Accelerating Diffusion Models for Discriminative Vision and Language Learners

Anonymous ACL submission

Abstract

Text-to-image diffusion models have demonstrated impressive generative capabilities, indicating they internalize substantive image-text representations. While these models have shown promise results, their potential in downstream discriminative applications is largely uncharted. In this paper, we delve into the capabilities of these diffusion models and improve the efficiency of using them as zero-shot vision and language learners. Towards this, we introduce a novel hierarchical sampling strategy that significantly optimizes the computational demands of these zero-shot diffusion models, making them faster and more feasible for real-world applications. Our work showcases the potential of text-to-image diffusion models as powerful tools for zero-shot image-text matching and sets the stage for more practical and effective applications of these models in real-world settings.

1 Introduction

Advances in large-scale machine learning models have allowed them to be trained on extensive internet-scale datasets and applied as zero-shot learners, removing the need for task-specific training. These models, exemplified by work such as Radford et al. (2021); Ilharco et al. (2021); Li et al. (2023b), can now handle a wide range of tasks without additional fine-tuning.

Among these advances, another line of work such as generative text-to-image models built on denoising diffusion probabilistic techniques, including Imagen (Saharia et al., 2022), Dalle-2 (Ramesh et al., 2022), and Stable Diffusion (Rombach et al., 2022; Podell et al., 2023), has attracted significant attention. They can produce realistic, high-resolution images from diverse text prompts, suggesting that they have learned useful representations of image-text data.

Despite this progress, their application to discriminative tasks remains underexplored, and their performance relative to other pre-trained models is not well understood. Some recent work (Li et al., 2023a) has investigated Stable Diffusion as a generative classifier using a re-weighted variant of its variational lower bound. However, this classification process, which involves multiple denoising steps at varying noise levels for each class, is computationally expensive. We aim to address these limitations by employing the Stable Diffusion model for discriminative tasks and introducing methods to accelerate its use as a zero-shot vision and language learners.

In this paper, we present simple but effective sampling techniques that reduce computational effort by up to a factor of 2. We systematically evaluate our methods on three benchmark classification datasets, demonstrating that our approach can significantly improve inference speed while maintaining comparable classification accuracy. This improvement brings us closer to making such diffusion-based zero-shot classifiers practical tools for a broad range of discriminative applications.

2 Preliminaries

This section provides a brief overview of diffusion models and how they can be used for zero-shot classification.

Diffusion Models: Diffusion models are latent variable generative models defined by a forward and a reverse Markov chain (Norris, 1998). Suppose we have data distributed as $q(\boldsymbol{x}_0)$, with $\boldsymbol{x}_0 \in R^d$. The forward process gradually adds Gaussian noise to generate a sequence of noisy variables $\boldsymbol{x}_{1:T} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$:

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}). \quad (1)$$

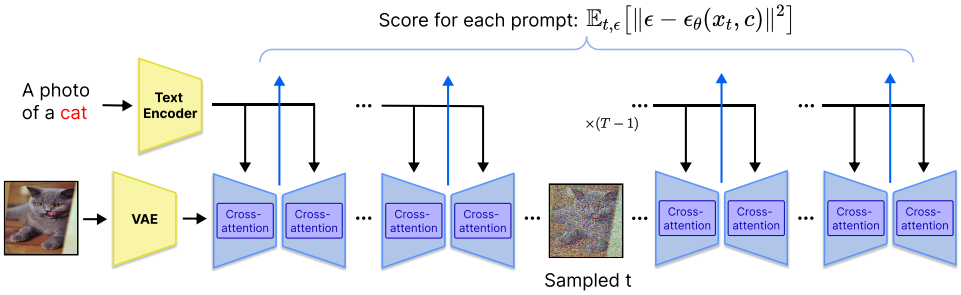

Figure 1: Diagram illustrating the use of diffusion models for zero-shot classification. Scores are calculated for each text prompt, which are derived from class labels, at every sampled time step. The class corresponding to the lowest expected score is subsequently predicted.

The reverse process removes noise step by step, starting from $\text{Normal}(\boldsymbol{x}_T; 0, \boldsymbol{I})$:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=0}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t). \quad (2)$$

Following Kingma and Welling (2013), training involves optimizing a variational lower bound, which can be expressed as:

$$\mathcal{L}_{\text{Diffusion}} = E_{\boldsymbol{x}_t, \epsilon, \mathbf{c}, t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c})\|_2^2\right], \quad (3)$$

with $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, $\epsilon \sim \text{Normal}(0, \boldsymbol{I})$, $t \sim \mathcal{U}([0, T])$, and $\mathbf{c}$ a text embedding.

Using Diffusion Models for Zero-shot Classification In this section, we explain how a text-to-image diffusion model can be adapted as a zero-shot classifier for evaluation on downstream tasks. Figure 1 illustrates the idea.

Given an image $\boldsymbol{x}$, the goal is to predict the most probable class assignment

$$\begin{aligned} \hat{y} &= \arg\max_{\mathrm{y}_i} p\left(y = \mathrm{y}_i \mid \boldsymbol{x}\right) \\ &= \arg\max_{\mathrm{y}_i} p\left(\boldsymbol{x} \mid y = \mathrm{y}_i\right) \cdot p\left(y = \mathrm{y}_i\right) \\ &= \arg\max_{\mathrm{y}_i} \log p\left(\boldsymbol{x} \mid y = \mathrm{y}_i\right), \end{aligned} \quad (4)$$

where we assume a uniform prior $p\left(y = \mathrm{y}_i\right) = \frac{1}{k}$ that can be dropped from the arg max.

Convert the label y_i from each class name into text prompts using a dataset-specific template (e.g. $\mathrm{y}_i \rightarrow \mathbf{c}_i$: A photo of a y_i). Then we can convert eq. 4 to be solved via VLB (Kingma and Welling, 2013) by:

$$\begin{aligned} \hat{y} &= \arg\max_{\mathrm{y}_i} \log p_\theta\left(\boldsymbol{x} \mid y = \mathrm{y}_i\right) \\ &\approx \arg\min_{\mathrm{y}_i} \mathcal{L}_{\text{Diffusion}}\left(\boldsymbol{x}, \mathrm{y}_i\right) \\ &= \arg\min_{\mathrm{y}_i \in [\mathrm{y}_i]} E_{t,\epsilon}\left[\boldsymbol{w}_t \left\|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, \mathbf{c}\right)\right\|^2\right], \end{aligned} \quad (5)$$

and $\boldsymbol{w}_t$ is a weight assigned to the timestep t.

3 Accelerated Sampling

In this section, we introduce an improved, hierarchical sampling strategy that enhances the efficiency of the sampling process for using pretrained diffusion models as classifier and optimizes the process of class prediction.

Monte-Carlo Estimation of Expectation The expectation in Eq. 5 is approximated using Monte Carlo estimation. We start by sampling the time step t and then deriving x_t in accordance with the forward diffusion process (Eq. 1): $\boldsymbol{x}_t \sim q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right)$.

Class Scoring and Prediction Upon obtaining a noisy image, we apply Stable Diffusion to denoise and predict $\boldsymbol{x}$ from $\boldsymbol{x}_t$, yielding $\hat{\epsilon} = \epsilon\theta\left(\boldsymbol{x}_t, \mathbf{c}, t\right)$. We designate the squared error of the prediction, $||\epsilon - \hat{\epsilon}||_2^2$, as the score for $(\boldsymbol{x}, \mathrm{y}_i)$. We compute this score for each class N times. The final step involves weighting the scores based on the corresponding $\boldsymbol{w}_t$ and averaging them across all sampled timesteps to generate a prediction score for each class.

Hierarchical Sampling Strategy In contrast to the conventional sampling strategy suggested by Li et al. (2023a), which allocates equal sample numbers to each class at every timestep, our approach places emphasis on classes with higher prediction probabilities.

Our strategy maintains a beam of classes, initially sized to C/b, where C is the total number of classes in the dataset and b is the `BeamFactor` hyper-parameter which determines the number of classes to retain during the sampling process. Given an input text prompt $\mathbf{c}$, the process begins by sampling N instances from the starting timestep

Dataset	Diffusion Classifier		Accelerated Diffusion Classifier	
	Accuracy	Inference Time	Accuracy	Inference Time
CIFAR10	86.6	110h	85.5	57h
STL10	94.7	39h	91.0	18h
FGVC	23.9	19h	23.1	11h

Table 1: Comparison of zero-shot classification performance and inference time between Diffusion Classifier and Accelerated Diffusion Classifier. The Accelerated Diffusion Classifier exhibits comparable performance to the Diffusion Classifier, albeit with the inference time for each dataset being approximately twice as long.

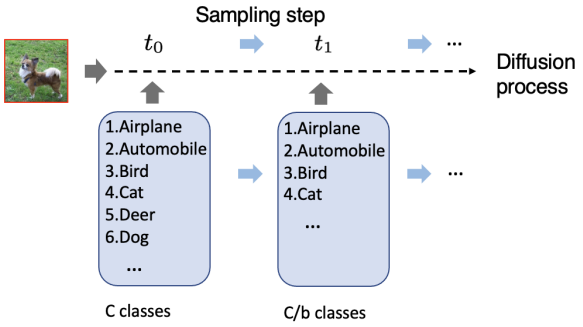

Figure 2: The illustration of the proposed accelerated sampling. Starting from the initial time step and progressing to the final one, we retain a progressively diminished number of class labels throughout the process.

Dataset	CLIP ResNet-50	Accelerated Diffusion Classifier
CIFAR10	75.6	85.5
STL10	94.3	91.0
FGVC	19.3	23.1

Table 2: The comparison of zero-shot classification performance with CLIP ResNet-50 performance. As can be observed from the three datasets, the performance of accelerated diffusion classifier is comparable.

Algorithm 1 Accelerated Diffusion Classification

Given: Image $\boldsymbol{x}$, conditioning inputs $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^{n}$, starting time step t_0, ending time step T, time interval Δt, sampling points for each time step N, and BeamFactor b.
Initialize $\texttt{Score}[\mathbf{c}_i] = \texttt{list}()$ for each class name $\mathbf{c}_i$
Calculate 'BeamSize' = C/b
for time steps $t = t_0, t_0 + \Delta t, \ldots, T$ **do**
  for Sampling $n \in N$ **do**
    Sample $\epsilon \sim \mathcal{N}(0, I)$
    $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x} + \sqrt{1 - \bar{\alpha}_t}\epsilon$
    for conditioning $\mathbf{c}_i \in \mathcal{C}$ **do**
      $\texttt{Score}[\mathbf{c}_i]$.append($\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}_i)\|^2$)
    end for
    Refine $\mathcal{C}$ with 'BeamSize' classes of C with lowest mean error.
  end for
end for
$\hat{y} \leftarrow \arg\min_{\mathbf{c}_i} \sum_t \text{mean}(\texttt{Score}[\mathbf{c}_i])$

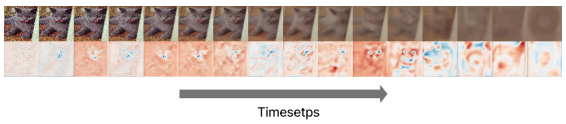

Figure 3: Visualization of VAE-decoded latent features and corresponding heatmaps at different sampled timesteps. The prompt is *"a photo of a dog."*

t_0. However, unlike previous methods, this process is not repeated for every class. Instead, we retain only the top C/b classes that demonstrate the highest performance after each timestep. This selective approach continues until a single class consistently achieves the highest probability across t_s additional samplings or each timestep has been sampled for N times. This process is visually represented in Figure 2.

A detailed description of the complete algorithm is provided in Algorithm 1.

4 Experiments

Baselines We compared our accelerated diffusion classifier with the baseline of standard diffusion classifier introduced in Li et al. (2023a).

Dataset We evaluate the zero-shot classification performance across three datasets: CIFAR10 (Krizhevsky and Hinton, 2009), STL10 (Coates et al., 2011), and FGVC-Aircraft (Maji et al., 2013).

4.1 Experimental Results

In Table 1, we show that the Accelerated Diffusion Classifier exhibits comparable performance to the Diffusion Classifier, albeit with the inference time for each dataset being approximately twice as long. In Table 2, we shows the comparison with CLIP using ResNet-50 as the backbone and ours can achieve competitive performance as well.

Figure 3 provides a qualitative analysis. The upper row shows VAE-decoded images of latent features at various sampled timesteps, while the

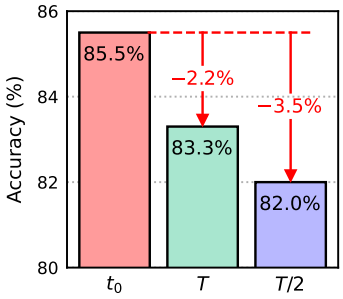

Figure 4: The compare of different sampling start timestep.

lower row depicts their corresponding heatmaps. As the timesteps progress, the image becomes increasingly coherent, clearly depicting the features of a dog. This highlights the alignment between the text prompt and the visual representation, showcasing the effectiveness of the hierarchical sampling in capturing semantic correlations.

4.2 Ablation Study: Sampling from Other Directions

We additionally conducted experiments to investigate the impact of alternative hierarchical sampling directions. The two alternative strategies we employed were: 1. Sampling in reverse order, that is, we begin with the noisy image and gradually reduce the sampling numbers as the images become cleaner. 2. Initiating sampling from the midpoint $T/2$ and then sampling time t from a uniform distribution in the range $[T/2 - \Delta t, T/2 + \Delta t]$.

The results of these experiments are displayed in Figure 4. It can be observed that starting from t_0, corresponding to the cleanest image, yields the highest accuracy. This intuitively makes sense as noiseless images contain more information that can be better aligned with the text prompts.

5 Related Work

Diffusion Probabilistic Models (DPMs) Diffusion Probabilistic Models (DPMs), encompassing diffusion (Sohl-Dickstein et al., 2015) and score-based generative models (Song and Ermon, 2019), have emerged as potent tools for image generation in recent years. The evolution of DPMs over the past couple of years has been marked by significant enhancements, particularly in sampling techniques like classifier-free guidance (Ho and Salimans, 2021). DPMs conventionally leverage convolutional U-Net architectures (Ronneberger et al., 2015), incorporating cross-attention layers. These works have demonstrated the diffusion models' potential ability via minimizing the objective of the distance between the predicted noise and the ground truth during training. These objectives can be extended to implementing diffusion models as classifiers (Li et al., 2023a; Clark and Jaini, 2023) and we further improve the efficiency in this work.

Generative Models for Discriminative Tasks The potential of generative models in discriminative tasks has been a focal point in recent research in the fields of natural language processing and machine learning. One prevalent approach involves fine-tuning the model for a specific discriminative task. For instance, Dai et al. (2021) improved performance on several discriminative tasks, including named entity recognition and machine translation, by fine-tuning a large transformer-based language model. In a similar vein, Yang et al. (2019) introduced a model that amalgamates a generative model for text generation and a discriminative model for sentiment analysis. Wang et al. (2018) proposed a method for adapting a generative model for language translation to the discriminative task of language classification, yielding superior results compared to established baselines. There are also instances of using generative models for discriminative tasks, such as initializing a discriminative model to enhance performance (Mao et al., 2019), or pre-training a discriminative model for domain adaptation (Chen et al., 2020). Recent studies (Li et al., 2023a; Clark and Jaini, 2023) propose the application of pre-trained diffusion models for zero-shot classification. Also, Wei et al. (2023) have restructured diffusion models as masked autoencoders, achieving state-of-the-art classification accuracy in video tasks.

6 Conclusion

Our work offers novel insights into the capabilities of stable diffusion. We posit that text-to-image generative diffusion models can learn powerful representations and serve as an efficient and fast vision and language learner. The hierarchical sampling strategy we introduce serves as a stepping stone towards making these models more accessible and practical for a wider range of applications, thereby unlocking new potential for their deployment in real-world scenarios.

7 Limitations

Our approach relies on pre-trained text-to-image diffusion models, which may inherit biases and ethical concerns from the data used during their initial training. These biases could influence predictions in unintended ways, particularly in applications involving sensitive or diverse datasets. While our method improves the efficiency of using such models for discriminative tasks, addressing these underlying biases and ensuring ethical deployment remains an important area for future research.

References

Yizhe Chen, Nan Duan, Wenbing Hu, Defeng Lian, Zhoujun Lu, Bin Xu, Yongyang Liu, Bing Dai, Ruslan Salakhutdinov, and Quoc Le. 2020. Pre-training with conditional generative adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*, 8:237–249.

Kevin Clark and Priyank Jaini. 2023. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

Zhilin Dai, Zihang Yang, Alec Fan, Jing Gao, Jaime Carbonell, and Quoc Le. 2021. Fine-tuning generative models for discriminative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical Report.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. 2023a. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Junhua Mao, Bin Xu, Yongyang Liu, Yiming Yang, Bing Dai, Ruslan Salakhutdinov, and Quoc Le. 2019. Improving discriminative models with generative pre-training. In *Advances in Neural Information Processing Systems*.

James R Norris. 1998. *Markov chains*. 2. Cambridge university press.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv: 2307.01952*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. pages 234–241.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*.

Yiming Wang, Amanpreet Singh, Jian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Adapting generative models for discriminative tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. 2023. Diffusion models as masked autoencoders. *arXiv preprint arXiv:2304.03283*.

Zichao Yang, Zhilin Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc Le. 2019. Combining generative and discriminative models for improved sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.