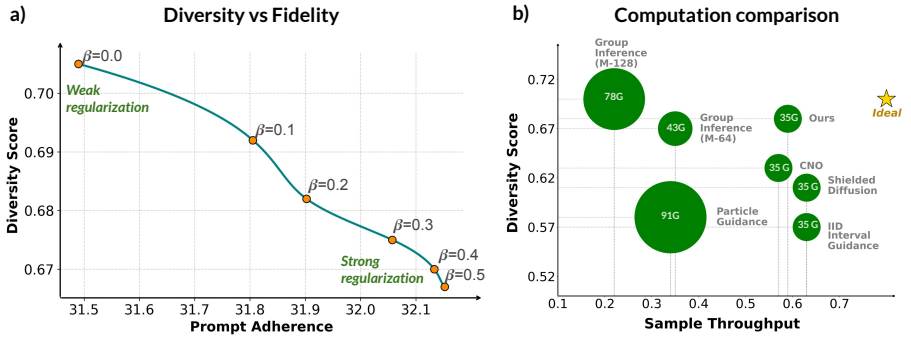


Fig. 10: a) Faithfulness to the Prompt and Diversity Pareto front for our method. Varying β offers a tradeoff between prompt adherence and diversity, higher beta promotes stronger regularization and improves prompt adherence and vice-versa. **b) Computational Cost v/s. Diversity on FLUX.1-dev**: We show a breakdown of the memory consumption, sample throughput and diversity of all methods. Radius of the bubbles indicate VRAM consumption. Our method achieves high diversity alongwith good sample throughput.

increases, inability to discard the similar samples causes diversity to regress toward I.I.D. baseline performance. Our method avoids these sampling bottlenecks, preserving output faithfulness and variance across large-scale generations.

4.5 Computation Analysis

We evaluate the efficiency and computational overhead of our proposed method relative to the baselines. All the benchmarking experiments were conducted using the FLUX.1-dev [30] model on a single NVIDIA H200 GPU. Our evaluation specifically characterizes the trade-off between enhanced generative diversity and incremental computational cost in terms of sample throughput and GPU VRAM used. To ensure statistical reliability, we generate 400 samples per method, incorporating a preliminary warmup phase, and record the average sample throughput and peak VRAM utilization. We present the comparison in Fig.10 **b)**. In this visualization, bubble size is proportionate to the VRAM utilization, the x-axis is the sample throughput, which is the number of samples generated per second, and the y-axis is the DINO diversity score. Our method achieves high diversity with a computational footprint nearly identical to the I.I.D. baseline. Group Inference has marginally higher diversity than ours at a cost of smaller throughput and more VRAM, limiting its practical utility. All the other baselines, though efficient, do not achieve good diversity. Our method achieves an excellent tradeoff in terms of quality, efficiency, and memory required.

4.6 User Study

We conduct a user study to subjectively evaluate the samples generated from our method against three baselines **a)** Interval Guidance, **b)** CFG and **c)** Group Inference across three dimensions: **i)** *Image Quality*, **ii)** *Prompt Alignment* and **iii)** *Image Diversity*. The study consists of 20 comparison pairs evaluated by 15 volunteers. Results of the user study are visualized in Fig. 11.

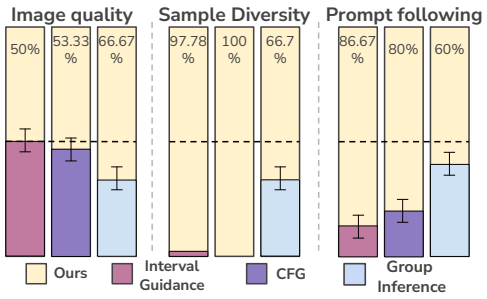


Fig. 11: User Study: User preferences show that our method substantially outperforms baselines in sample diversity while delivering competitive image quality and prompt following scores.

5 Conclusion and Discussion

Limitations. While our approach excels in effectively improving diversity, it has some limitations. As an inference-time intervention, our method is inherently limited by the base model and may inherit biases present in the pretrained model. Additionally, because the approach relies on batch-level computations, it is not effective in streaming image generation.

Conclusion. We introduce a training-free framework designed to systematically rectify diversity collapse in pretrained flow models. Our methodology is built on the key insight that diversity collapse stems from the collapse of internal features of the flow model. To address this, we present a lightweight feature self-guidance mechanism that effectively disperses these features, coupled with a manifold regularization step to ensure that updated features remain anchored to the data distribution. This two-step approach maintains high structural fidelity and alignment with input prompts while functioning as a seamless, plug-and-play module across various conditional flow generative models. Our method takes an important step towards efficiently and effectively improving the pretrained flow models and can fuel more research in investigating the internal features of the flow models. As a future direction, we aim to extend this framework to enhance diversity in other modalities such as 3D model synthesis, video generation, and the generation of diverse molecular structures.

Acknowledgement

We thank Tejan Karmali and Vaibhav Agrawal for reviewing the manuscript. This work is supported by Google, Kotak IISc AI-ML Centre, and Rishubh Parihar is supported by PMRF by the Govt. of India.

References

1. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Transactions on Graphics (TOG) **42**(6), 1–10 (2023)
2. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7877–7888 (2025)
3. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 843–852 (2023)
4. Berrada, T., Romero-Soriano, A., Drozdzal, M., Verbeek, J., Alahari, K.: Entropy rectifying guidance for diffusion and flow models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
5. Corso, G., Xu, Y., De Bortoli, V., Barzilay, R., Jaakkola, T.: Particle guidance: non-iid diverse sampling with diffusion models. arXiv preprint arXiv:2310.13102 (2023)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021), https://transformer-circuits.pub/2021/framework/index.html
8. Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable image generation. Advances in Neural Information Processing Systems **36**, 16222–16239 (2023)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
10. Eyring, L., Karthik, S., Dosovitskiy, A., Ruiz, N., Akata, Z.: Noise hypernetworks: Amortizing test-time compute in diffusion models. arXiv preprint arXiv:2508.09968 (2025)
11. Eyring, L., Karthik, S., Roth, K., Dosovitskiy, A., Akata, Z.: Reno: Enhancing one-step text-to-image models through reward-based noise optimization. Advances in Neural Information Processing Systems **37**, 125487–125519 (2024)
12. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410 (2022)
13. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023)

15. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
16. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
17. Harrington, A., Koepke, A., Karthik, S., Darrell, T., Efros, A.A.: It's never too late: Noise optimization for collapse recovery in trained diffusion models. arXiv preprint arXiv:2601.00090 (2025)
18. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
21. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
22. Jalali, M., Lei, H., Gohari, A., Farnia, F.: Sparke: Scalable prompt-aware diversity guidance in diffusion models via rke score. arXiv preprint arXiv:2506.10173 (2025)
23. Kazimi, T., Dunlop, C., Yanardag, P.: Diverse video generation with determinantal point process-guided policy optimization. arXiv preprint arXiv:2511.20647 (2025)
24. Khan, M.A.H., Jain, Y., Bhattacharyya, S., Vineet, V.: Test-time prompt refinement for text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 6506–6516 (October 2025)
25. Kim, B., Um, S., Ye, J.C.: Diverse text-to-image generation via contrastive noise optimization. arXiv preprint arXiv:2510.03813 (2025)
26. Kim, J., Yoon, T., Hwang, J., Sung, M.: Inference-time scaling for flow models via stochastic generation and rollover budget forcing. arXiv preprint arXiv:2503.19385 (2025)
27. Kirchhof, M., Thornton, J., Béthune, L., Ablin, P., Ndiaye, E., Cuturi, M.: Shielded diffusion: Generating novel and diverse images using sparse repellency. arXiv preprint arXiv:2410.06025 (2024)
28. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
29. Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models. Advances in Neural Information Processing Systems **37**, 122458–122483 (2024)
30. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024), accessed: 2025-10-01
31. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025), accessed: 2026-2-01
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
34. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)

35. Luo, G., Granskog, J., Holynski, A., Darrell, T.: Dual-process image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17972–17983 (2025)
36. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., et al.: Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732 (2025)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Parihar, R., Bhat, A., Basu, A., Mallick, S., Kundu, J.N., Babu, R.V.: Balancing act: Distribution-guided debiasing in diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6668–6678 (2024)
39. Parmar, G., Patashnik, O., Ostashev, D., Wang, K.C., Aberman, K., Narasimhan, S., Zhu, J.Y.: Scaling group inference for diverse and high-quality generation. arXiv preprint arXiv:2508.15773 (2025)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
42. Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., Weber, R.M.: Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. arXiv preprint arXiv:2310.17347 (2023)
43. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)
44. Snell, C., Lee, J., Xu, K., Kumar, A.: Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314 (2024)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
46. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
47. Tang, Z., Peng, J., Tang, J., Hong, M., Wang, F., Chang, T.H.: Inference-time alignment of diffusion models with direct noise optimization. arXiv preprint arXiv:2405.18881 (2024)
48. Tong, S., Zheng, B., Wang, Z., Tang, B., Ma, N., Brown, E., Yang, J., Fergus, R., LeCun, Y., Xie, S.: Scaling text-to-image diffusion transformers with representation autoencoders. arXiv preprint arXiv:2601.16208 (2026)
49. Um, S., Ye, J.C.: Minority-focused text-to-image generation via prompt optimization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20926–20936 (2025)
50. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
51. Wang, R., He, K.: Diffuse and disperse: Image generation with representation regularization. arXiv preprint arXiv:2506.09027 (2025)

52. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
53. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
54. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025)
55. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
56. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
57. Yun, T., Zhang, D., Park, J., Pan, L.: Learning to sample effective and diverse prompts for text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23625–23635 (2025)
58. Zhao, B.N., Xiao, Y., Xu, J., Jiang, X., Yang, Y., Li, D., Itti, L., Vineet, V., Ge, Y.: Dreamdistribution: Learning prompt distribution for diverse in-distribution generation. In: The Thirteenth International Conference on Learning Representations (2025)
59. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
60. Zhou, X., Li, Q., Hu, X., Chen, H., Gu, S.: Guiding a diffusion transformer with the internal dynamics of itself. arXiv preprint arXiv:2512.24176 (2025)
61. Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., Li, H.: From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15329–15339 (October 2025)