Entropy Rectifying Guidance for Diffusion and Flow Models

Tariq Berrada Ifriqi[1,2] **Adriana Romero-Soriano**[1,3,4,5]
Michal Drozdzal[1] **Jakob Verbeek**[1] **Karteek Alahari**[2]
[1] FAIR at Meta [2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France
[3] McGill University [4] Mila, Quebec AI institute [5] Canada CIFAR AI chair
tariqberrada@meta.com

Abstract

Guidance techniques are commonly used in diffusion and flow models to improve image quality and input consistency for conditional generative tasks such as class-conditional and text-to-image generation. In particular, classifier-free guidance (CFG) is the most widely adopted guidance technique. It results, however, in trade-offs across quality, diversity and consistency: improving some at the expense of others. While recent work has shown that it is possible to disentangle these factors to some extent, such methods come with an overhead of requiring an additional (weaker) model, or require more forward passes per sampling step. In this paper, we propose *Entropy Rectifying Guidance* (ERG), a simple and effective guidance method based on inference-time changes in the attention mechanism of state-of-the-art diffusion transformer architectures, which allows for simultaneous improvements over image quality, diversity and prompt consistency. ERG is more general than CFG and similar guidance techniques, as it extends to unconditional sampling. We show that ERG results in significant improvements in various tasks, including text-to-image, class-conditional and unconditional image generation. We also show that ERG can be seamlessly combined with other recent guidance methods such as CADS and APG, further improving generation results.

1 Introduction

Diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Dhariwal and Nichol, 2021) and flow models (Lipman et al., 2023; Ma et al., 2024; Esser et al., 2024) are state-of-the-art generative modeling tools for various modalities, ranging from images (Rombach et al., 2022; Podell et al., 2024; Chen et al., 2024; Esser et al., 2024), to audio (Wang et al., 2023; Levy et al., 2023), and video (Jin et al., 2025; Polyak et al., 2024). These models generate data by starting with a simple prior and iteratively removing noise – a process called "denoising". See Kingma and Gao (2023); Lipman et al. (2023) for details on diffusion and flow matching. These models can be conditioned on various inputs to control the generative process, e.g., in text-to-image models. Guidance techniques such as classifier guidance (Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2021) are commonly used to improve sample quality and consistency with input conditioning. In the sampling process, these techniques combine a conditional signal with an unconditional one, and control the strength of conditioning through a scaling parameter. Although such guidance techniques are a crucial component in achieving state-of-the-art results, they are also known to negatively affect the diversity of generated samples given a particular prompt (Sadat et al., 2024; Karras et al., 2024; Kynkäänniemi et al., 2024; Saharia et al., 2022). Moreover, too high guidance scales may lead to overly saturated images, affecting the quality of generated images (Saharia et al., 2022). An extensive analysis on the trade-offs of image generation quality, diversity, and consistency was presented by Astolfi et al. (2024). To mitigate these quality-diversity-consistency trade-offs, more advanced

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

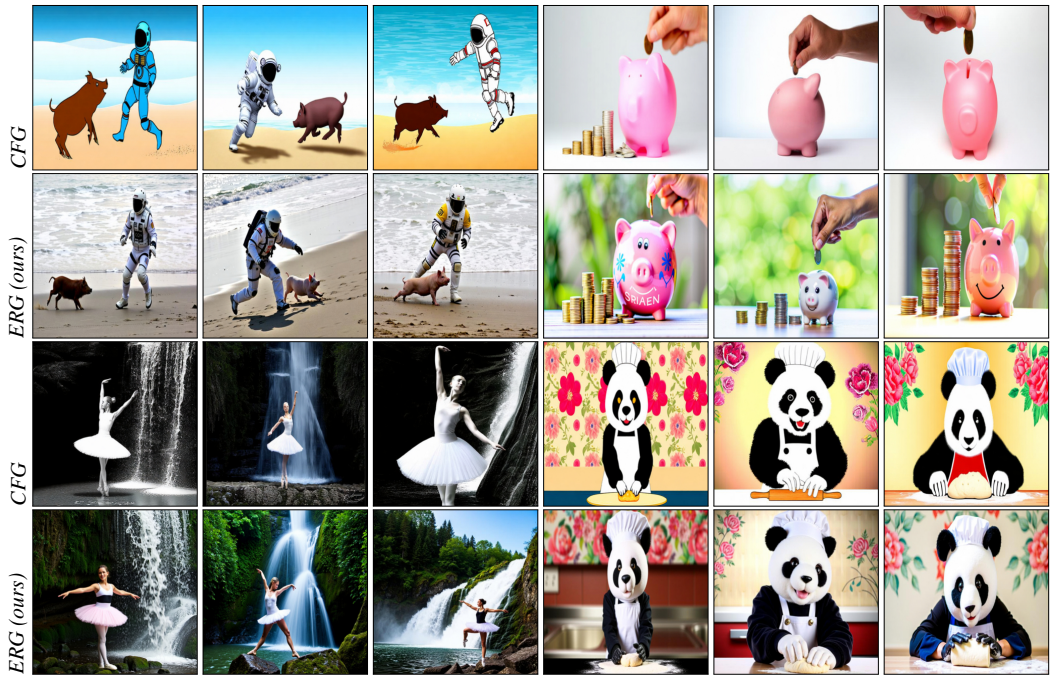

Figure 1: **Qualitative comparison of classifier-free guidance (CFG) and our Entropy Rectifying Guidance (ERG).** The images generated using ERG (second and fourth row) exhibit greater quality and diversity than standard CFG. Images are generated using 50 Euler steps; each column corresponds to a different random seed for the generations.

guidance techniques have recently been proposed, see e.g. Sadat et al. (2025); Karras et al. (2024). Most guidance techniques , however, require spending part of the training cycles on unconditional generation for the guidance to work, even if unconditional generation is not a goal in itself, and are not applicable to unconditional sampling (Sadat et al., 2025; Ho and Salimans, 2021). Others rely on a second model with weaker performance than the main model, thereby increasing memory requirements (Karras et al., 2024).

In our work, we build upon the work of Karras et al. (2024) and Hong (2024), and propose Entropy Rectifying Guidance (ERG): a simple and effective method to obtain *both* a *strong* and a *weak* predictive signal from a *single model* that leverages attention layers, where the model may be conditional or unconditional. In particular, our method uses the Hopfield energy formulation of attention (Ramsauer et al., 2021; Hong, 2024) and applies a temperature scaling to the softmax function of the attention layers in order to obtain the weak predictive signal. This scaling does not require any adaptations in model training, and may be applied to pre-trained denoising models, as well as their accompanying text encoders. Moreover, motivated by this energy interpretation of attention layers, we also consider iterative re-application of attention layers, and rescaling the residual attention update. Our experiments show that manipulating the attention layers in this manner results in simultaneous improvements in sample quality, diversity, and consistency, contrary to the trade-offs observed by Astolfi et al. (2024) for classifier-free guidance.

In summary, our contributions are the following:

1. We propose Entropy Rectifying Guidance (ERG), a guidance mechanism based on modifying the energy landscape of the attention layers.
2. Since our guidance mechanism does not require unconditional inference, it is directly applicable to any attention-based diffusion or flow model, including unconditional, class-conditional, and text-to-image models.
3. Experimentally, we find that ERG significantly improves image quality *and* diversity while retaining the same prompt consistency as the standard classifier-free guidance ($+30$ points in density, $+4$ points in VQAScore) on COCO at 512 resolution using our 1.9B text-to-image model.

2 Related work and background

In this section, we briefly review background material on diffusion and flow models, and related work on guidance techniques to improve sampling as well as the Hopfield energy formulation of attention.

2.1 Diffusion and flow matching models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) form a flexible class of generative models whose underlying principle is to map samples $\boldsymbol{\epsilon}$ from a trivial unit Gaussian prior $p_0 = \mathcal{N}(0, \mathbf{I})$ to samples from a learned model p_1 of the data distribution. The forward process is defined as: $\mathbf{x}_t = \alpha_t \mathbf{x}_1 + \sigma_t \boldsymbol{\epsilon}$ with $t \in [0, 1]$, where $\mathbf{x}_1 \sim p_1$, and α_t is a decreasing function of "time" t while σ_t an increasing function of t.

Flow matching methods (Lipman et al., 2023) assume that $\alpha_0 = \sigma_1 = 1$ and $\alpha_1 = \sigma_0 = 0$. Using these assumptions, during the reverse process $\mathbf{x}_t$ interpolates between $\boldsymbol{\epsilon}$ at $t = 0$ and $\mathbf{x}_1$ at $t = 1$. In contrast, score-based diffusion models (Ho et al., 2020; Dhariwal and Nichol, 2021) set α_t and σ_t implicitly through different formulations of stochastic differential equations (SDE) where $\mathcal{N}(0, \mathbf{I})$ is the equilibrium distribution. Additionally, they consider $t \in [0, T]$ with T large enough so that $\mathbf{x}_T$ is approximately distributed as a unit Gaussian random variable.

2.2 Guidance mechanisms

Classifier guidance. To enable high quality conditional generation, Dhariwal and Nichol (2021) proposed to guide the sampling process by leveraging gradients from pre-trained auxiliary classifier $p(c|\mathbf{x})$ in each denoising step. They use the classifier to define the (scaled) joint score function as $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, c) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + w \nabla_{\mathbf{x}_t} \log p(c|\mathbf{x}_t)$, where $p(\mathbf{x}_t)$ is an unconditional data model, and w is a scalar parameter regulating the strength of the classifier guidance. While classifier guidance allows to improve input consistency and image quality (at the expense of diversity), it requires an auxiliary classification model that is robust to inputs $\mathbf{x}_t$ with varying amounts of noise.

Classifier-free guidance. To avoid the need for an auxiliary noise-robust classifier, Ho and Salimans (2021) proposed classifier-free guidance (CFG). In this case, during the training process, two generative models are learned, one conditional $p(\mathbf{x}|c)$ and one unconditional $p(\mathbf{x}|\emptyset)$. In practice, the unconditional model is trained by dropping conditioning information with a small probability. The score function used for sampling is extrapolated towards the conditional prediction and away from the unconditional prediction $\nabla_{\mathbf{x}}^{\text{CFG}} \log p(\mathbf{x}|c) = w \nabla_{\mathbf{x}} \log p(\mathbf{x}|c) + (1 - w) \nabla_{\mathbf{x}} \log p(\mathbf{x}|\emptyset)$. While CFG improves image quality and input consistency with respect to classifier guidance, it tends to come at the cost of a reduction in diversity (Astolfi et al., 2024; Ho and Salimans, 2021; Sadat et al., 2024). Moreover, CFG often leads to generation artifacts, such as over-saturation as the guidance scale w grows (Sadat et al., 2025).

Advanced guidance techniques. Several improved variants of classifier-free guidance have been proposed recently. Hong et al. (2023) presented Self-Attention Guidance (SAG), a guidance mechanism based on feeding a modified intermediate sample $\mathbf{x}_t$ when performing inference for unconditional prediction. The modification consists in blurring $\mathbf{x}_t$ in regions that are most attended to by the model's self-attention. This method has been developed for the U-Net architecture, hence applying it to more recent diffusion transformer architectures requires a hyperparameter search to understand which attention layers should be used for this method.

Smoothed energy guidance (SEG) (Hong, 2024) contrasts conditional prediction with a "weaker" conditional prediction obtained by altering the attention's softmax energy with a Gaussian kernel applied to queries. This method is developed for U-Net-style architectures and the softmax alteration applies to the self-attention layers in the middle block of the U-Net. In the conditional case, SEG uses a linear combination of the conditional, unconditional, and energy-smoothed unconditional prediction, whereas in our approach we only use the conditional and smoothed conditional term. Thus, SEG requires an additional function evaluation with respect to ERG for conditional inference. For unconditional inference, both approaches require only two function evaluations. In a similar spirit, Ahn et al. (2024) propose a guidance method based on manipulating the attention mechanism, by replacing the attention matrix with an identity mapping inside the denoiser U-Net.

Karras et al. (2024) proposed AutoGuidance, a method that uses a smaller/weaker version of the same conditional model for classifier-free guidance, resulting in better diversity and image quality. Like ours, their approach can also be applied to unconditional sampling. However, their method requires access to an earlier checkpoint of the model or, for best results, training a separate model with lower capacity, as well as accessing two models when sampling, which may increase the memory footprint.

Rather than considering modifications of the unconditional model term, Sadat et al. (2024) proposed the "condition-annealed diffusion sampler" (CADS) to increase the diversity of generations while maintaining sample quality. This is achieved by adding Gaussian noise to the conditioning tokens during inference, using a piecewise-linear decreasing schedule on the noise amplitude. Sadat et al. (2025) propose "adaptive projected guidance" (APG), a variant of CFG that resolves the over-saturation problem by emphasizing guidance orthogonal to the conditional prediction, rescaling the guidance term, and introducing a negative momentum term. The latter two approaches can be easily combined with our approach, and we consider such combinations in our experiments.

Lastly, several works (Kynkäänniemi et al., 2024; Chung et al., 2024a; Pavasovic et al., 2025; Wang et al., 2024) explore non-constant weight schedules for CFG. Such methods are complementary to our work, as our method only operates at the architecture level by rectifying the attention updates. Although they therefore could be combined with our approach, we defer this to future work.

2.3 Hopfield energy formulation of attention

The Hopfield network (Hopfield, 1982) is a dense associative memory model that aims to associate an input with its most similar pattern. More specifically, it constructs an energy function to model an energy landscape that contains basins of attraction around the desired patterns. Modern Hopfield energy networks (Ramsauer et al., 2021) introduce a new family of energy functions that improve the storage capacity of the model and make it compatible with continuous embeddings. Specifically, the following energy functional matches a continuous d-dimensional state (query) pattern $\boldsymbol{\xi} \in \mathbb{R}^d$ with N stored (key) patterns $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N) \in \mathbb{R}^{d \times N}$ as $E(\boldsymbol{\xi}; \mathbf{X}) = \frac{1}{2}\boldsymbol{\xi}^\top \boldsymbol{\xi} - \text{LogSumExp}\left(\mathbf{X}^\top \boldsymbol{\xi}, \beta\right)$, where $\text{LogSumExp}(\mathbf{x}, \beta) = \beta^{-1} \log\left(\sum_{i=1}^d \exp(x_i)\right)$, where x_i are the elements of the vector $\mathbf{x}$, and β a scalar hyperparameter defining the sharpness of the approximation of the maximum in the LogSumExp operation. Intuitively, the first term imposes a finite norm on the queries while the second term measures the alignment between the state patterns (queries) and stored patterns (keys). Using the Concave-Convex Procedure (CCCP) (Yuille and Rangarajan, 2003), an iterative update rule, which converges to the global minima of the energy, can be derived as $\boldsymbol{\xi}_{l+1} = \mathbf{X}\ \text{SoftMax}(\beta \mathbf{X}^\top \boldsymbol{\xi}_l)$, where $\text{SoftMax}(\mathbf{x}) = \exp(\mathbf{x} - \text{LogSumExp}(\mathbf{x}, 1))$. Equivalently, each iterative update step can be seen as a gradient update in the negative direction of the energy $\nabla_{\boldsymbol{\xi}} E(\boldsymbol{\xi}; \mathbf{X}) = \boldsymbol{\xi} - \mathbf{X}\ \text{SoftMax}(\beta \mathbf{X}^\top \boldsymbol{\xi})$. Taking a gradient descent step on this energy landscape with a step size γ results in an update of the form $\boldsymbol{\xi}_{l+1} = \boldsymbol{\xi}_l - \gamma\left(\boldsymbol{\xi}_l - \mathbf{X}\text{SoftMax}(\beta \mathbf{X}^\top \boldsymbol{\xi}_l)\right)$, and for $\gamma = 1$ we recover the CCCP update.

Ramsauer et al. (2021) show that the CCCP update is related to the standard attention operation as follows. Assuming there are S state (query) patterns, and N stored (key) patterns that can be mapped to keys, queries and values using linear transformations, he state pattern can be obtained through a concatenation: $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_S] \in \mathbb{R}^{d \times S}$. Then the attention map is given by $\Xi_{l+1} = \mathbf{X}\text{SoftMax}(\mathbf{X}^\top \Xi_l)$. The keys, queries and values are obtained via linear projections as $\mathbf{K} = \mathbf{X}W_K^\top \in \mathbb{R}^{N \times d}$, $\mathbf{V} = \mathbf{X}W_V^\top \in \mathbb{R}^{N \times d}$ and $\mathbf{Q} = \boldsymbol{\Xi}_{l+1} W_Q^\top \in \mathbb{R}^{S \times d}$, respectively. By setting $\beta = \frac{1}{\sqrt{d}}$ and substituting this into the CCCP update rule, we obtain: $\mathbf{Q}_{l+1} = \text{SoftMax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{K} = \text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}\left(W_K W_V^{-1}\right)^\top$. Hence, we observe the equation of attention up to a linear transformation, which provides an interpretation of the attention mechanism through the Hopfield energy lens. Such a formulation has been used in the Energy Based Cross-Attention method (EBCA) (Park et al., 2024) for adaptive context control in order to incorporate additional contexts into the generative process of a conditional diffusion model.

3 Guidance via entropy rectification

Similarly to previous approaches (Ho and Salimans, 2021; Hong, 2024; Karras et al., 2024), our approach is based on contrasting the (conditional) denoising estimate with a less powerful one. In particular, using the Hopfield energy interpretation of attention (Ramsauer et al., 2021), see

Section 2.3, we manipulate the energy landscape of the attention layer, by rectifying the entropy of the associations in the attention operation. The modified attention layers lead to lower quality predictions, which are used as the contrasting term for guidance. This approach does not require a second model, and can be used for both conditional and unconditional sampling. We refer to our approach as Entropy Rectifying Guidance (ERG).

3.1 Manipulating the energy landscape

Our method manipulates the energy landscapes by introducing two new test-time hyperparameters, α and τ in the energy function:

$$E(\boldsymbol{\xi}; \mathbf{X}) = \frac{1}{2}\boldsymbol{\xi}^\top\boldsymbol{\xi} - \alpha \cdot \text{LogSumExp}\left(\mathbf{X}^\top\boldsymbol{\xi}, \tau \cdot \beta\right), \tag{1}$$

where $\beta = \frac{1}{\sqrt{d}}$ is the default temperature of the attention attention update. The temperature rescaling parameter τ controls the sharpness of the softmax attention, and α the relative importance of the similarity between the state matching term compared to the norm of the state patterns. Temperature rescaling with τ is similar to the Gaussian blurring of the attention maps introduced in SEG (Hong, 2024), but allows for non-local smoothing of the attention maps. Additionally, the view of the attention layer as a CCCP update of the energy function, allows for consideration of different methods to minimize the energy landscape. For instance, taking K gradient descent steps with step size γ, as illustrated in Algorithm 1.

Algorithm 1 Entropy rectifying guidance

Require: $K \in \mathbb{N}$ number of gradient update steps.
Require: $\gamma > 0$ step size.
Require: $\alpha \in \mathbb{R}$ State pattern matching weight.
Require: $\tau > 0$ attention temperature.
Require: $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{d \times N}$ keys and values.
Require: $\mathbf{Q} \in \mathbb{R}^{d \times S}$ queries.
$k \leftarrow 0$
while $k < K$ **do**
$\mathbf{Q} \leftarrow \mathbf{Q} - \gamma \left(\mathbf{Q} - \alpha \cdot \text{softmax}\left(\tau \cdot \beta \mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}\right)$
$k \leftarrow k + 1$
end while

Using different settings of the hyperparameters α, γ, τ and K allows us to manipulate the attention operation, and obtain noise estimates that deviate from the trained model. We expect these to be weaker estimates compared to those provided by the model, as it was trained with standard attention layers, i.e. with $\alpha = \gamma = \tau = K = 1$. When applying this rectification mechanism to the denoiser model, we refer to the method as image-ERG, or I-ERG for short. In particular, we apply it to the negative/unconditional prediction part of the classifier-free guidance, and only on certain layers of the network that we will identify in our experiments. Additionally, we impose a kickoff threshold κ on the time steps in which guidance is applied.

3.2 Manipulating the energy of the text encoder

Besides the image denoising model, we can also manipulate the energy landscape of attention-based text encoders to obtain a weak version of the conditional embeddings. Let c be the text prompt that is used as conditioning. The text tokens are obtained by feeding the prompt to the text encoder: $\text{Enc}(c) \in \mathbb{R}^{d_t}$, such as Llama (Grattafiori et al., 2024) or T5 (Chung et al., 2024b). Text tokens are then fed to the denoiser model through cross-attention layers.

To obtain a contrasting signal for guidance, we manipulate the energy landscape in the self-attention layers of the text encoder following Algorithm 1. More specifically, for every self-attention layer in the text encoder, we introduce a temperature hyperparameter in the softmax function. This enables us to change the strength by which keys and queries are being matched, resulting in a modified prompt embedding at the output. For the remainder of the manuscript, we refer to this method condition-ERG, or C-ERG for short. For simplicity reasons, for the text encoder, we only consider changing the temperature but not the step size γ, pattern matching weight α, and number of update steps K.

3.3 Guidance update

When combining the energy modulations in the text-encoder and denoising model, we obtain our Entropy Rectifying Guidance (ERG) update:

$$\Delta_{\text{ERG}}(\mathbf{x}, c, t; \Theta_\xi) = w \cdot D(\mathbf{x}, \boldsymbol{\phi}_c, t) + (1 - w) \cdot D^{\boldsymbol{\xi}}(\mathbf{x}, \boldsymbol{\phi}_c^\tau, t; \Theta_\xi), \tag{2}$$

where D is the learned denoiser model with parameters Θ that are omitted for simplicity, $D^{\boldsymbol{\xi}}$ is the denoiser model where the attention layers have been replaced by the modified version presented in Equation (1), and $\Theta_\xi = \{\alpha, \gamma, \tau, K\}$ the set of hyperparameters introduced by ERG. We use $\boldsymbol{\phi}_c$ to denote the prompt embedding produced by the text encoder, while $\boldsymbol{\phi}_c^\tau$ denotes the embedding obtained with the modified attention layers.

Compared to standard CFG, the main differences are that (i) we replace the unconditional text embeddings with conditional embeddings obtained with the entropy-rectified attention mechanism (C-ERG), and (ii) the denoiser model for the negative/unconditional predictions also uses modified attention layers (I-ERG). Finally, the changes to the image denoiser are only applied after a certain point during sampling in order to not overly penalize the negative components of the ERG update at the start of sampling. For the text-encoder we apply the temperature scaling throughout the denoising process, so that at any stage we obtain a noise prediction that can be contrasted with the vanilla denoising signal.

Note that our approach can be combined with other approaches, e.g., CADS (Sadat et al., 2024) and APG (Sadat et al., 2025). We will explore such combinations in our experiments.

4 Experimental evaluation

4.1 Experimental setup

Datasets and architectures. We experiment with class-conditional and text-to-image models trained using rectified flow-matching (Lipman et al., 2023; Esser et al., 2024). We use a face-blurred version of ImageNet (Deng et al., 2009) to train class-conditional models at 256 and 512 resolution based on the XL/2 variant of the DiT architecture (Peebles and Xie, 2023), which is composed of 28 attention blocks with hidden dimension of 1152, resulting in 790M parameters. For the text-to-image model, we use an architecture similar to MMDiT (Esser et al., 2024), and train a 512 resolution model on a mix of a proprietary dataset of 320M text-image pairs and YFCC100M (Thomee et al., 2016), where all faces in YFCC100M have been blurred. Similar to MMDiT (Esser et al., 2024), the model uses a mix of different text encoders: Llama3-8B (Grattafiori et al., 2024) and Flan-T5-XL (Chung et al., 2024b). During training, each of the text encoders is disabled with a probability of $\sqrt{0.1}$, so that the probability of both encoders being disabled is around 10%. We enable both text encoders during inference time for text-to-image generation, and disable both text-encoders for unconditional image generation experiments. The architecture of the model is made of 38 blocks with a hidden dimension of $1,536$, resulting in approx 1.9B parameters. For both datasets, we recaption the images using both Florence-2 Large (Xiao et al., 2024) to obtain medium-length captions and PaliGemma-3B (Beyer et al., 2024) for shorter COCO-style captions. Additional techniques to improve training efficiency of this model, such as conditioning mechanisms and pre-training strategies, were adopted from Berrada et al. (2024). For the class-conditional model we use the asymmetric autoencoder of Zhu et al. (2023), while for the larger text-to-image model we use the SD3 autoencoder (Esser et al., 2024).

Metrics. We consider metrics for quality, diversity, and consistency (Astolfi et al., 2024). We measure *sample quality* with FID (Heusel et al., 2017) and density (Naeem et al., 2020); *sample diversity* is measured with coverage (Naeem et al., 2020) and FID; and *prompt consistency* is measured with CLIPScore (Hessel et al., 2021) and VQAScore (Lin et al., 2024). For evaluation of text-to-image and unconditional generation, we use the 40k COCO'14 validation image-caption pairs. For the class-conditional models, we sample 50 images for each of the 1,000 ImageNet classes and use the ImageNet validation set as a reference. All evaluated models are sampled using the Euler method with 50 sampling steps. We use the EvalGIM (Hall et al., 2024) library for all evaluations.

Baselines. In addition to the standard classifier-free guidance, we compare our method to several recent state-of-the-art guidance techniques: Condition-Annealed Diffusion Sampler (CADS) (Sadat et al., 2024), Adaptive Projected Guidance (APG) (Sadat et al., 2025), Smooth Energy Guidance (SEG) (Hong, 2024), and Auto-Guidance (Karras et al., 2024). For APG, we follow the recommendations from the paper and set $\gamma_{\text{APG}} = -0.5$, $\eta_{\text{APG}} = 0.0$, $r_{\text{APG}} = 5.0$. For CADS, we perform a grid search over $\tau_1^{\text{CADS}} \in [0.6, 0.8], \tau_2^{\text{CADS}} \in [0.8, 1.0], s^{\text{CADS}} \in [0.25, 1.0], \psi^{\text{CADS}} = 1.0$. Since SAG, PAG and SEG were developed specifically for the U-Net architecture, we adapt these method for diffusion transformers (Peebles and Xie, 2023) by applying the method in the attention layers

Table 1: **Comparison of ERG with other guidance approaches for text-to-image generation.** We compare ERG to other state-of-the-art guidance approaches and mark the best result in each column in bold in the top part of the table. In the bottom part of the table we evaluate combinations of ERG with APG and CADS, and bold results when they surpass the results in the upper part of the table.

Guidance \ Metric	FID (↓)	Density (↑)	Coverage (↑)	CLIPScore (↑)	VQAScore (↑)	NFE (↓)
CFG (Ho and Salimans, 2021)	12.81	98.24	71.12	26.45	70.15	2
APG (Sadat et al., 2025)	11.88	104.07	73.06	26.54	72.47	2
CADS (Sadat et al., 2024)	11.93	101.01	72.99	26.76	73.36	2
PAG* (Ahn et al., 2024)	12.75	107.21	72.20	26.80	73.32	2
SAG* Hong et al. (2023)	**11.68**	103.58	72.74	26.81	72.16	2
SEG* (Hong, 2024)	16.87	87.77	61.91	**26.86**	73.59	3
AutoGuidance (Karras et al., 2024)	16.62	87.02	62.75	26.59	73.53	2
ERG (ours)	13.62	**120.25**	**73.21**	**26.86**	**73.96**	2
ERG (ours) + APG	**11.37**	115.08	**80.50**	26.74	73.55	2
ERG (ours) + CADS	12.87	**128.54**	76.23	26.75	73.45	2

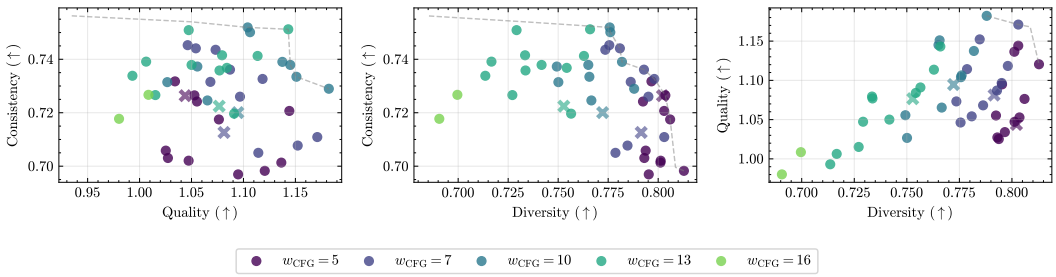

Figure 2: **Pareto fronts on consistency-diversity-quality for text-to-image generation.** Comparing ERG + APG (dots) with APG (crosses). We sweep over different guidance scales (each marked with a different color), and hyper-parameters α, γ, τ for ERG. Dashed lines trace the Pareto fronts for each plot. We measure consistency with VQAScore, quality with density and diversity with coverage.

of the middle blocks of the transformer; we refer to these methods with an asterisk superscript. Additional details on the choice of layers are provided in Appendix D.4. For AutoGuidance, we follow the recommendations from the paper and use an earlier checkpoint of the same model, at approximately $1/16$-th of the training, as the weaker model. To ensure a fair comparison, we select the best performing guidance strength for baselines as well as our method using the rank-scoring algorithm detailed in Appendix A.2. Note that this is different from reporting the optimal score achieved for each metric, which might not correspond to any particular run because of inherent trade-offs between the different facets of the generations.

4.2 Main experimental results

Throughout our experiments, we modify attention layers in both the text encoder (C-ERG) and the image denoising model (I-ERG), unless specified otherwise.

Text-to-image generation. In Table 1, we compare ERG with recent state-of-the-art guidance mechanisms on the text-to-image generation task. ERG demonstrates excellent performance, outperforming baselines such as CFG, SAG, PAG and SEG* in most metrics. Specifically, ERG considerably boosts image quality as reported for Density (e.g., +22 points when compared to CFG). Additionally, ERG achieves the highest consistency scores: $+0.4$ points in CLIPScore and $+3.8$ in VQAScore when compared to standard CFG. Moreover, ERG combined with APG achieves the overall best diversity, measured by FID and Coverage, and also improves Density over prior methods. These results suggest that our ERG approach is successfully able to boost the generation quality of the model in all three facets of generation (quality, diversity, and consistency). Moreover, the number of function evaluations (NFE) required for our approach is only two per inference step, which is comparable to other methods except SEG which requires three.

Following Astolfi et al. (2024), we plot Pareto fronts for quality (measured by Density), diversity (Coverage), and consistency (VQAScore) metrics in Figure 2 when using ERG + APG. We find all

Figure 3: **Unconditional generation results.** Compared to not using guidance (top), our ERG generates more realistic and detailed images and more coherent structure (bottom). Images obtained from T2I model at 512 with empty prompt as input. Samples in each column use the same seed.

Table 2: **Unconditional generation.** Comparing ERG to other approaches compatible with unconditional sampling.

	FID (↓)	Density (↑)	Coverage (↑)
No guidance	101.50	8.99	3.63
SAG (Hong et al., 2023)	39.25	46.22	30.70
PAG* (Ahn et al., 2024)	41.50	45.65	30.51
SEG* (Hong, 2024)	37.75	55.56	34.79
AutoGuidance (Karras et al., 2024)	39.50	48.26	34.71
ERG (ours)	**36.25**	**55.84**	**51.59**

Table 3: **Class-conditional generation.** Comparison of ERG with other guidance methods for models trained for 256 and 512 resolution.

	Res.	FID (↓)	Density (↑)	Coverage (↑)
CFG (Ho and Salimans, 2021)	256	**3.67**	127.03	85.81
PAG* (Ahn et al., 2024)		5.31	111.94	81.04
SAG (Hong et al., 2023)		3.78	131.89	86.35
SEG* (Hong, 2024)		6.15	132.22	84.51
ERG (ours)		**3.67**	**141.96**	**86.72**
CFG (Ho and Salimans, 2021)	512	5.65	146.97	**86.70**
PAG* (Ahn et al., 2024)		4.65	134.49	86.50
SAG (Hong et al., 2023)		4.81	120.09	83.91
SEG* (Hong, 2024)		6.59	160.11	81.85
ERG (ours)		**4.56**	**163.63**	86.13

the points belonging to the Pareto fronts correspond to ERG + APG, which improves in all three facets of the generations w.r.t. APG, and provides significant boosts for quality and consistency. This can also be observed in the qualitative comparison between APG and ERG + APG in Figure 13 and Figure 14 in the supplementary material. Similarly, in Figure 1 the images sampled using ERG show better visual quality than those sampled with CFG.

Unconditional generation. In the unconditional model sampling experiment, we compare I-ERG with sampling without guidance and using applicable methods: AutoGuidance, SEG*, PAG* and SAG*. We provide quantitative evaluation results in Table 2, where we find that ERG outperforms all other methods, with significant boosts in FID, Density, and Coverage over other methods. The qualitative examples in Figure 3 clearly exhibit artifacts in terms of structural coherence in all the objects present in the generated images when not using guidance, which disappear when using ERG.

Class-conditional generation. For class-conditional generation, we observe similar trends as those seen for text-to-image and unconditional sampling in Table 3. In particular, at 256 resolution, we find improvements in Density and Coverage. FID remains similar to CFG but is better compared to all other methods tested. At 512 resolution, ERG is best across all metrics, except for coverage where ERG is slightly behind CFG and SEG*.

4.3 Analysis and ablations

Text attention energy. To isolate the effect of the temperature re-scaling in the text-encoder and image denoising components, we experiment with C-ERG and disable temperature rescaling in the image denoising model. In Figure 4, we vary the SoftMax temperature τ_c used in the text encoder, and consider generation performance for different guidance strengths ω. For all metrics we find a somewhat symmetrical behavior around $\tau_c = 1$, which corresponds to an unguided prediction because in this case there is no difference with the normal conditional prediction.

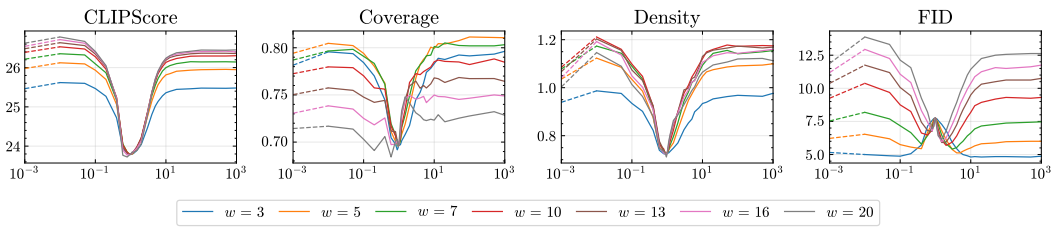

Figure 4: **Temperature rescaling in the conditioning for text-to-image generation (C-ERG).** We vary the text encoder's attention temperatures τ_c. Each curve corresponds to a different guidance strength w. The left-most point on each curve represents the result for standard CFG.

Table 4: **Impact of the different components of ERG.** We accumulate different components of ERG, namely C-ERG, then I-ERG through denoiser entropy rectification and objective reweighting, finally all are merged into ERG.

τ_c	τ_i	γ	FID (↓)	Density (↑)	Coverage (↑)	CLIP (↑)	VQA (↑)
✗	✗	✗	**12.81**	98.24	71.12	26.45	70.15
✓	✗	✗	13.06	109.52	72.06	26.73	73.10
✓	✓	✗	13.62	120.25	73.21	**26.86**	73.96
✓	✓	✓	13.62	**123.65**	**74.07**	26.81	**74.67**

Table 5: **Multi-step gradient descent for optimizing the energy landscape.** we ablate different values for K and γ.

K	1	1	1	10
γ	1	1.5	0.5	0.1
FID (↓)	13.62	13.62	13.06	**12.68**
Density (↑)	120.25	**123.65**	121.95	119.07
Coverage (↑)	73.21	**74.07**	72.63	72.11
CLIP (↑)	**26.86**	26.81	26.16	26.77
VQA (↑)	73.96	**74.67**	73.01	73.95

The CLIPScore, Coverage and Density metrics generally improve when moving away from $\tau_c = 1$, making it either larger or smaller. For FID, the best values are obtained with low guidance scales and intermediate temperature values, the general trend shows improved FID using C-ERG compared to standard guidance. Compared with standard classifier-free guidance (left-hand side of the dashed lines), we find that all metrics can be improved for any guidance scale, provided with the right temperature. While $\tau_c < 1$ results in higher CLIPScore, $\tau_c > 1$ results in higher Coverage, indicating that tuning τ_c provides an easy way to control the diversity-consistency tradeoff.

Combining the different parts. In Table 4, we combine different parts of ERG and measure the effects on different facets of image generation. Our results show that all components show positive effects across all metrics, at the expense of a slight degradation in FID. Compared to the CFG baseline (first row), most of the improvement in CLIPScore and VQA are brought by the conditional entropy rectification (τ_c, second row), while the improvements in Coverage and Density mostly come from rectifying the attention in the denoiser to obtain (τ_i, third row). Finally, a further modest improvement in Density, Coverage and VQA is brought by the update step size (γ, fourth row).

Multi-step gradient descent. In Table 5 we consider the effect of varying number of updates K for each attention operation along with the update step size. We find small variations in metrics with respect to the baseline $K = \gamma = 1$ with slight improvements when setting $\gamma = 1.5$, using multiple gradient descent steps did not induce significant gains. Therefore, we used $K = \gamma = 1$ in our default setup in our experiments, unless specified otherwise.

5 Conclusion

We presented *Entropy Rectifying Guidance (ERG)*, a novel guidance mechanism for sampling diffusion and flow models which significantly improves sample quality without sacrificing diversity and consistency performance. In particular, by manipulating the energy landscape of the attention layers in the diffusion transformer and the text encoder at inference time, ERG significantly boosts the performance of different models when studying their quality-consistency-diversity trade-offs and is applicable to different modalities such as text-to-image, class conditional and unconditional models. Furthermore, ERG outperforms recent state-of-the-art guidance methods such as CADS, APG, SEG and AutoGuidance, while requiring the same compute as standard CFG. ERG can be combined with approaches such as APG and CADS, which further improves results.

Acknowledgements. Karteek Alahari was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

References

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, 2024.

Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism Pareto fronts of conditional image generative models. *arXiv preprint*, 2406.10429, 2024.

Tariq Berrada, Pietro Astolfi, Melissa Hall, Reyhane Askari Hemmat, Yohann Benchetrit, Marton Havasi, Matthew J. Muckley, Karteek Alahari, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. On improved conditioning mechanisms and pre-training strategies for diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint*, 2407.07726, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint*, 2406.08070, 2024a.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, (25), 2024b.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint*, 2407.21783, 2024.

Melissa Hall, Oscar Mañas, Reyhane Askari, Mark Ibrahim, Candace Ross, Pietro Astolfi, Tariq Berrada Ifriqi, Marton Havasi, Yohann Benchetrit, Karen Ullrich, Carolina Braga, Abhishek Charnalia, Maeve Ryan, Mike Rabbat, Michal Drozdzal, Jakob Verbeek, and Adriana Romero Soriano. Evalgim: A library for evaluating generative image models. *arXiv preprint*, 2412.10604, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems*, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *International Conference on Computer Vision*, 2023.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, 2024.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Mark Levy, Bruno Di Giorgi, Floris Weers, Angelos Katharopoulos, and Tom Nickson. Controllable music production with diffusion models and guidance gradients. In *Advances in Neural Information Processing Systems*, 2023.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint*, 2404.01291, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Machine Learning*, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for Bayesian context update in text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. Understanding classifier-free guidance: High-dimensional theory and non-linear generalizations. *arXiv preprint*, 2502.07849, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *arXiv preprint*, 2410.13720, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlovic, Geir Kjetil Sandve, Victor Greiff, David P. Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *International Conference on Learning Representations*, 2025.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and sheng zhao. AUDIT: Audio editing by following instructions with latent diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Conference on Computer Vision and Pattern Recognition*, 2024.

A. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.

Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric VQGAN for StableDiffusion. *arXiv preprint*, 2306.04632, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: we claim: (i) a novel guidance technique to sample diffusion/flow models, (ii) our guidance technique is applicable to unconditional sampling, and (iii) improved sample quality/diversity/consistency compared to previous methods. The claims are clearly stated in the abstract and introduction. Novelty is backup-up by a clear discussion of relevant related work, and performance claims are backed up by extensive experimental results on multiple data sets and models.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: In section B of the appendix, we discuss the limitations of our method in detail.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: Our paper does not introduce theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our experiments are accompanied with the full set of hyperparameters used to reproduce each experiment. Namely Section 4.1 and Appendix E discuss the choice of hyperparameters in detail. We also provide pseudo-code for our method in Algorithm 1 and Algorithm 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our paper provides pseudo-code and hyperparameter choices for popular open-source models such as Stable Diffusion 3, we evaluate our method on COCO dataset which is open-source as well. Some models were trained on public data (ImageNet) others on a mix of public and proprietary data (YFCC + 320M proprietary text-image pairs).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 provides a comprehensive context for the setup in which experiments were conducted. Our main results in Section 4.2 are accompanied with extensive ablations in Section 4.3 where each experiment is accompanied with a thorough discussion of the setup of the experiment and an analysis of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: While we do not provide statistical significance tests, we evaluate our method in different settings (class-conditional, text-to-image), different models (internal models of different scales, open-source models), and also models trained for different resolutions (256 and 512), across all these setups, we find our method to result in significant improvements. Accordingly, we believe that the significance of our method is validated by the extent of our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our method consists in a modified sampling algorithm for diffusion and flow models, the new sampling method has a comparable NFEs to standard classifier-free guidance, which we report in Table 1. The main difference in compute compared would be induced by the number of gradient descent steps K in the multistep energy update, this variable is set to $K = 1$ in most of our experiments and recommended setups, hence the computational cost is practically unchanged compared to standard classifier-free guidance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our internal model was trained on mitigated datasets (CSAM filtering + face blurring), we only used compliant datasets for evaluation (MS COCO), which is

under Creative Commons Attribution 4.0 License. Our implementation on top of open-source models does not alter the model weights, but just provides changes to the guidance mechanism used for inference.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We propose a method to improve the quality of generative models, hence our method can be used to generate high-quality, realistic content which can be used maliciously to spread misinformation and disinformation, creating deepfakes, identity theft etc. By releasing our code transparently, we provide a way for researchers to study and counter the potential harmful effects of our method being misused, allowing for the development of defense strategies.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any new models or datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We carefully cite the respective authors of each method/model and dataset that was referenced in our manuscript. We additionally provide links to the different assets used in Table 14.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: We do not introduce new datasets/code/models with our paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not perform any crowdsourcing or use human subjects in our work.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not perform any crowdsourcing or use human subjects in our work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our work does not make use of LLMs as an important, original or non-standard component of the method we propose.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.