One-step Diffusion Models with Bregman Density Ratio Matching

Yuanzhi Zhu^{*1} **Eleftherios Tsonis**^{*1} **Lucas Degeorge**^{*1,2,3} **Vicky Kalogeiton**^{1}
[1]LIX, École Polytechnique, CNRS, IPP [2]LIGM, École Nationale des Ponts et Chaussées, CNRS, IPP [3]AMIAD

Abstract

Diffusion and flow models achieve high generative quality but remain computationally expensive due to slow multi-step sampling. Distillation methods accelerate them by training fast student generators, yet most existing objectives lack a unified theoretical foundation. In this work, we propose Di-Bregman, a compact framework that formulates diffusion distillation as Bregman divergence-based density-ratio matching. This convex-analytic view connects several existing objectives through a common lens. Experiments on CIFAR-10 and text-to-image generation demonstrate that Di-Bregman achieves improved one-step FID over reverse-KL distillation and maintains high visual fidelity compared to the teacher model. Our results highlight Bregman density-ratio matching as a practical and theoretically-grounded route toward efficient one-step diffusion generation.

1 Introduction

Diffusion and flow models [1, 15, 21, 24, 41, 43, 44] have become a cornerstone of generative modeling, attaining state-of-the-art performance across modalities and tasks [5, 7, 9, 10, 10, 25, 34, 36, 47]. Yet their sampling process remains prohibitively slow, often requiring hundreds of network evaluations per sample. This has motivated an active line of research on distillation: training fast student generators that reproduce a pre-trained teacher's output in one or few steps. Current approaches can be broadly categorized as Ordinary Differential Equation (ODE)-based [12, 22, 38, 42], which learn consistency mappings along the teacher's probability-flow ODE, and distribution-based [27, 55, 60], which directly match the generator's output distribution to that of the teacher or data. ODE-based methods enforce sufficient but unnecessary conditions for one-step generation, whereas distribution-based methods relax these constraints and capture a broader solution space. Variational Score Distillation (VSD) [49] and Distribution Matching Distillation (DMD) [55] define objectives based on reverse-Kullback-Leibler (KL) divergence between student and teacher models. f-distill [52] reframed these methods through the lens of f-divergences. Despite this progress, a general perceptive that explains these objectives in a simple mathematical form remains missing.

We introduce Di-Bregman, a general framework that formulates diffusion distillation as Bregman divergence-based density-ratio matching. The central insight is that aligning the student distribution $q(x)$ with the teacher $p(x)$ can be viewed as driving the ratio $r(x) = \frac{q(x)}{p(x)}$ toward constant one, under a suitable convex function h. This perspective yields a closed-form gradient (Theorem 1) with weighting $h''(r)r$. Under this formulation, familiar objectives, such as KL- or MSE-based distillation arise as specific choices of h. The result is a concise, interpretable expression that connects multiple existing formulations within a single theoretical framework.

Beyond theory, Di-Bregman remains practical. To get the weightning coefficient $h''(r)r$, we estimate density ratios through a simple classifier trained to distinguish student samples from real data, enabling efficient training without repeated teacher simulation and allowing optional adversarial

*share the same office

EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM).

refinement. Preliminary results on both unconditional image and text-to-image generation demonstrate that our approach attains improved one-step FID than reverse-KL distillation and maintains visual fidelity comparable to the multi-step teacher models.

In summary, our contributions are:

- We introduce a unified formulation of diffusion distillation based on Bregman density-ratio matching, which yields a closed-form gradient interpretation,
- We propose a practical classifier-based training procedure that effectively instantiates this formulation and validate it on early benchmarks.

2 Preliminaries

2.1 Variational Score Distillation

Variational Score Distillation (VSD) [49] was introduced to mitigate mode-seeking and over-saturation [2] issues observed when using Score Distillation Sampling (SDS) for 3D asset generation [35]. Importantly, the VSD objective is defined on the *final* samples produced by a generator, rather than on intermediate sampler states. This final-sample focus naturally motivates efforts to distil powerful multi-step pre-trained model into compact few-step or one-step generators via VSD-style objectives; several recent works have followed this route [27, 31, 55].

Concretely, VSD can be viewed as minimizing a time-averaged divergence between the noisy marginal produced by the student generator and the corresponding noisy marginal of a pretrained reference model. Writing q_t and p_t for the generator and reference noisy marginals at time t, respectively, the gradient of a typical score-distillation loss admits the following approximation:

$$\nabla_\theta \mathcal{L}_{\text{VSD}} = \mathbb{E}_t \left[\nabla_\theta \text{KL}(q_t \parallel p_t) \right] \approx -\mathbb{E}_{t,\epsilon} \left[w(t) \big(s_\phi(x_t, t) - s_\psi(x_t, t) \big) \frac{\mathrm{d}G_\theta(\epsilon)}{\mathrm{d}\theta} \right], \tag{1}$$

where $w(t)$ is a scalar weighting function over timesteps, and the noisy state x_t is obtained by applying the forward diffusion kernel at time t to the generator output $G_\theta(\epsilon)$ using another independent Gaussian noise. $s_\phi(\cdot, t)$ and $s_\psi(\cdot, t)$ denote the pre-trained score function on reference data and the auxiliary score function on the student-generated data evaluated at timestep t, respectively. Intuitively, the score difference $s_\phi - s_\psi$ provides a learning signal that pushes the student's generated noisy marginals toward those of the pre-trained teacher model, and backpropagating through G_θ to update the generator parameters θ.

2.2 Bregman Divergence for Density Ratio Matching

Given two probability distributions $p^*(x)$ and $q^*(x)$, the goal of *density ratio matching* is to learn a ratio model $r_\theta(x)$ that approximates the true density ratio $r^*(x) := \frac{q^*(x)}{p^*(x)}$ based on i.i.d. samples drawn from both distributions.

The *Bregman divergence* provides a flexible and theoretically grounded measure for comparing functions such as density ratios. It generalizes the notion of squared Euclidean distance to a broad class of divergences that share similar geometric and convexity properties [2, 46]. Let h be a differentiable and strictly convex function. The Bregman divergence associated with h between two functions r and r^* is defined as [18, 46]:

$$D_h(r \| r^*) = \int p(x) \Big[h(r(x)) - h(r^*(x)) - h'(r^*(x))(r(x) - r^*(x)) \Big] \, \mathrm{d}x. \tag{2}$$

This divergence is positive-definite, which means it is always non-negative and equals zero if and only if $r(x) = r^*(x)$ almost everywhere with respect to $p(x)$, which is the density implicitly defined in $r(x)$. Many well-known divergences arise as special cases of the Bregman divergence for specific choices of the convex function h. For instance, the *squared loss* corresponds to $h(r) = \frac{1}{2}r^2$, leading

[2] The SDS objective tends to produce solutions corresponding to the mode of the averaged likelihood, leading to mode-seeking behavior. Moreover, a high Classifier Free Guidance (CFG) scale can cause over-saturated and over-smoothed generation results.

to least-squares density ratio estimation [45] and the *KL divergence* corresponds to $h(r) = r\log r - r$. More instances can be found in Sec. 2.2. This unifying framework allows density ratio estimation to be interpreted as minimizing a Bregman divergence under different convex function h, providing a general connection between statistical divergences and convex analysis [2, 6].

Table 1: Examples of different $h(r)$ in Bregman divergence and the corresponding $h''(r)r$. The choices of $h(r)$ are from [18, 32].

Name	$h(r)$	$h''(r)r$	$h''(e^{-l})e^{-l}$
LR	$r\log r - (1+r)\log(1+r)$	$\frac{1}{1+r}$	$\sigma(l)$
KL	$r\log r - r$	1	1
BE	$-\log r$	$1/r$	e^{l}
LS	$r^2/2$	r	e^{-l}
SBA	$\frac{r^{1+\lambda}-r}{\lambda(\lambda+1)}$	r^{λ}	$e^{-\lambda l}$

3 Method

In this section, we introduce a general distillation framework, termed Di-Bregman, which is derived from the Bregman divergence for density ratio matching formulation in Sec. 2.2. The core idea is to align the student distribution $q(x)$, induced by a one-step generator G_θ, with the teacher distribution $p(x)$. Since the student distribution $q(x)$ is implicitly defined by the generator through the push-forward measure of the prior, i.e., $x = G_\theta(\epsilon)$ with $\epsilon \sim \mathcal{N}(0, I)$, the distribution $q(x)$ and its density ratio depend on the generator parameters θ. Let $r(x) = \frac{q(x)}{p(x)}$ denote the density ratio between the student and teacher distributions. Perfect alignment hence corresponds to $r(x) = 1$ for all x, which motivates minimizing a divergence between $r(x)$ and the target ratio 1 in Eq. (2):

$$D_h(r\|1) = \int p(x)\Big[h(r(x)) - h(1) - h'(1)(r(x)-1)\Big]\,\mathrm{d}x. \tag{3}$$

Minimizing this divergence with respect to θ encourages the student generator G_θ to produce samples whose induced distribution $q(x)$ matches the teacher distribution $p(x)$.

Following prior work on one-step and few-step distillation of diffusion models [27, 31, 55], we can derive the analytical form of the gradient of the Bregman divergence in Eq. (3). The resulting expression corresponds to a weighted variant of the gradient used in the KL-based objective (Eq. (1)), where the weight is a function of the density ratio $r(x)$, analogous to the formulation in f-distill [52].

To further generalize this result as in VSD, we consider the intermediate distributions p_t and q_t obtained via the diffusion forward process. This allows the Bregman-based distillation gradient to be evaluated at arbitrary diffusion timesteps. The following theorem formally characterizes the gradient of the Bregman divergence in this general setting.

Theorem 1 (Gradient of Bregman divergence)**.** *Let p_t be a reference (teacher) marginal density at time t and let $q_t = q_{\theta,t}$ be the marginal induced by the generator G_θ at time t. These intermediate densities are obtained via the forward diffusion process. Define the intermediate density ratio $r_t(x) := \frac{q_{\theta,t}(x)}{p_t(x)}$. Assume that h is twice continuously differentiable. Then the gradient of the Bregman divergence $D_h(r_t\|1) = \mathbb{E}_{p_t}[h(r_t)] - h(1)$ with respect to θ admits the following form:*

$$\nabla_\theta D_h(r_t\|1) = -\mathbb{E}_\epsilon\Big[w(t)h''(r_t(x_t))\,r_t(x_t)\,\big(\nabla_{x_t}\log p_t(x_t) - \nabla_x \log q_{\theta,t}(x_t)\big)\,\nabla_\theta G_\theta(\epsilon)\Big], \tag{4}$$

where $w(t)$ is a weight function.

The corresponding proof can be found in Appendix B.

In practice, the density ratio on noisy data, $r_t(x) = \frac{q_t(x)}{p_t(x)}$, can be estimated using a classifier trained to distinguish samples from the student generator G_θ and those from the teacher model or reference dataset. Under the common assumption that the pre-trained teacher model already captures the data distribution well, it is often both preferable and computationally cheaper to draw real samples

Figure 2: Images generated with only one-step by model trained with Di-Bregman. More images are shown in Appendix E

directly from the dataset rather than repeatedly sampling from the teacher. The discriminator loss is hence:

$$\min_{G_\theta} \max_{D_\eta} \mathbb{E}_{x_{\mathrm{gt}} \sim p_{\mathrm{data}}, t \sim p_{t_{\mathrm{GAN}}}} [\log D_\eta((x_{\mathrm{gt}})_t, t)] + \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t \sim p_{t_{\mathrm{GAN}}}} [\log(1 - D_\eta(G_\theta(\epsilon)_t, t))]. \tag{5}$$

For a discriminator output $D_\eta = \sigma(l_t(x))$, where $l_t(x)$ denotes the classifier logits at noise level t, the optimal output satisfies $\sigma(l_t^*(x)) = \frac{p_t(x)}{p_t(x)+q_t(x)}$. This implies that the density ratio can be recovered as $r_t(x) = e^{-l_t(x)}$. We provide common used $h(r)$ and corresponding $h''(e^{-l})e^{-l}$ in Sec. 2.2. In this framework, the trained classifier not only provides an estimate of the local density ratio but can also be repurposed as a discriminator for adversarially training the student generator.

Compared to f-distill [52], our framework places fewer constraints on the convex function, which yields greater flexibility in choosing divergence families for distillation. Recently, Uni-Instruct [48] proposes a unifying view that connects integral f-divergences [52, 55] and score-based divergences [28, 60]. However, Di-Bregman is complementary to this line of work: it provides a Bregman-divergence perspective that admits a broader class of function h and recovers many existing objectives as special cases. Together, these formulations offer a more complete picture of distribution-based diffusion distillation.

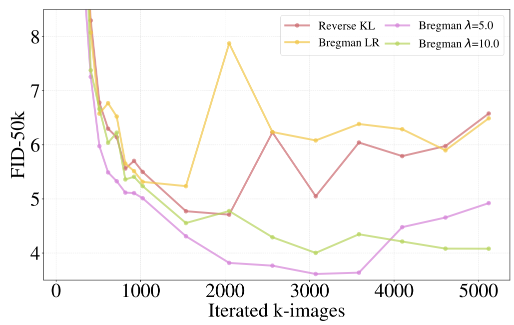

Figure 1: **Evolution of one-step FID against number of iterated images**: Di-Bregman achieves a lower one-step FID.

4 Experiments

To evaluate the effectiveness of the proposed method, we conduct experiments on both unconditional image and text-to-image generation tasks. Quantitative results are reported on the CIFAR-10 dataset using an EDM teacher [16], while qualitative results are presented for text-to-image generation with a Stable Diffusion v1.5 [36] teacher. As shown in Fig. 1, when applying the SBA-type Bregman divergence with $\lambda = 5$, our method achieves a lower one-step Fréchet Inception Distance (FID) compared to the baseline reverse KL distillation approach. In addition, Fig. 2 illustrates representative one-step samples generated by our distilled text-to-image model, demonstrating high visual quality and fidelity to the text prompts. More experimental details (D), qualitative (E), quantitative (F) results and additional ablations (F) are provided in Appendix.

5 Conclusion

We introduced Di-Bregman, a generalized framework for diffusion model distillation grounded in Bregman divergences. Empirically, our method improves one-step generation quality on CIFAR-10 and produces competitive visual results in text-to-image synthesis, demonstrating both its theoretical generality and practical effectiveness.

Algorithm 1 Di-Bregman Distillation

Require: Pre-trained teacher model ϕ, auxiliary model ψ, discriminator heads η, condition dataset $\mathcal{D}_c$, ground truth dataset $\mathcal{D}_d$, loss weights w_{GAN} (optional)

$\theta \leftarrow \text{copyWeights}(\phi)$, $\psi \leftarrow \text{copyWeights}(\phi)$ // initialize models
repeat
    ### Generate one-step image samples x_θ
    Sample $\epsilon \sim \mathcal{N}(0,1)$, $c \sim \mathcal{D}_c$
    $x_\theta = G_\theta(\epsilon, c)$
    ### Update generator θ
    Sample $t \sim \mathcal{U}[0,1]$, $x_t \sim q_{t|0}(x_t|x_\theta)$ // Forward
    Calculate true and auxiliary score $s_\phi(x_t, c)$ and $s_\psi(x_t, c)$
    Calculate the density ratio r_t using discriminator head logit output $r_t(x) = e^{-l_t(x)}$
    `# calculate Di-Bregman loss gradient`
    $\nabla_\theta \mathcal{L}_{\text{Di-Bregman}}(\theta) \longleftarrow -\mathbb{E}_\epsilon\Big[h''(r_t(x_t))\, r_t(x_t)\, \big(s_\phi(x_t, c) - s_\psi(x_t, c)\big)\, \nabla_\theta G_\theta(\epsilon)\Big].$
    `# calculate GAN loss (optional)`
    $\mathcal{L}_{\text{GAN}}(\theta) \longleftarrow \mathbb{E}_{\epsilon,t}[-\log(D_\eta(x_t, t))].$
    `# calculate total loss and update`
    Update θ using gradient of $\mathcal{L}_{\text{gen}}(\theta) = \mathcal{L}_{\text{Di-Bregman}}(\theta) + w_{\text{GAN}}\mathcal{L}_{\text{GAN}}(\theta)$
    ### Update auxiliary model ψ
    Sample $t' \sim \mathcal{U}[0,1]$, $x_{t'} \sim q_{t|0}(x_{t'}|x_\theta(x_{\text{init}}, c))$
    Update ψ with standard denoising score match loss to leard x_θ
    ### Update discriminator η
    Sample $t'' \sim \mathcal{U}[0, 0.95]$, calculate $x_{t''} \sim q_{t|0}(x_{t''}|x_\theta)$
    Sample real data $(x_{\text{gt}}, c) \sim \mathcal{D}_d$, calculate $(x_{\text{gt}})_{t''}$
    Update η with GAN objective (Eq. (5))
until *convergence*
Return one-step generator θ

References

[1] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[3] Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *arXiv preprint arXiv:2505.18825*, 2025.

[4] Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025.

[5] Luc Boudier, Loris Manganelli, Eleftherios Tsonis, Nicolas Dufour, and Vicky Kalogeiton. Training-free synthetic data generation with dual ip-adapter guidance. In *British Machine Vision Conference (BMVC)*, 2025.

[6] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[7] Robin Courant, Xi Wang, David Loiseaux, Marc Christie, and Vicky Kalogeiton. Pulp motion: Framing-aware multimodal camera and human motion generation. *arXiv preprint arXiv:2510.05097*, 2025.

[8] Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *European Conference on Computer Vision*, pages 176–192. Springer, 2025.

[9] Lucas Degeorge, Arijit Ghosh, Nicolas Dufour, David Picard, and Vicky Kalogeiton. How far can we go with imagenet for text-to-image generation? *arXiv*, 2025.

[10] Nicolas Dufour, Victor Besnier, Vicky Kalogeiton, and David Picard. Don't drop your samples! coherence-aware training benefits conditional diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6264–6273, 2024.

[11] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

[12] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

[13] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.

[14] 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. *Advances in neural information processing systems*, 30, 2017.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[16] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[17] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

[18] Yeongmin Kim, Heesun Bae, Byeonghu Na, and Il-Chul Moon. Preference optimization by estimating the ratio of the data distribution. *arXiv preprint arXiv:2505.19601*, 2025.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[22] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[23] Qiang Liu. Icml tutorial on the blessing of flow. *International conference on machine learning*, 2025.

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[25] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

[26] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

[27] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023.

[28] Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024.

[29] Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2025.

[30] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[31] Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7807–7816, 2024.

[32] Frank Nielsen and Richard Nock. Sided and symmetrized bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

[33] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[35] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[37] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[38] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[39] Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. *Advances in Neural Information Processing Systems*, 37:36046–36070, 2025.

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *neurips*, 2022.

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[42] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] 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*, 2020.

[45] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 20, 2008.

[46] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

[47] Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2609–2619, 2025.

[48] Yifei Wang, Weimin Bai, Colin Zhang, Debing Zhang, Weijian Luo, and He Sun. Uni-instruct: One-step diffusion model through unified diffusion divergence instruction. *arXiv preprint arXiv:2505.20755*, 2025.

[49] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

[50] Sirui Xie, Zhisheng Xiao, Diederik P Kingma, Tingbo Hou, Ying Nian Wu, Kevin Patrick Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *arXiv preprint arXiv:2405.16852*, 2024.

[51] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024.

[52] Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

[53] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510*, 2024.

[54] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

[55] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024.

[56] Haoyang Zheng, Xinyang Liu, Cindy Xiangrui Kong, Nan Jiang, Zheyuan Hu, Weijian Luo, Wei Deng, and Guang Lin. Ultra-fast language generation via discrete diffusion divergence instruct. *arXiv preprint arXiv:2509.25035*, 2025.

[57] Mingyuan Zhou, Yi Gu, and Zhendong Wang. Few-step diffusion via score identity distillation. *arXiv preprint arXiv:2505.12674*, 2025.

[58] Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Long and short guidance in score identity distillation for one-step text-to-image generation. *arXiv preprint arXiv:2406.01561*, 2024.

[59] Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. *arXiv preprint arXiv:2410.14919*, 2024.

[60] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024.

[61] Yuanzhi Zhu, Xingchao Liu, and Qiang Liu. Slimflow: Training smaller one-step diffusion models with rectified flow. In *European Conference on Computer Vision*, pages 342–359. Springer, 2025.

[62] Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Di[M]o: Distilling masked diffusion models into one-step generator. *arXiv preprint arXiv:2503.15457*, 2025.

[63] Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Soft-di [m] o: Improving one-step discrete image generation with soft embeddings. *arXiv preprint arXiv:2509.22925*, 2025.

A Limitations and Future Works.

This work primarily presents preliminary results. In future studies, we plan to extend our approach to a wider range of teacher models and conduct comprehensive comparisons with state-of-the-art methods. Moreover, while our current experiment only use the classifier in Eq. (4), we aim to incorporate adversarial training based on it to further enhance the performance of one-step generation.

B Derivation

Theorem 1 (Gradient of Bregman divergence)**.** *Let p_t be a reference (teacher) marginal density at time t and let $q_t = q_{\theta,t}$ be the marginal induced by the generator G_θ at time t. These intermediate densities are obtained via the forward diffusion process. Define the intermediate density ratio $r_t(x) := \frac{q_{\theta,t}(x)}{p_t(x)}$. Assume that h is twice continuously differentiable. Then the gradient of the Bregman divergence $D_h(r_t\|1) = \mathbb{E}_{p_t}[h(r_t)] - h(1)$ with respect to θ admits the following form:*

$$\nabla_\theta D_h(r_t\|1) = -\mathbb{E}_\epsilon\Big[w(t)h''(r_t(x_t))\, r_t(x_t)\, \big(\nabla_{x_t}\log p_t(x_t) - \nabla_x \log q_{\theta,t}(x_t)\big)\, \nabla_\theta G_\theta(\epsilon)\Big]. \quad (6)$$

where $w(t)$ is a weight function.

Proof. Recall that

$$D_h(r_t\|1) = \int p_t(x)\Big[h(r_t(x)) - h(1) - h'(1)\big(r_t(x) - 1\big)\Big]\mathrm{d}x,$$

where $r_t(x) = q_{\theta,t}(x)/p_t(x)$ and p_t do not depend on θ. Differentiating under the integral sign and using that

$$\int p_t(x)\, \nabla_\theta r_t(x)\, \mathrm{d}x = \nabla_\theta \int q_{\theta,t}(x)\, \mathrm{d}x = \nabla_\theta 1 = 0,$$

we obtain

$$\nabla_\theta D_h(r_t\|1) = \int p_t(x)\, h'(r_t(x))\, \nabla_\theta r_t(x)\, \mathrm{d}x = \int h'(r_t(x))\, \nabla_\theta q_{\theta,t}(x)\, \mathrm{d}x. \quad (7)$$

Next, we express $q_{\theta,t}$ as the pushforward of a base noise $\epsilon \sim p(\epsilon)$ through the generator at time t, $x_t = F(G_\theta(\epsilon), z)$ with fixed forward process $F(x, z) = \alpha_t x + \sigma z$ and $z \sim \mathcal{N}(0, I)$:

$$q_{\theta,t}(x) = \int p(\epsilon)\, \delta\big(x - x_t\big)\, \mathrm{d}\epsilon, \quad (8)$$

where δ is the Dirac delta. Differentiating equation 8 w.r.t. θ and using the chain rule for distributions yields

$$\nabla_\theta q_{\theta,t}(x) = \int p(\epsilon)\, \nabla_\theta \delta\big(x - x_t\big)\, \mathrm{d}\epsilon = -w(t) \int p(\epsilon)\, \nabla_x \delta\big(x - x_t\big)\, \nabla_\theta G_\theta(\epsilon)\, \mathrm{d}\epsilon. \quad (9)$$

Substituting equation 9 into equation 7 gives

$$\nabla_\theta D_h(r_t\|1) = -\int w(t) h'(r_t(x))\, \left[\int p(\epsilon)\, \nabla_x \delta\big(x - x_t\big)\, \nabla_\theta G_\theta(\epsilon)\, \mathrm{d}\epsilon\right] \mathrm{d}x \quad (10)$$

$$= -\int p(\epsilon)\, \left[w(t) \int h'(r_t(x))\, \nabla_x \delta\big(x - x_t\big)\, \mathrm{d}x\right] \nabla_\theta G_\theta(\epsilon)\, \mathrm{d}\epsilon. \quad (11)$$

We now integrate by parts in x (assuming boundary terms vanish):

$$\int h'(r_t(x))\, \nabla_x \delta\big(x - x_t\big)\, \mathrm{d}x = -\int \delta\big(x - x_t\big)\, \nabla_x h'(r_t(x))\, \mathrm{d}x.$$

Hence

$$\begin{aligned} \nabla_\theta D_h(r_t\|1) &= \int p(\epsilon)\, \left[\int w(t)\delta\big(x - x_t\big)\, \nabla_x h'(r_t(x))\, \mathrm{d}x\right] \nabla_\theta G_\theta(\epsilon)\, \mathrm{d}\epsilon \\ &= \mathbb{E}_\epsilon\big[w(t) \nabla_x h'(r_t(x_t))\, \nabla_\theta G_\theta(\epsilon)\big]. \end{aligned} \quad (12)$$

Apply the chain rule $\nabla_x h'(r_t) = h''(r_t)\nabla_x r_t$ and $\nabla_x r_t = r_t(\nabla_x \log q_{\theta,t} - \nabla_x \log p_t)$ yields Eq. (4) stated in the theorem. $\square$

C Related Works: Diffusion Distillation

Distillation methods for accelerating diffusion and flow models fall into two broad families. *ODE-based* distillation exploits the teacher's Probability Flow ODE (PF-ODE) to derive regression-style objectives for a student model [3, 4, 11, 12, 13, 17, 22, 26, 30, 38, 42, 53, 61]. These approaches frame distillation as learning an ODE-consistent mapping, often enabling stable one- or few-step samplers which preserves the coupling induced by teacher models' PF-ODE. By contrast, *distribution-based* methods align the student generator's output distribution with the teacher's multi-step sampling distribution (or with a specified data distribution) without relying on an explicit PF-ODE. This class covers divergence- and adversarial-style matching techniques [8, 27, 29, 31, 39, 48, 49, 50, 51, 54, 55, 56, 57, 58, 59, 60, 62, 63]. Compared to distribution-based methods, ODE-based formulations optimize more indirect objectives that enforce consistency with an underlying continuous-time dynamics. These ODE constraints are *sufficient but not necessary* for correct one-step generation. Consequently, ODE-based methods are more restrictive, while distribution-based formulations directly match the target distribution and thus allow a broader family of solutions and greater modeling flexibility. In f-distil, [52] extend the VSD framework from reverse Kullback–Leibler (KL) divergence to more general f-divergence and use discriminator to estimate the density ratio. A notable feature of many distribution-based methods is that they match not only the final data distribution but also the intermediate noisy-data distributions encountered during sampling; this property has also been referred to as *Interpolation Distillation* [23].

D Experimental Setup

Datasets and Pre-trained Teacher Models. Our experiments to demonstrate the effectiveness of Di-Bregman are performed on the CIFAR-10 [19] 32×32 for unconditional generation and on the LAION [40] and COCO [20] datasets for text-to-image generation. The pre-trained teacher models are adopted from the official checkpoints from previous works, EDM [16], and Stable Diffusion v1.5 [36].

Implementation Details. All experiments are conducted on a single NVIDIA H100 GPU. For CIFAR-10 (32×32) experiments, we adopt the U-Net architecture of NCSN++ [44]. The implementation is based on the SiDA framework [59], where the discriminator is built upon the auxiliary model encoder, and the mean feature vector is used as the predicted discriminator logits. For the text-to-image experiment, we use Stable-Diffusion v1.5 [36], a 900M-paramenter U-net-based model, trained on LAION [40] and distilled at 512×512 resolution. All results presented in the paper are one-step generated using our distilled generator.

Evaluation Metrics. The metrics we use for quantitative results on CIFAR-10 are Fréchet Inception Distance (FID) [14] and Inception Score (IS) [37]. In our experiments, FID is computed with 50,000 generated samples compared against the training set using Clean-FID [33], while IS is calculated from the same generated images based on their Inception features.

E Additional qualitative results

In Fig. 3, we provide additional qualitative comparisons, where our one-step student produces visually coherent and faithful samples, closely matching the teacher output across diverse prompts. Additional uncurated CIFAR-10 samples from our Di-Bregman model are shown in Fig. 6, demonstrating diverse one-step generation, with an FID of 3.61.

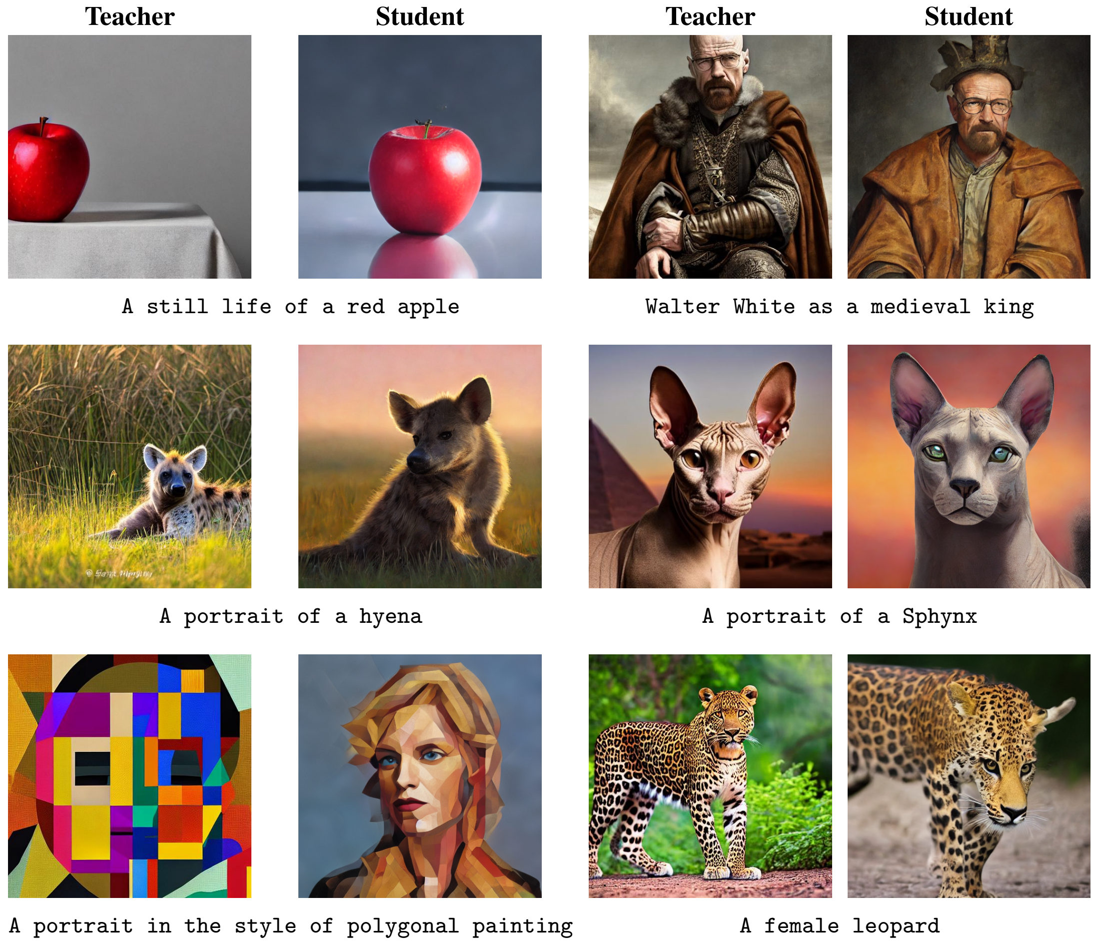

Figure 3: **Qualitative comparison at 512×512: Teacher 50 NFEs (first, third columns) vs Student 1 NFE (second, last columns) for six prompts (left and right blocks per row).** The teacher is the Stable Diffusion v1.5 [36] model.

F Additional quantitative results

In this section, we provide some additional quantitative results from our one-step models.

Figures 4 and 5 show the evolution of one-step FID and IS, respectively, across different values of the Bregman parameter λ. We observe that Di-Bregmanconsistently improves over the reverse KL baseline for several λ configurations. In particular, settings such as $\lambda = 3.0$, $\lambda = 5.0$, and $\lambda = 10.0$ yield the lowest one-step FID (Fig. 4) and the highest one-step IS (Fig. 5), confirming the robustness of Di-Bregmanacross a range of divergence parameters. Lower λ values (e.g., $\lambda \leq 1.0$) tend to perform closer to the baseline, while negative $\lambda = -1.0$ underperforms. These results demonstrate that Di-Bregmanoffers consistent improvements in sample quality metrics over the reverse KL distillation method.

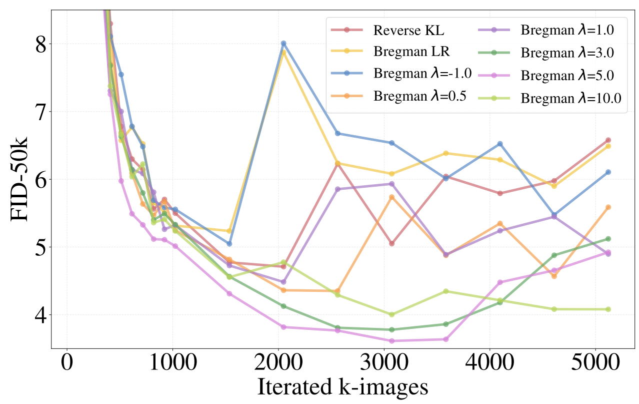

Figure 4: **Evolution of one-step FID against number of iterated images**. For $\lambda = 3.0, \lambda = 5.0$ or $\lambda = 10.0$ Di-Bregman achieves a lower one-step FID.

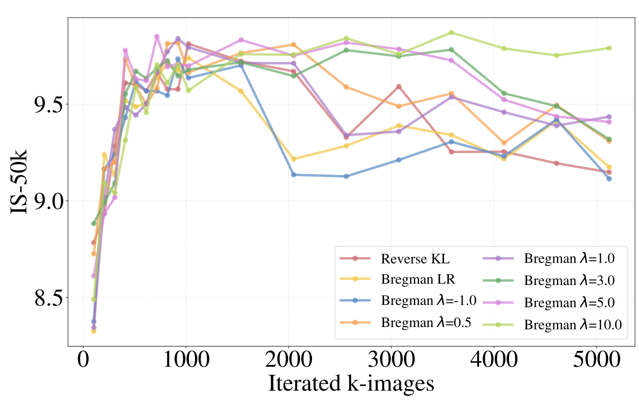

Figure 5: **Evolution of one-step IS against number of iterated images**. For $\lambda = 3.0, \lambda = 5.0$ or $\lambda = 10.0$ Di-Bregman achieves a higher one-step IS.

Figure 6: Uncurated samples from unconditional CIFAR-10 32×32 using Di-Bregman with single step generation (FID=3.61).

G General Divergence with Fixed Target Ratio r_t^*:

$$\nabla_\theta D_h(r_t \| r_t^*) = \int p_t(x)\, h'(r_t(x))\, \nabla_\theta r_t(x)\, \mathrm{d}x - \int p_t(x) h'(r_t^*)\, \nabla_\theta r_t(x)\, \mathrm{d}x. \tag{13}$$

The second term does not equal to 0 because $h'(r_t^*)$ is not a constant anymore.

The general final gradient is hence:

$$\nabla_\theta D_h(r_t \| r_t^*) = -\mathbb{E}_\epsilon \Big[h''(r_t(x_t))\, r_t(x_t) \left(\nabla_{x_t} \log p_t(x_t) - \nabla_x \log q_{\theta,t}(x_t) \right) \nabla_\theta G_\theta(\epsilon) + h''(r_t^*) \nabla_\theta r_t^*(x_t) \Big]. \tag{14}$$

We can interpret the target density ratio r^* as a normalized reward function, hence the optimal distribution satisfies $q^* = pr^*$ and we can use Eq. (13) to direct distill a reward tilted distribution.

In practice, employing an unnormalized reward $R(x_t)$ entails estimating its normalization constant, $\mathbb{E}_{x_t \sim p_t} R(x_t)$, which is typically intractable for arbitrary reward functions.